



École Doctorale  
d'Informatique,  
Télécommunications  
et Électronique de Paris

# Thèse

présentée pour obtenir le grade de docteur

de Telecom ParisTech

Spécialité : Signal et Image

**Marco PALEARI**

**Informatique Affective:  
Affichage, Reconnaissance, et Synthèse  
par Ordinateur des Émotions**

Soutenue le 12 octobre 2009 devant le jury composé de

Niculae Sebe  
Jean-Claude Martin  
Ryad Chellali  
Gaël Richard  
Benoit Huet

President  
Rapporteur  
Rapporteur

Directeur de thèse





École Doctorale  
d'Informatique,  
Télécommunications  
et Électronique de Paris

# Thesis

submitted in partial fulfillment of the requirements for the  
degree of Doctor of Philosophy

from Telecom ParisTech

Track : Multimedia

**Marco PALEARI**

**Affective Computing.  
Display, Recognition, and Computer  
Synthesis of Emotions**

Presented October the 12th 2009 to

Niculae Sebe  
Jean-Claude Martin  
Ryad Chellali  
Gaël Richard  
Benoit Huet

President  
Examiner  
Examiner

Thesis Advisor



*Don't part with your illusions.  
When they are gone you may still exist, but you  
have ceased to live.*

Mark Twain

*The question is not whether intelligent machines  
can have emotions, but whether machines can be  
intelligent without any emotions*

Marvin Minsk

*La ragione umana, anche senza il pungolo della  
semplice vanità dell'onniscienza, è perpetuamente  
sospinta da un proprio bisogno verso quei problemi  
che non possono in nessun modo esser risolti da un  
uso empirico della ragione... e così in tutti gli  
uomini una qualche metafisica è sempre esistita e  
sempre esisterà, appena che la ragione s'innalzi alla  
speculazione.*

Immanuel Kant

---



## Abstract

Affective Computing refers to computing that relates to, arises from, or deliberately influences emotions and has its natural application domain in highly abstracted human-computer interactions. Affective computing can be divided into three main parts, namely display, recognition, and synthesis.

The design of intelligent machines able to create natural interactions with the users necessarily implies the use of affective computing technologies. We propose a generic architecture based on the framework "Multimodal Affective User Interface" by Lisetti and the psychological "Component Process Theory" by Scherer which puts the user at the center of the loop exploiting these three parts of affective computing.

We propose a novel system performing automatic, real-time, emotion recognition through the analysis of human facial expressions and vocal prosody. We also discuss about the generation of believable facial expressions for different platforms and we detail our system based on Scherer theory. Finally we propose an intelligent architecture that we have developed capable of simulating the process of appraisal of emotions as described by Scherer.

## Résumé

L'informatique Affective regarde la computation que se rapporte, surgit de, ou influence délibérément les émotions et trouve son domaine d'application naturel dans les interactions homme-machine à haut niveau d'abstraction.

L'informatique affective peut être divisée en trois sujets principaux, à savoir: l'affichage, l'identification, et la synthèse. La construction d'une machine intelligente capable d'interagir de façon naturelle avec son utilisateur passe forcément par ces trois phases. Dans cette thèse nous proposons une architecture basée principalement sur le modèle dite "Multimodal Affective User Interface" de Lisetti et la théorie psychologique des émotions nommé "Component Process Theory" de Scherer.

Dans nos travaux nous avons donc recherché des techniques pour l'extraction automatique et en temps-réel des émotions par moyen des expressions faciales et de la prosodie vocale. Nous avons aussi traité les problématiques inhérentes la génération d'expressions sur de différentes plateformes, soit elles des agents virtuel ou robotique. Finalement, nous avons proposé et développé une architecture pour des agents intelligents capable de simuler le processus humaine d'évaluation des émotions comme décrit par Scherer.

---





## Acknowledgments

This document is the result of more than three years of research. During these three years, here at EURECOM I have been interacting with some of the most wonderful people I know and I would like you all to know.

I am greatly indebted with my Dr. Benoit Huet who succeeded Christine L. Lisetti and Brian Duffy in the direction of this thesis and allowed me to finish it in a friendly and stimulating environment. I would like to thank Dr. Brian Duffy, with whom I started the Ph.D. but that had, unfortunately, to leave too soon. I would like to also thank Professor Christine L. Lisetti, my original thesis director, who allowed me to start a Ph.D. in this fantastic environment and, by letting me know the topic of Affective Computing, greatly influenced my research. I also would like to express my gratitude to Professor Jean-Luc Dugelay and Professor Dirk Sloock with whom I had the chance to work and to learn something.

I would really like to thank the members of my jury. I am really thankful to Professor Gaël Richard (Telecom ParisTech), Professor Nicu Sebe (University of Trento), Professor Jean-Claude Martin (LIMSI-CNRS, University of Paris XI), and Professor Ryad Chellali (Italian Institute of Technology) for having the patience of reading this long document and giving me, I am sure, very interesting feedbacks and ideas to think about.

I would really like to thank all of my colleagues and friends here at EURECOM: I spent some great time with you. Some people in particular will stay in my hearth forever for I have spent with them some of the best moments of my life: thank you Kostas for the support and the company; thank you Nesli and Ufuk, you are really wonderful people; thank you Ani and all of my other office-mates for being so patients with me, thank you Federico, Remi, Emilie, Gerardine, Josephine, and Cyril for the nights spent role-playing; thanks Virginia for the long afternoons spent discussing very relevant work-related matters; thanks all of my team of sailing starting with Daniele and Dimitri for allowing us to go to the final of the French championship and again thank EURECOM for supporting us in this adventure. Thank you Amandine, Olivier, Rachid, Antony, Carmelo, and all of my Master's students (Sofi, Silvia, Lavi, Pauline, and all the others) for the interesting work we shared but most of all for your friendship. Thank you to all of my other colleagues: secretaries, IT service, soccer buddies, volley-ball buddies, beach-and-guitar buddies, and, believe it or not, most of the students too: without you this experience would not have been so great; unfortunately, there is no space enough for reporting all of your names but you all are wonderful people.

Most of all, I want to thank my family, my parents, my brother, and my sister for always believing that soon or later people will accept that I am smarter than Einstein: I love you all too. Thanks too, to my friends in Bra and Turin for believing intelligent robots are among us, or simply for letting me have something to miss while I was here having fun in French Riviera: you are the bests.

Finally, thanks to my other family here in French Riviera, with whom I shared most of my difficulties and of my great moments and without whom I might as well have ended my adventure earlier; in particular thanks to Alessandro, Corrado, and Xiaolan with whom I shared the last three years but also to all of my previous housemates.

---



---

# Contents

<b>1</b>	<b>Introduction générale</b>	<b>13</b>
1.1	Affective Computing . . . . .	13
1.2	Émotions . . . . .	14
1.3	Scénarios d'Application . . . . .	15
1.3.1	Taches de l'informatique affective . . . . .	19
1.4	Contributions de la Thèse . . . . .	20
1.4.1	Partie 1: Affichage d'émotion . . . . .	21
1.4.2	Partie 2: Reconnaissance d'émotion . . . . .	21
1.4.3	Partie 3: Synthèse d'émotion . . . . .	23
1.4.4	Reconnaissance biométrique . . . . .	24
1.4.5	Transcription automatique de la musique . . . . .	25
1.5	Contenu du document . . . . .	26
<b>2</b>	<b>General Introduction</b>	<b>31</b>
2.1	Affective Computing . . . . .	31
2.2	Emotions . . . . .	32
2.3	Application Scenarios and Emotion Related Tasks . . . . .	33
2.4	Thesis Contributions . . . . .	34
2.5	Outline . . . . .	36
<b>3</b>	<b>Psychological Background</b>	<b>39</b>
3.1	Affective Computing . . . . .	39
3.1.1	Emotion and the Human Brain . . . . .	39
3.1.1.1	Emotion and Perception . . . . .	39
3.1.1.2	Descartes Error . . . . .	40
3.1.2	Emotion and Communication . . . . .	41
3.1.3	Emotion and Human-Computer-Interactions . . . . .	42
3.2	Applications of Affective Computing . . . . .	43
3.2.1	e-Learning . . . . .	43
3.2.2	Medicine and Tele-medicine . . . . .	44
3.2.3	Gaming . . . . .	45
3.2.4	Indexing and Retrieval . . . . .	47
3.2.5	Personal Intelligent Agent . . . . .	48
3.2.6	Other Scenarios . . . . .	49
3.3	Affective States . . . . .	50
3.3.1	What is an emotion? . . . . .	51
3.3.2	Structured Model of Emotions . . . . .	52
3.3.2.1	Discrete Categories . . . . .	52

---

3.3.2.2	Dimensional Emotion Descriptions . . . . .	53
3.3.2.3	Componential Models . . . . .	56
<b>4</b>	<b>Emotional Display</b>	<b>63</b>
4.1	Introduction . . . . .	63
4.2	Relevant Work . . . . .	63
4.2.1	Scherer's Component Process Theory for Emotional Expressions . . . . .	64
4.2.2	Emotional Facial Expressions . . . . .	66
4.2.2.1	Facial Expression Models . . . . .	67
4.2.2.2	Action Unit and FACS . . . . .	67
4.2.2.3	MPEG-4 FAPs . . . . .	68
4.2.2.4	3D Graphics Animation Engines . . . . .	71
4.2.2.5	GRETA . . . . .	71
4.2.2.6	X-Face . . . . .	72
4.2.2.7	Galatea . . . . .	73
4.2.2.8	Haptek . . . . .	74
4.2.2.9	Robotic Platforms . . . . .	75
4.2.2.10	AIBO . . . . .	75
4.2.2.11	Papero . . . . .	76
4.2.2.12	Kismet . . . . .	76
4.2.2.13	Philips iCat . . . . .	77
4.2.3	Emotional Speech . . . . .	78
4.2.3.1	Speech Models . . . . .	78
4.2.3.2	Emotional Speech Engines . . . . .	81
4.2.3.3	Festival . . . . .	81
4.2.3.4	Loquendo . . . . .	82
4.2.3.5	Mary . . . . .	82
4.3	Building Believable Facial Expressions . . . . .	82
4.3.1	Cherry, the Haptek Avatar . . . . .	83
4.3.1.1	Step 1 - Converting SECs to AUs . . . . .	83
4.3.1.2	Step 2 - Converting AUs to Haptek parameters . . . . .	83
4.3.1.3	Step 3 - Finding the right intensities for the AUs . . . . .	86
4.3.1.4	Step 4 - Exploiting temporal information . . . . .	86
4.3.1.5	Implementation details . . . . .	88
4.3.2	Cleo, the Philips Robot . . . . .	90
4.3.2.1	Step 1 - Converting SECs to AUs . . . . .	90
4.3.2.2	Step 2 - Converting AU to iCat servo controls . . . . .	90
4.3.2.3	Step 3 - Finding the right intensities for the AUs . . . . .	90
4.3.2.4	Step 4 - Exploiting temporal information . . . . .	90
4.3.2.5	Implementation details . . . . .	91
4.4	Results . . . . .	92
4.4.1	Open Questions for Research in Psychology . . . . .	92
4.4.2	User Study . . . . .	92
4.4.2.1	Cherry, the Haptek Avatar . . . . .	92
4.4.2.2	Cleo, the Philips Robot . . . . .	93
4.5	Concluding Remarks . . . . .	95

---

<b>5</b>	<b>Emotional Recognition</b>	<b>99</b>
5.1	Introduction . . . . .	99
5.2	Relevant Work . . . . .	99
5.2.1	Emotion Recognition via Facial Expressions . . . . .	100
5.2.2	Emotion Recognition via Speech Signal . . . . .	103
5.2.3	Physiological Signals and Alternative Emotion Recognition Systems	106
5.2.4	Multimodal Emotion Recognition Systems . . . . .	106
5.3	ARAVR: Automatic Real-Time Audio-Video Emotion Recognition . . .	106
5.3.1	Databases . . . . .	107
5.3.2	eNTERFACE'05 . . . . .	108
5.3.3	Facial Expression Feature Extraction . . . . .	110
5.3.4	Step 1: Face Detection . . . . .	110
5.3.5	Step 2: ROI definition . . . . .	111
5.3.6	Step 3: Feature Point Extraction . . . . .	114
5.3.7	Prosodic Expression Feature Extraction . . . . .	116
5.3.8	Feature Vector Construction . . . . .	117
5.3.9	Machine Learning . . . . .	119
5.3.10	Multimodal Fusion . . . . .	121
5.3.11	AMMAF: A Multimodal Multilayer Affect Fusion Paradigm . . . .	124
5.4	Results . . . . .	126
5.4.1	Metrics . . . . .	126
5.4.2	Dense Motion Flow vs. Feature Point . . . . .	127
5.4.3	Audio vs. Video vs. Audio-Video . . . . .	128
5.4.4	Feature vector: statistical or polynomial analysis . . . . .	129
5.4.5	Comparison Among Features . . . . .	130
5.4.6	Feature Extraction . . . . .	131
5.4.7	Analysis Procedure . . . . .	132
5.4.8	Results . . . . .	133
5.4.9	Multimodal Fusion . . . . .	144
5.4.10	Detectors or Classifier . . . . .	147
5.4.11	Machine Learning Approaches . . . . .	148
5.4.12	Data Post-Treatment . . . . .	148
5.4.13	Resulting System . . . . .	153
5.5	Perspectives: SAMMI . . . . .	156
5.6	Concluding Remarks . . . . .	161
<b>6</b>	<b>Emotion and Artificial Intelligence</b>	<b>163</b>
6.1	Introduction . . . . .	163
6.2	Relevant Work . . . . .	164
6.2.1	Basics on Artificial Intelligence . . . . .	165
6.2.1.1	BDI . . . . .	165
6.2.1.2	Belief Networks . . . . .	167
6.2.1.3	Decision Networks . . . . .	168
6.2.1.4	Dynamic Belief and Decision Networks . . . . .	168
6.2.1.5	Utility Theory . . . . .	170
6.2.2	CPT and OCC Model of Emotions for AI . . . . .	170
6.2.3	Existing AI with Emotions . . . . .	171
6.2.3.1	GRETA . . . . .	171

---

6.2.3.2	EMA . . . . .	171
6.2.3.3	VALERIE . . . . .	175
6.3	Generic AI: ALICIA . . . . .	177
6.3.1	BDI+E . . . . .	178
6.3.1.1	The reactive layer . . . . .	180
6.3.1.2	Behavioral Level . . . . .	180
6.3.1.3	The deliberative layer . . . . .	181
6.3.1.4	The physical layer . . . . .	182
6.3.2	Modules descriptions . . . . .	183
6.3.2.1	Jimmy and Dr. Tom Example . . . . .	190
6.3.3	Demo Implementation . . . . .	191
6.3.3.1	Proof of concept scenarios . . . . .	194
6.4	Concluding Remarks . . . . .	195
<b>7</b>	<b>Summary and Conclusion</b>	<b>197</b>
<b>I</b>	<b>APPENDICES</b>	<b>201</b>
<b>A</b>	<b>AU Table</b>	<b>203</b>
<b>B</b>	<b>Alternative Feature Point Detection</b>	<b>209</b>
B.1	Introduction . . . . .	209
B.2	ROI Definition . . . . .	209
B.3	Feature Point extraction . . . . .	209
B.3.1	Eyebrow ROI processing . . . . .	210
B.3.2	Eye ROI processing . . . . .	213
B.3.3	Mouth ROI processing . . . . .	213
B.4	Conclusions . . . . .	214
<b>C</b>	<b>Biometrics</b>	<b>215</b>
C.1	System . . . . .	216
C.1.1	Expression analysis . . . . .	218
C.2	Experimental results and analysis . . . . .	219
C.3	Conclusions . . . . .	222
<b>D</b>	<b>Automatic Music Transcription</b>	<b>223</b>
D.1	Introduction . . . . .	223
D.2	Guitar Transcription . . . . .	224
D.2.1	Automatic Fretboard Detection . . . . .	224
D.2.2	Fretboard Tracking . . . . .	225
D.2.3	Hand Detection . . . . .	225
D.2.4	Audio Visual Information Fusion . . . . .	226
D.3	Prototype . . . . .	227
D.4	Future Work . . . . .	228
D.5	Conclusions . . . . .	229
	<b>Bibliography</b>	<b>250</b>

---

# Chapter 1

## Introduction générale

Les ordinateurs et les autres appareils électroniques sont de plus en plus présents dans notre vie au quotidien. Améliorer la qualité des interactions homme-machine et des communications à travers d'une machine devient donc, un sujet de recherche très pertinent et très important.

L'un des sujets les plus chauds dans le domaine de l'interaction homme-machine et en général l'informatique-centré-l'homme est, maintenant, informatique affective, un sujet de recherche intéressant à la frontière de l'informatique, de la psychologie et des sciences cognitives. Les raisons qui motivent le traitement, l'affichage, et la compression des émotions sont nombreuses (Picard [1997]).

### 1.1 Affective Computing

*Informatique affective est informatique qui se réfère à, en résulte, ou influence délibérément les émotions. [...] Informatique affective comprend également de nombreuses autres choses, comme donner à un ordinateur la capacité de reconnaître et exprimer des émotions, développer sa capacité de répondre intelligemment à l'émotion de l'homme, et de lui permettre de régler et d'utiliser ses émotions.*

Ceci est la définition originale par Rosalind Picard de l'informatique affective dans le livre "*Affective Computing*" (Picard [1997]).

Il y a beaucoup de raisons motivant informatique affective.

Les chercheurs ont démontré que, pendant les interactions homme-machine de tous les jours les gens ont tendance à supposer que les ordinateurs ont des capacités émotionnelles (Covey [2004]). Reeves and Nass [1998] ont montré que nous interagissons souvent avec les ordinateurs comme si eux étaient capables de comprendre les émotions. Koda and Maes [1996] a démontré que, dans certains domaines, les gens préfèrent interagir avec des machines capables d'afficher une sorte de capacités émotionnelles. Fait intéressant, Koda et Maes démontrent aussi que des gens qui ont affirmé que l'ordinateur ne doit pas afficher des émotions préfèrent, enfin, les interactions avec ce genre d'agents émotionnels. Covey [2004] va au-delà de ces résultats en affirmant que, d'une quelconque façon les humains "ont besoin" de croire que toute être intelligent aille quelque sorte de comportement affectif. En effet d'après Covey les humains n'arrivent pas à imaginer l'intelligence sans émotions car cela serait trop contre nature et effrayant. Il est intéressant de noter que plus la machine ressemble à l'homme (à savoir un robot par rapport à un

---

avatar ou une interface textuelle) plus nous avons tendance à percevoir le comportement de la machine aussi intelligente que émotionnelle.

Deuxièmement, de nombreux chercheurs ont démontré que les émotions ont une forte influence sur de nombreuses fonctions cognitives telles que: la prise de décision (Damasio [2005]), la perception (Halberstadt et al. [1995]) et son interprétation (Bouhuys et al. [1995]), les communications (Mehrabian and Wiener [1967]), et beaucoup d'autres (Hudlicka and Fellous [1996], Lisetti and Gmytrasiewicz [2002]). Goleman [2006] soutient que l'intelligence émotionnelle (à savoir l'ensemble de toutes les capacités liées aux émotions) est généralement plus corrélées au "*succès dans la vie*" que le quotient intellectuel (QI).

Toutes ces raisons motivent la nécessité de rechercher et construire des ordinateurs de plus en plus capables de reconnaître, traiter et afficher les émotions afin d'améliorer le naturel, et l'efficacité des interactions homme-machine, des communication par ordinateur, et de la informatique-centre-à-l'homme (human-centered-computing) en général.

En particulier, nous soutenons que, dans de nombreuses situations, les logiciels devraient être conçus en gardant à l'esprit que l'homme doit toujours être mis au milieu d'une boucle dans laquelle l'interaction avec l'ordinateur ou le robot est "couplé avec" et "médiée par" les émotions comme dans la figure 1.1. Ces considérations doivent, par conséquence, être gardées à l'esprit pendant toutes les étapes de la conception d'un agent informatique, du tout premier dessin jusqu'à l'implémentation finale.

## 1.2 Émotions

Les émotions sont étudiées dans de nombreux domaines différents. Psychologues, physiologistes, biologistes, anthropologues, sociologues, philosophes et même des éthologues étudient les émotions (Scherer and Ekman [1984]).

La définition de "l'émotion" n'est donc pas claire; cependant la communauté scientifique semble d'accord sur le fait que les émotions soient des phénomènes psychophysiques, qui ont des origines dans les réactions aux événements et qui sont motivés par les théories de l'évolution darwinienne comme les autres caractéristiques physiques des êtres vivants (Darwin [1872]).

En d'autres termes, les émotions surgissent comme un processus d'évaluation des événements qui nous entourent et qui ont deux composantes différentes: la première est *cognitive* (elle a lieu dans notre cerveau et influence sur notre façon de penser), et la deuxième est *physique* (il a toujours origine dans le cerveau, mais influence notre corps, nos manières d'agir, la façon dont nous percevons, etc.)

Pendant les longues années de recherche sur les émotions, les chercheurs ont défini de nombreux systèmes de classification de ces phénomènes. Trois grandes familles peuvent être délimitées, à savoir:

- **catégories émotionnelles discrètes:** les émotions sont caractérisées par un mot clé ou une étiquette identifiant une famille d'émotions (joie, par exemple);
- **modèles dimensionnels d'émotions:** les émotions sont caractérisées par une position dans un espace multidimensionnel dans lequel l'axe de base représentent des concepts émotionnels tels que *Valence* ou *Intensité* (par exemple le bonheur peut être représenté comme  $B = [0.8, 0.7]$  car c'est une émotion caractérisée par des valeurs de valence et intensité positives);



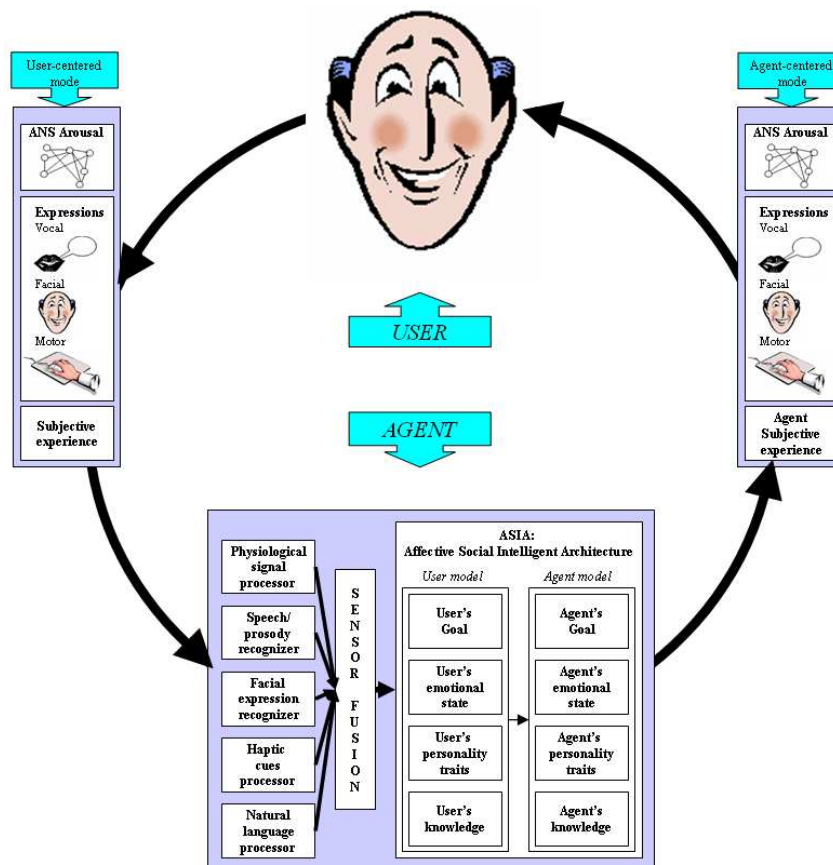


Figure 1.1: Human Centered Affective Multimodal User Interface. (adapté à partir de Lisetti and Nasoz [2002])

- **modèles des émotions par composantes:** les émotions sont caractérisées par un ensemble de composantes qui représentent les phases d'appréciation des émotions elles-mêmes (par exemple, le bonheur est l'émotion résultant d'un événement inattendu qui est agréable car il facilite l'obtention des objectifs de l'agent).

### 1.3 Scénarios d'Application

Il y a des nombreux exemples prototypiques de l'utilisation des émotions en informatique tel que la télé-médecine, l'e-learning, les jeux, l'indexation et la recherche de média, l'intelligence artificielle en générale, les systèmes de messagerie instantanée, et autre forme de télé-communication.

**E-learning.** Presque toutes les écoles ne disposent pas d'assez de temps ou de budget pour permettre à un instructeur de s'asseoir et aider chaque élève individuellement. Cela peut être frustrant pour les élèves qui éprouvent des difficultés à comprendre un certain sujet. Fournir un assistant personnel virtuel (un Intelligent Tutoring Systems (ITS)), peut résoudre en partie ce problème car de cette façon, chaque élève peut avoir un tuteur personnel se référant à lui pour l'aider lorsqu'il a des soucis avec la matière.

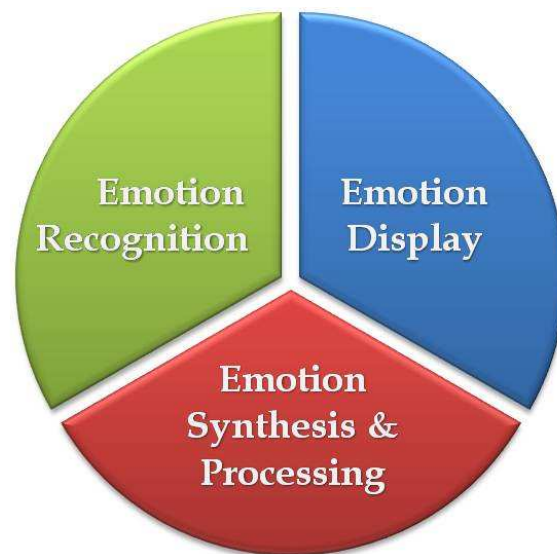


Figure 1.2: La computation affective et ses trois composantes

Si les émotions ne sont pas toujours considérées dans les ITS, plusieurs chercheurs (Picard [1997], Lisetti and Gmytrasiewicz [2002]) soutiennent que les émotions sont fondamentales dans la communication et dans les interactions humaines en général. Peu d'études ont appliqué les émotions pour les environnements d'apprentissage (Kapoor et al. [2001], Kapoor and Picard [2005], Conati [2002], Conati and Zhou [2002]), démontrant ainsi que est possible augmenter le plaisir et l'efficacité de l'apprentissage des élèves qui utilisent ce genre d'applications.

En effet, il est bien connu qu'un niveau de stress soigneusement choisis généralement permet d'améliorer les performances de l'homme (voir la figure 3.1) (Hebb [1966]), et notamment les performances pendant l'apprentissage (Yerkes and Dodson [1908]).

Dans les télé-applications de l'e-learning, à savoir les applications où les élèves interagissent avec des professeurs réels à travers un outil informatique, l'information sur l'état affectif des étudiants pourrait se perdre en raison des limites intrinsèques aux communications par ordinateur. En présence de ce genre d'applicatif émotionnel, l'enseignant peut, par conséquent, tirer profit de la disponibilité de l'information sur l'état affectif des élèves et modifier le style d'enseignement, faisant des pauses, et réagissant généralement de façon naturelle, par rapport à cet information avec le résultat d'améliorer l'efficacité de son enseignement.

**Médecine et de télé-médecine.** Le rire et le bonheur sont bien connus pour avoir des effets positifs sur le corps humain. Un état affectif positif aide à lutter contre le stress et réduire la douleur en libérant les "*endorphines*", mais il a également un effet positif sur les systèmes cardio-vasculaire et respiratoire, détend les muscles et renforce le système immunitaire en augmentant le nombre de "*cellules T*" et en abaissant niveaux de "*cortisol sérique*" ((Adams and McGuire [1986], Goodman [1992], Fry [1994], Martin [2002, 2004], Davidson et al. [2002], Lisetti and LeRouge [2004]).

Alors qu'en face-à-face, il est facile pour le médecin d'évaluer l'état émotionnel du patient et de réagir en conséquence, il pourrait être difficile d'accéder aux mêmes informations via une interface informatique. Des interfaces dédiées et des logiciels d'évaluation

de l'émotion doivent alors être créés pour donner au médecin les informations nécessaires en lui permettant de réagir correctement aux états émotionnelles négatifs.

**Jeux Vidéo.** Les jeux sont de plus en plus complexes et ils emploient des graphiques de pointe, des adversaires artificiellement intelligents, de manettes à retour de force, et systèmes d'affichage immersifs. Les émotions jouent un rôle clé dans l'expérience d'utilisateur, à la fois dans des jeux développés uniquement pour des fins de divertissement, et dans des jeux développés pour l'éducation, la formation, la thérapie ou la réadaptation.

En effet, l'informatique affective influence déjà l'état de l'art des jeux en plusieurs manières significatives:

- les personnages de jeu utilisent de plus en plus des expressions faciales vraisemblables pour raconter l'histoire;
- les personnages dans le jeu sont de plus en plus intelligents et commencent à simuler, les émotions comme des réactions aux événements dans l'histoire;
- les développeurs de jeux essaient d'influencer les émotions du joueur en créant artificiellement de sentiment de calme et de rythme (les événements de l'histoire, les musiques et les environnements sont souvent conçus et choisis avec cette fin);

Une autre façon intéressante d'utiliser les émotions dans l'environnement de jeu serait de détecter explicitement l'émotion du joueur et de modifier dynamiquement le jeu en fonction de celle-ci (Jones and Sutherland [2008]). Par exemple, un jeu sur ordinateur peut réagir et ajouter de nouveaux adversaires si l'humeur du joueur est en train de devenir trop calme en rapprochant l'état d'ennui ou, au contraire, diminuer la difficulté du jeu si le joueur était trop excité, stressé et fatigué.

**Indexation des multimédia.** De nos jours une énorme quantité de données est générée tous les jours par les utilisateurs du réseau Internet. Cette énorme quantité de données est générée par l'utilisateur et par conséquent indexées, recherchées, et récupérées grâce à des étiquettes de texte qui sont fournies par les utilisateurs lors du téléchargement des fichiers ou liées automatiquement aux médias grâce au texte qui les entoure.

Ce système n'est pas efficace, surtout car les étiquettes ont tendance à ne pas représenter complètement les médias et parce qu'une grande partie d'entre elles sont manquantes ou trompeuses. L'état de l'art sur l'indexation et la récupération de données est basé sur une technique de récupération d'information "basée-sur-le-contenu". Les systèmes de recherche actuels extraient les informations directement par le média et représentent, par exemple, une chanson par son tempo et timbre, une image par les couleurs, et un film par une combinaison de toutes ces caractéristiques.

Bien que les systèmes "basée-sur-le-contenu" ont l'avantage de représenter réellement le contenu des médias les chercheurs se battent encore pour définir des modes d'analyse capables de résoudre toutes les questions liées au thème de l'indexation et la récupération. Une question, que nous n'allons pas aborder dans le présent document concerne la capacité de ces technologies de fonctionner à l'échelle de contenu de la taille d'Internet.

La problématique d'intérêt de notre recherche est celle communément appelée "*fossé sémantique*". Les utilisateurs ont tendance à utiliser le langage naturel pour formuler

des requêtes, par exemple un utilisateur va rechercher “coucher de soleil romantique sur la plage” et non un tableau où les couleurs dominantes sont le rouge, orange et violet. D’une manière générale le fossé sémantique caractérise la différence entre les deux descriptions d’un objet par différentes représentations linguistiques; dans notre cas, celles qui sont d’une part, le langage humain, naturel, et d’autre part les informations représentatives de l’image à partir du contenu (par exemple les couleurs ou les textures).

Dans toutes les expressions de l’art, et donc dans des films, des images et la musique, les émotions sont une source fondamentale d’information. En fait, comme Ian Maitland, réalisateur et monteur, a déclaré: “*Un film est tout simplement une série d’émotions s’enchaînent avec une intrigue*” (Picard [1997]).

Une preuve de cela est le fait que nous avons tendance à décrire les genres cinématographiques avec de noms d’émotions: par exemple on dirait un film “romantique”, une comédie “drôle” ou un film d’horreur “effrayant”. De même on aurait tendance à aborder la musique avec des descriptions affective: une chanson “romantique”, de la musique rock “énergique”, ou un blues “triste”.

Une raison en est que la mémoire humaine a tendance à relier des événements avec des significations affective similaires et de rappeler des événements ayant un sens affectif fort avec plus de détails (Cahill and McGaugh [1995], McGaugh and Cahill [2002]).

L’extraction de l’information émotionnelle à partir du contenu d’un film, d’une image ou d’une chanson est, par conséquent, très pertinents pour l’indexation multimédia (Salway and Graham [2003], Chan and Jones [2005], Paleari and Huet [2008], Park and Lee [2008], Shan et al. [2009]). Les émotions représentent, peut être, une des composantes manquantes pour combler le fossé sémantique. En effet, les émotions sont des caractéristiques très abstraite sémantiquement mais sont aussi calculable par du traitement automatique.

En fait, la construction des systèmes de recherche en fonction ou aidé par des émotions nous permettra de rechercher exactement le type de médias que nous cherchons. Les systèmes de recommandation automatique pourraient apprendre nos préférences affectives et proposer la meilleure musique pour nos goûts généraux ou même adapter dynamiquement la musique à notre humeur.

Dans un futur assez proche les ordinateurs dans notre foyer pourrait donner le meilleur éclairage, et changez les plateaux à ceux que mieux s’adaptent à notre état d’esprit tout en remplissant l’environnement avec une musique douce.

Bien sur, dans tous ces scénarios les émotions doivent être utilisés en association avec d’autres informations sémantiques pour supporter l’indexation et contrôler l’environnement.

**Autres Scénarios.** Comme nous l’avons dit au début de ce chapitre les émotions sont fondamentales pour la prise de décision humaine, les communications, la mémoire, et bien d’autres fonctions cognitives. Il est donc clair que les ordinateurs peuvent prendre avantage des compétences émotionnelles de nombreuses façons différentes et que les scénarios possibles pour l’informatique affective ne sont limitées qu’à notre imagination.

Par exemple, l’informatique affective pourrait être utilisée pour aider les personnes autistes. Les gens affecté par l’autisme ont souvent du mal à comprendre les émotions des gens qui les entourent et, par conséquent, à réagir de notre façon naturelle. Ce fait provoque souvent des troubles sociaux pour les personnes autistes et aux personnes qui les entourent. Les ordinateurs peuvent alors aider ces gens en reconnaissant les émotions que les entourent ou en leur donnant de suggestions sur les réactions plus normales à ces émotions.

---

Un autre scénario pourrait être celle qui est généralement connu sous le nom “Affective Miroir” (Picard [1997]). Dans ce cas, les ordinateurs pourraient nous aider à préparer une réunion importante, un entretien, ou un spectacle.

Pour conclure, nous avons démontré que, dans de nombreux scénarios, comme ceux que nous avons passés en revue, nous aimerions que les ordinateurs soient capables de penser d’une manière plus proche des humains et de considérer donc les émotions comme un des paramètres. Dans certains autres scénarios les capacités affectives ne seront pas très pertinentes, par exemple nous n’auront probablement jamais besoin d’affecter les émotions d’une calculatrice. Enfin, il y aurait des scénarios dans lesquels on préférerait plutôt que les ordinateurs agissent de manière complètement rationnelle, ce qui pourrait être le cas de l’ordinateur de contrôle de l’état d’une centrale nucléaire.

Il est, néanmoins, la responsabilité des chercheurs en informatique affective de travailler pour trouver des solutions techniques pour permettre à l’ordinateur d’agir émotionnellement pour tous ces scénarios dans lesquels le calcul affectives seront nécessaires pour améliorer la naturalité, le plaisir et l’efficacité des interactions homme-machine.

### 1.3.1 Taches de l’informatique affective

En analysant ces scénarios, trois tâches principales peuvent être distingué (voir figure 1.2):

1. **Affichage des émotions:** l’aptitude à communiquer des émotions est au coeur d’un bon nombre de ces scénarios d’application. Pendant les communications naturelles les personnes utilisent les émotions pour renforcer les concepts et encourager des réactions. Si nous voulons créer des interactions homme-machine, aussi bien naturelles, agréables et efficaces, alors nous avons besoin de donner aux ordinateurs la capacité d’afficher les émotions d’une façon qui pourrait être perçu par l’homme comme naturelle, agréable et efficace. Dans les communications humaines, les messages émotionnels sont transférés à travers plusieurs canaux: les expressions faciales, prosodie vocale, la posture, les gestes et la sémantique des mots ne sont que quelques-unes des nombreuses façons dont l’homme se sert naturellement pour communiquer le contenu émotionnel. Il existe deux secteurs dans lesquels ces capacités sont déjà exploitées: il s’agit de ceux relatifs aux industries du cinéma et des jeux vidéo. Dans ces deux domaines la puissance graphique des ordinateurs est utilisé pour afficher des personnages qui avec leur vraisemblance font le succès, ou l’échec, d’une production.
2. **Reconnaissance des émotions:** l’aptitude à reconnaître les émotions semble être au centre de la plupart des scénarios que nous avons présentées aussi bien que l’aptitude à les afficher. Dans des applications telles que la télé-médecine et l’apprentissage par ordinateur, les médecins et les enseignants désirent avoir des notions par rapport l’état émotif des utilisateurs qui interagissent avec les machines. Par exemple, un professeur apprendra différemment en fonction du fait que l’étudiant soit sur le point de s’ennuyer ou que le même sujet se sente inspiré, et concentrée. De même, un médecin voudra garder sous contrôle l’humeur de ses patients pour prévoir des actions dans le cas où l’humeur soit trop mauvaise. Interagissant à distance via un ordinateur ne facilite pas la reconnaissance naturelle des émotions humaines; des techniques doivent alors être trouvées pour faciliter les communications de



l'homme avec son médecin ou professeur qui rendent accessible ces informations aux utilisateurs professionnels (dans ces cas les médecins et les professeurs).

3. **Synthèse et traitement des émotions:** les capacités de comprendre les émotions, avec ceux de les synthétiser, ou bien de "sentir" les émotions, toutes appartiennent à cette famille. Tout scénario impliquant l'intelligence artificielle (AI) comme le "pervasive-computing", les assistants personnels et intelligent, ou des jouets de pointe (par exemple les animaux de compagnie robotiques tels que le robot Aibo de Sony), a besoin de cette capacité à un certain niveau de complexité. Dans ces scénarios, il ne suffit pas de comprendre et d'afficher des émotions car les ordinateurs doivent aussi réagir avec pertinence aux émotions des utilisateurs et, parfois, commencer des comportements émotionnels afin de provoquer des réactions. En outre, comme Damasio et beaucoup d'autres chercheurs (Damasio [2005], Gmytrasiewicz and Lisetti [2002]) montrent, les mêmes concepts de l'intelligence et la rationalité ne peut être complètement séparée de celle des émotions et il est donc impossible de simuler une chose sans aussi créer l'autre.

## 1.4 Contributions de la Thèse

Dans cette thèse nous avons étudié le thème de l'informatique affective avec une approche scientifique et inquisitive. Notre travail est divisé en trois parties distinctes relatives aux trois sujets différents de l'informatique affective que nous avons expliqué dans la section précédente (à savoir *l'affichage des émotions, la reconnaissance des émotions et la synthèse de l'émotion*).

Le premier chapitre de ce document traite l'affichage d'émotion et en particulier la production d'expressions faciales émotionnelles. Dans cette partie, nous mettons l'accent sur les théories qui sous-tendent la génération des émotions par les humains (en figure 1.3 nous pouvons observer un exemple du processus de génération d'expression faciale d'après Scherer [1984]) et essayer de reproduire le processus naturel de la génération de l'expression faciale d'un agent d'interface graphique 3D de Hapték et une plate-forme robotique appelé iCat qui a été développé par le laboratoires de recherche de Philips au Pays-Bas.

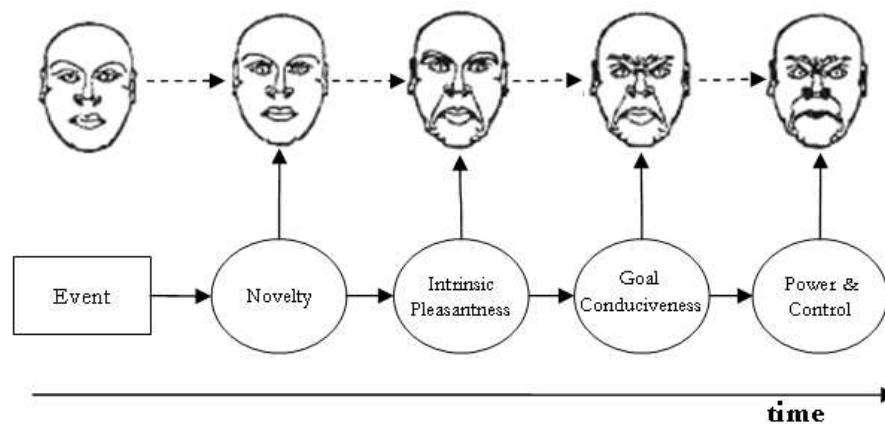


Figure 1.3: Le processus de génération d'une expression faciale (*rage*) d'après le CPT de Scherer

Le deuxième chapitre traite de la question de la reconnaissance des émotions. Pour résoudre cette tâche, nous mettons l'accent sur l'audio et la vidéo. Un ensemble de scénarios a été définie, afin de permettre à l'ordinateur d'extraire des informations relatives aux émotions par ces deux modalités. Dans ce cas, nous utilisons une base de données audio-vidéo multimodale que est librement disponibles sur Internet et nous appliquons des techniques différentes avec la contrainte de travailler en temps-réel ou près du temps-réel avec notre équipement standard.

Le troisième chapitre de ce document de synthèse aborde l'émotion et les liens possibles entre l'intelligence artificielle et les émotions. Cette partie représente une extension des travaux entrepris pendant mes études de maîtrise (Paleari [2005]). ALICIA est un agent d'interface base sur le concept du BDI (Belief, Desires, Intentions) et capable d'évaluer les événements entourant avec un processus d'évaluation similaire à celle décrit par Scherer [1984].

Ici, il suit une liste des principales contributions de notre recherche pour chacun de ces trois parties de la thèse:

#### 1.4.1 Partie 1: Affichage d'émotion

1. Génération de expressions faciales crédibles basé sur des théories psychologiques: dans la première partie de mon doctorat nous avons mis l'accent sur la génération des expressions faciales crédibles pour un visage en 3D (à savoir un avatar de Hapték comme celui montré en figures 1.4–1.8) et une plate-forme robotique (à savoir le robot iCat par Philips en figure ). Les expressions faciales ont été générées en basant sur les travaux de Scherer [1984] et en particulier de sa théorie du "appraisal" d'émotions par un processus à composantes (Paleari and Lisetti [2006a,b,c], Grizard et al. [2006], Paleari et al. [2007a]).

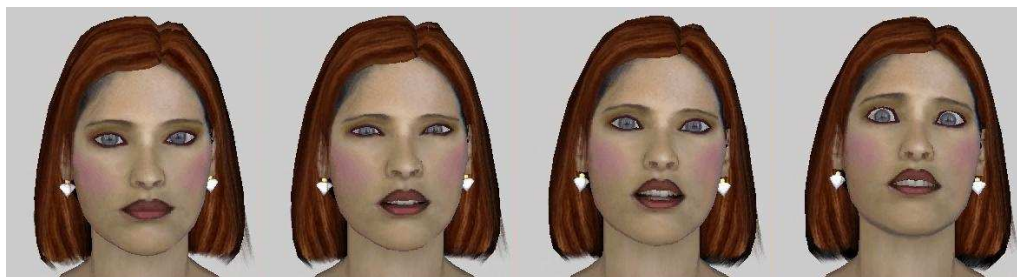


Figure 1.4: Une possible évolution de l'expression de peur avec Hapték

2. Génération d'une liste de questions pour les psychologues: pendant la génération des expressions faciales nous avons observé que certains détails sur la théorie des émotions par Scherer étaient manquantes ou peu claires. Nous avons donc crée une liste de questions pour les psychologues et nous avons été invité à l'Université de Genève pour discuter de ces questions avec Scherer et son équipe (Paleari et al. [2007a]).

#### 1.4.2 Partie 2: Reconnaissance d'émotion

1. Design de SAMMI, Semantic Affect-enhanced MultiMedia Indexing (figure 1.10): nous présentons un cadre destiné à l'indexation automatique des films et d'autres



Figure 1.5: Une possible évolution de l'expression de tristesse avec Hapték



Figure 1.6: Une possible évolution de l'expression de dégoût avec Hapték



Figure 1.7: Une possible évolution de l'expression de rage avec Hapték



Figure 1.8: Une possible évolution de l'expression de joie avec Hapték



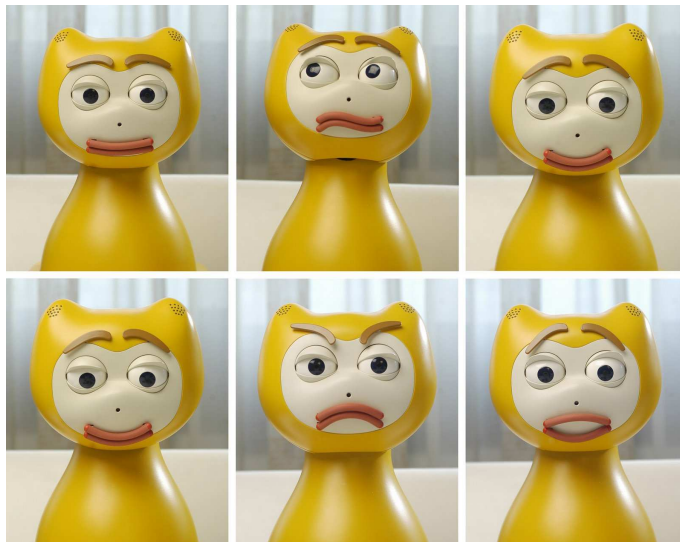


Figure 1.9: Expressions faciales par default sur l'iCat de Philips

éléments multimédia via les émotions contenues dans les médias même et interprété par les acteurs (Paleari et al. [2007b,a], Paleari and Huet [2008]).

2. Design de AMMAF, A Multimodal Multilayer Affect Fusion paradigm (figure 1.11): AMMAF est un système complexe de fusion multimodale pour la reconnaissance des émotions que essaye de résoudre la question de la synchronisation multimodale et que est capable d'aborder, au même temps, les différents phénomènes affectifs comme les émotions, l'humeur, et des traits de personnalité (Paleari and Lisetti [2006c]).
3. Design et développement de ARAVER, Automatic Real-Time Audio-Video Emotion Recognition (figure 1.12): ARAVER est un système d'interprétation et reconnaissance automatique des émotions de l'utilisateur user-indépendant qui analyse l'expression faciale et la prosodie vocale (Paleari et al. [2007a], Paleari and Huet [2008], Paleari et al. [2009a]). En particulier, les contributions principales de ce système sont:
  - système de détection et suivi automatique en temps-réel de point clef du visage indépendant de la personne;
  - études approfondies sur les caractéristiques et les vecteurs de caractéristiques pour la reconnaissance audio-visuelle d'émotion (Paleari and Huet [2008]).
  - études approfondies des classificateurs et des techniques de pré et post traitement pour la reconnaissance des émotions (Paleari and Huet [2008], Paleari et al. [2009a]);

### 1.4.3 Partie 3: Synthèse d'émotion

1. amélioration de VALERIE, Virtual Agent for Learning Environment Reacting and Interacting Emotionally:

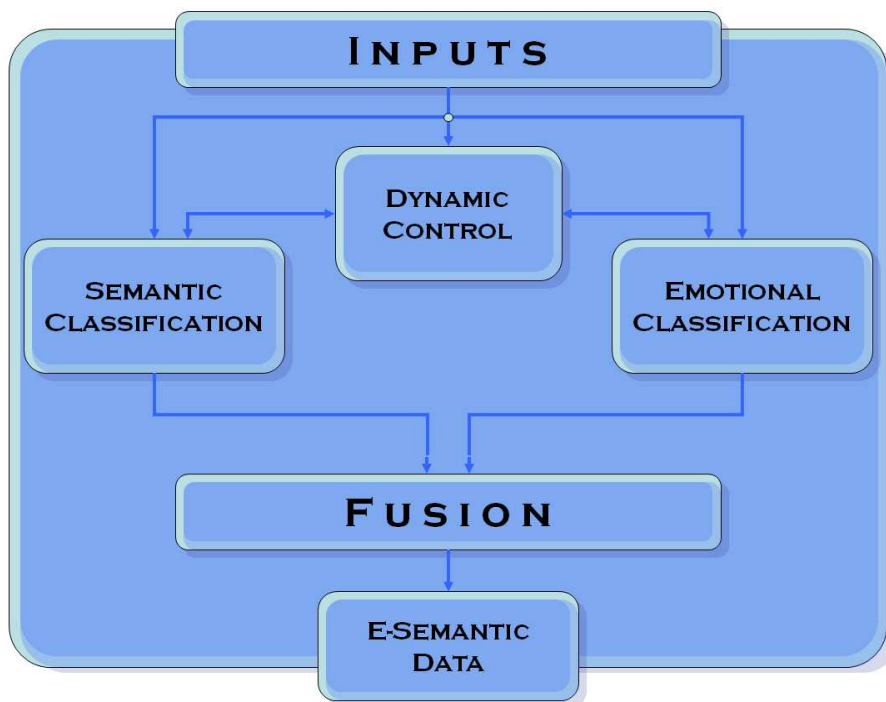


Figure 1.10: L'architecture de SAMMI

VALERIE (figure 1.13) est un développement "proof-of-concept" d'une simple intelligence artificielle pour un système de e-apprentissage qui fait usage de simples techniques d'estimation d'émotion et de réactions émotionnelles pour améliorer l'expérience de apprentissage (Paleari et al. [2005], Paleari and Lisetti [2006d]).

2. le développement d'ALICIA, agent d'intelligence artificielle utilisant les émotions: ALICIA (figure 1.14) est une complexe architecture d'intelligence artificielle capable de réagir émotionnellement à des stimuli environnementaux (Paleari et al. [2007a]). L'architecture de ALICIA est basée sur les théories psychologiques des émotions de Scherer [1984] et Leventhal and Scherer [1987], mais reprend également l'inspiration profonde à partir de Gratch and Marsella [2004].

En annexe à cette thèse, nous présentons deux autres contributions qui sont moins liées aux thèmes principaux de l'informatique affective sur lesquels nous avons travaillé:

#### 1.4.4 Reconnaissance biométrique

Typiquement les expressions faciales sont considérées comme l'un des problèmes dans le domaine de la reconnaissance faciale biométrique. Similairement, les caractéristiques faciales des différents sujets sont considérées comme un problème à surmonter pour la reconnaissance des expressions émotionnelles. En effet, il est beaucoup plus facile de comprendre l'expression d'un visage d'une personne connue qu'effectuer la reconnaissance des émotions d'un sujet inconnu dont on possède peu d'informations si non qui il appartient au genre humaine.

Dans ce travail, nous démontrons que la dynamique même des expressions faciales que nous avons utilisée pour reconnaître l'émotion est une caractéristique biométrique (Paleari

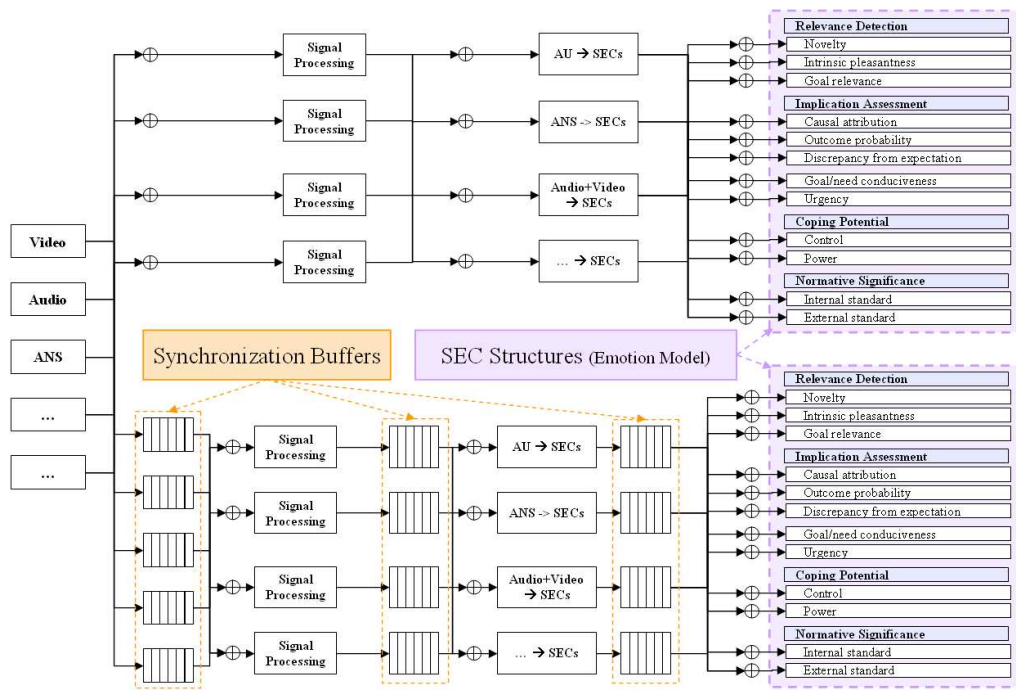


Figure 1.11: l'architecture de AMMAF

et al. [2009c]). En d'autres termes, la manière dont une personne exprime ses émotions par le biais des expressions du visage est unique et peut donc être utilisé pour la reconnaissance du sujet. De l'autre côté, la détection du sujet devrait être utile pour la tâche de reconnaissance des émotions automatique. À l'avenir, les deux systèmes vont travailler ensemble, le module de reconnaissance des émotions va aider le module de reconnaissance de la personne et vice versa.

### 1.4.5 Transcription automatique de la musique

La musique, comme le cinéma et toute autre forme d'art, sont parmi les plus pures formes de communication affective. Lors de l'exécution d'un morceau de musique, les artistes presque imperceptiblement modifier les tempos et la hauteur des notes au fin de communiquer une certaine émotion. Au même temps, le visage de l'artiste va aussi montrer l'émotion. Afin d'effectuer la reconnaissance des émotions d'un morceau de musique, deux systèmes pourraient être en interaction: un premier système analysera l'expression du visage de l'artiste (cela pourrait se faire grâce à notre système "ARAVÉR" de reconnaissance de l'émotion (voir la section 5 et Paleari et al. [2007a, 2009a], Paleari and Huet [2008], Paleari et al. [2009b]); un deuxième système détecte d'abord la transcription du morceau joué, reconnaît le morceau et compare ensuite la transcription détecté avec la base pour en extraire un autre information émotionnel.

Dans ce travail nous présentons une première étape vers la génération automatique des relevés de notes de musique avec un approche bimodale audio-visuel (Paleari et al. [2008a,b]). En figure 1.15 est possible observer l'interface de notre système de transcription automatique.

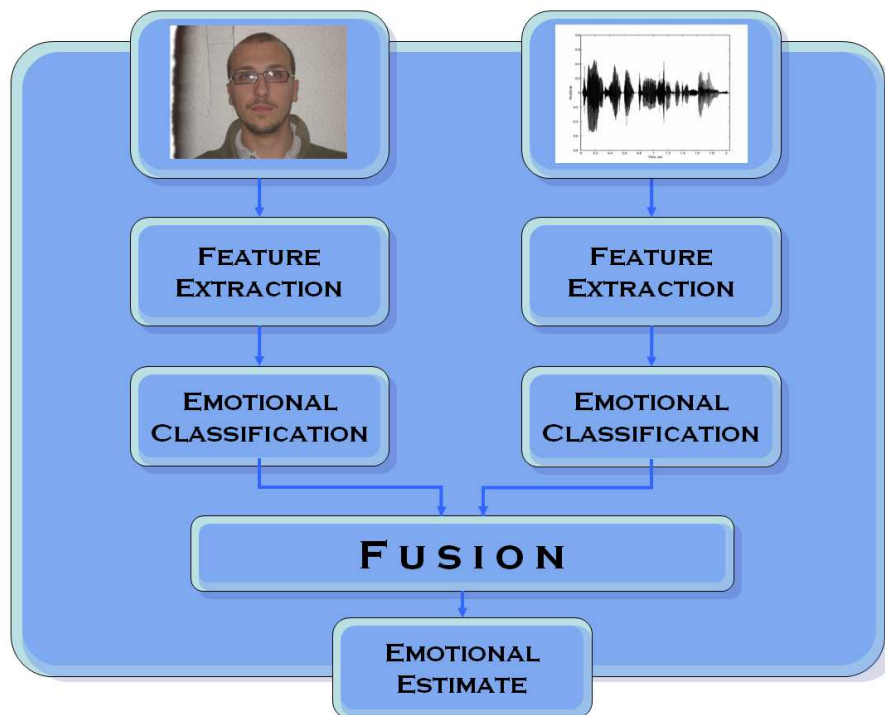


Figure 1.12: Architecture de ARAVER

## 1.5 Contenu du document

Dans les sections suivantes, nous présentons le sujet de l'informatique affective (section 3.1), nous montrons quelques-uns des scénarios d'application possibles (chapitre 3.2), et nous décrivons la nature des phénomènes affectifs travers quelques-unes des théories psychologiques les plus pertinentes (section 3.3). Le reste de cette thèse est organisé en trois parties: la chapitre 4 décrit le sujet de l'affichage d'émotions, la chapitre 5 traite le sujet de la reconnaissance d'émotion, et la chapitre 6 se concentre sur le thème du traitement et de la synthèse d'émotions et d'autre phenomènes affectif.

La chapitre 4 est en outre composé de quatre chapitres principaux: Le sectione 4.1 présente un aperçu général sur le sujet de l'affichage d'émotion; Le sectione 4.2 présente quelques théories sur l'expression émotionnelle et une partie relevant de l'état de l'art et des technologies de pointe les plus pertinents pour l'affichage émotionnel; Le chapitre 4.3 présente nos travaux de recherche sur ce sujet et en particulier détail le processus de génération d'expressions faciales; Enfin, les chapitres 4.4 et 4.5 présentent et discutent les résultats obtenus par les expressions du visage que nous avons développé.

La chapitre 5 est elle aussi composé de cinq chapitres principaux: Le sectione 5.1 présente un aperçu général sur le sujet de la reconnaissance des émotions; Le sectione 5.2 décrit l'état de l'art le plus pertinent aussi bien que certaines des techniques de pointe pour la reconnaissance des émotions; Le sectione 5.3 examine en détail les différentes parties du système temps-réel pour la reconnaissance multimodale d'émotion que nous avons nommé ARAVER, et du cadre pour la fusion multimodale appelée AMMAF; Le sectione 5.4 présente les résultats obtenus avec ARAVER. Enfin, le sectione 5.5 présente quelques perspectives pour ce domaine de recherche en décrivant SAMMI, un cadre d'indexation destiné au multimédia que utilise les émotions et le contenu sémantique

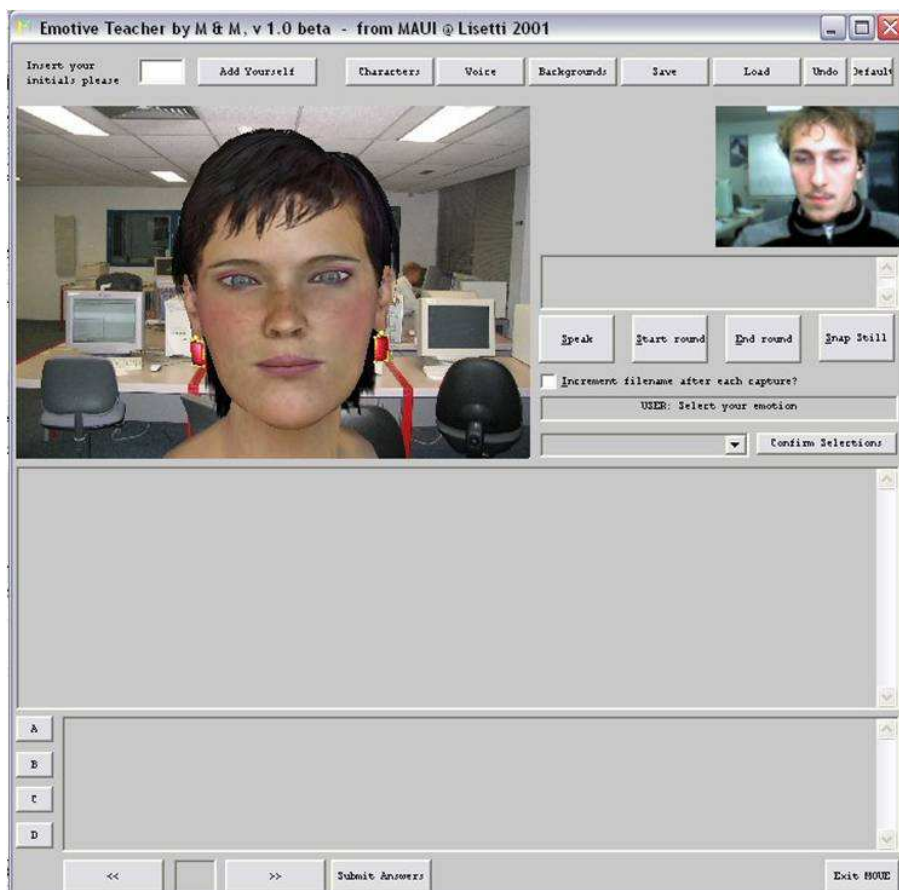


Figure 1.13: L'interface de VALERIE

des différents médias pour l'indexation et la recherche du contenu.

Le chapitre 6 est aussi divisé en quatre chapitres: La section 6.1 introduit le sujet de la synthèse de l'émotion; La section 6.2 aperçoit quelques théories d'émotions en ce qui concerne la synthèse émotionnelle et présente quelques travaux existants dans ce domaine; Le chapitre 6.3 présente l'architecture d'ALICIA et montre quelques scénarios que nous avons développés pour la tester; Enfin, la section 6.4 conclut le chapitre et discute les résultats préliminaires obtenus avec ALICIA.

Ces trois parties sont suivies de quelques conclusions générales (section 7) et par des annexes différentes. En particulier, B présente un travail préliminaire sur un système de détection et suivi des points clés du visage que nous avons étudié. L'annexe C présente une approche biométrique à la reconnaissance de personnes qui utilise les mêmes techniques utilisées pour la reconnaissance des émotions et qui démontre que l'aspect dynamique des expressions faciales peut être utilisé pour la reconnaissance d'utilisateur. Enfin, D présente un travail sur la transcription automatique de musique.

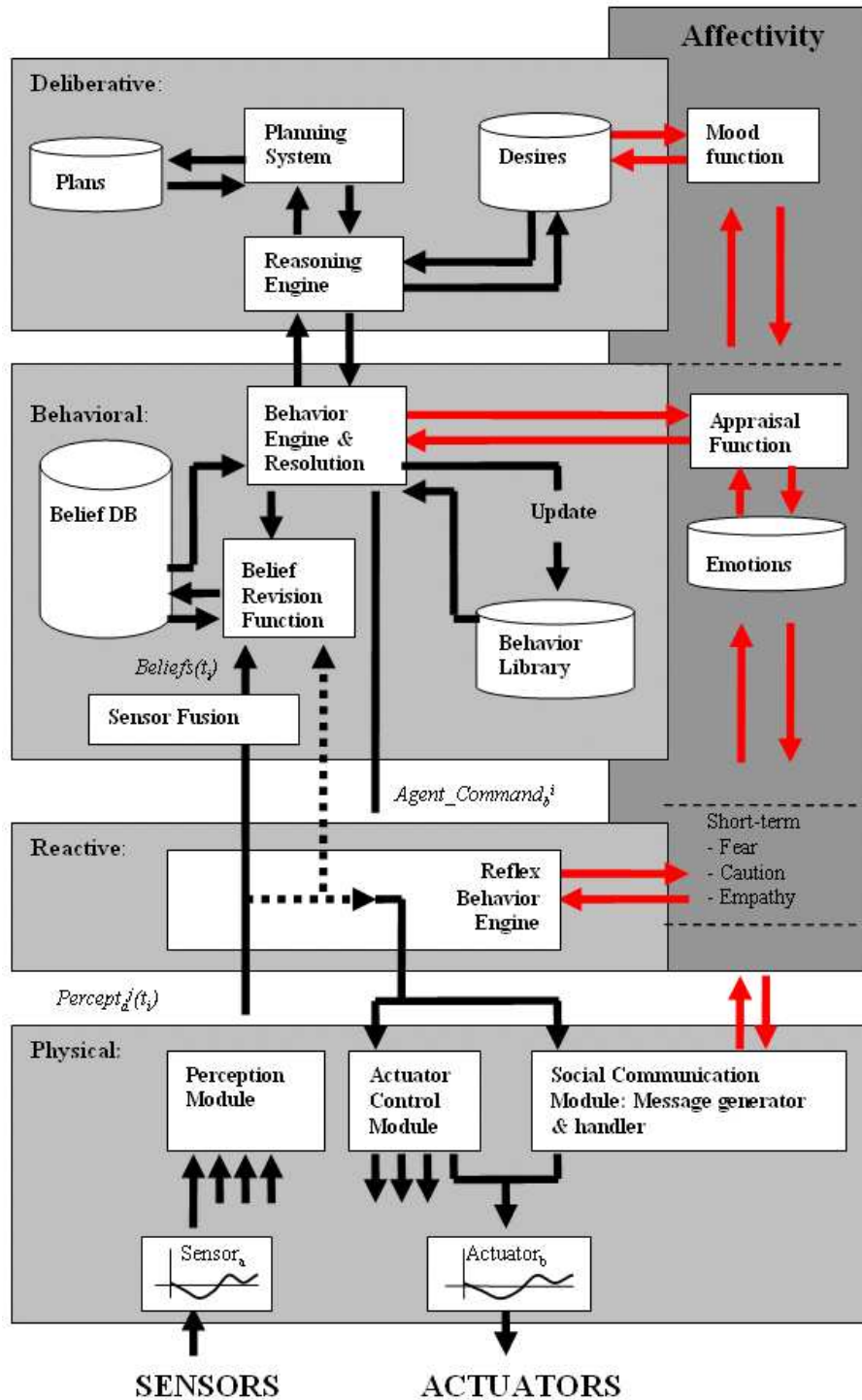


Figure 1.14: L'architecture de ALICIA et ces trois niveaux





Figure 1.15: L'interface du système de transcription automatique





## Chapter 2

# General Introduction

Computers and other electronic devices are becoming more and more common in our everyday lives. Improving the quality of the computer-mediated communications and human-computer interactions become, therefore, a very relevant subject of research.

One of the most hottest topics in the domain of human-computer interaction and generally human-centered-computing is now affective computing, an interesting research topic at the edges of computer science, psychology, and cognitive science. The reasons for the computer to process, show, and understand emotions are numerous (Picard [1997]).

### 2.1 Affective Computing

*Affective computing is computing that relates to, arises from, or deliberately influences emotions. [...] Affective computing also includes many other things, such as giving a computer the ability to recognize and express emotions, developing its ability to respond intelligently to human emotion, and enabling it to regulate and utilize its emotions.*

This is Picard's original definition of affective computing from the "Affective Computing" book (Picard [1997]).

The reasons motivating affective computing are manifolds.

Researchers have demonstrated that during everyday human-computer interactions people tend to assume that computer have emotional-like abilities (Covey [2004]). Reeves and Nass [1998] showed that we often interact with computers as if they were capable of understanding emotions. Koda and Maes [1996] demonstrated that in some domains people prefer to interact with machines capable of displaying some sort of emotional capabilities. Interestingly, Koda and Maes demonstrate that also people who affirmed that computer should not display emotions finally prefer interactions with emotional agents. Covey [2004] goes beyond these results arguing that in some way humans "need" to believe that any intelligent machine have some sort of emotional-like behavior. It is interesting to notice that the more the machine looks human (i.e. robot with respect to avatar or text interface) the more people perceives the machine's behavior as intelligent and emotional.

Secondly, many researchers demonstrated that emotions held a strong influence on many cognitive functions such as: decision making (Damasio [2005]), perception (Halberstadt et al. [1995]) and its interpretation (Bouhuys et al. [1995]), communications (Mehrabian and Wiener [1967]), and many others (Hudlicka and Fellous [1996], Lisetti and Gmy-

---

trasiewicz [2002]). Goleman [2006] argues that emotional intelligence (i.e. the set of all the skills related to emotions) is generally more correlated than the intelligence quotient (IQ) to the “*success in life*”.

All these reasons motivate the need of computers able to recognize, process, and display emotions in order to improve naturalness, pleasantness, and effectiveness of human–computer–interactions, computer–mediated communications, and human–centered computing in general.

In particular, we argue that in many situations software should be designed keeping in mind that the human shall always be put at the middle of a loop in which the interaction with the computer or robot is “coupled with” and “mediated by” the emotions as in figure 2.1. These consideration shall, therefore, be kept in mind at all the phases of design of a computer agent, from the very first sketch–up to the final implementation.

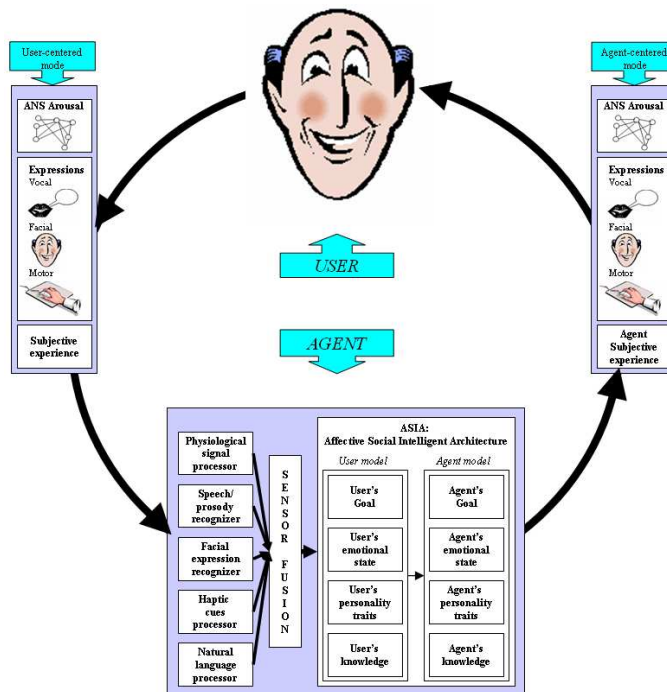


Figure 2.1: Human Centered Multimodal Affective User Interface. (adapted from Lisetti and Nasoz [2002])

## 2.2 Emotions

Emotions are studied in several different domains. Psychologists, physiologists, biologists, anthropologists, sociologists, philosophers and even ethologists study emotions (Scherer and Ekman [1984]).

The definition of “emotion” is therefore not clear, however the research community seems to agree on the fact that emotions are psycho–physical phenomena, that have origin in reactions to events and are motivated by the Darwinian evolution theories like any other physical characteristics of living beings (Darwin [1872]).

In other words emotions arise as a process of appraisal of the events surrounding us and have two different components: a first one is *cognitive* (it takes place in our brain and

influences the way we think), and a second one is *physical* (it has still origin in the brain but influences our body, the way we act, the way we perceive, etc.).

During the long years of research on emotions researchers have defined many different systems of classification of these phenomena. Three main families can be delineated; these are:

- **discrete emotional categories:** emotions are characterized by a keyword or a label identifying a family of emotions (e.g. happiness);
- **dimensional models of emotions:** emotions are characterized by a position in a multidimensional space in which the axis represent basic emotional concepts such as *valence* or *pleasantness* (e.g. happiness = [0.8, 0.7] is an emotion characterized by having positive pleasantness and positive valence);
- **componential models of emotions:** emotions are characterized by a set of components representing the phase of appraisal of the emotions themselves (e.g. happiness is the emotion arising from an unexpected event which is pleasant and facilitate one's goals).

## 2.3 Application Scenarios and Emotion Related Tasks

Prototypical examples of the use of emotions in computer science are tele-medicine, e-learning, gaming, indexing & retrieval, general artificial intelligence, instant messaging systems, and other form of tele-communication.

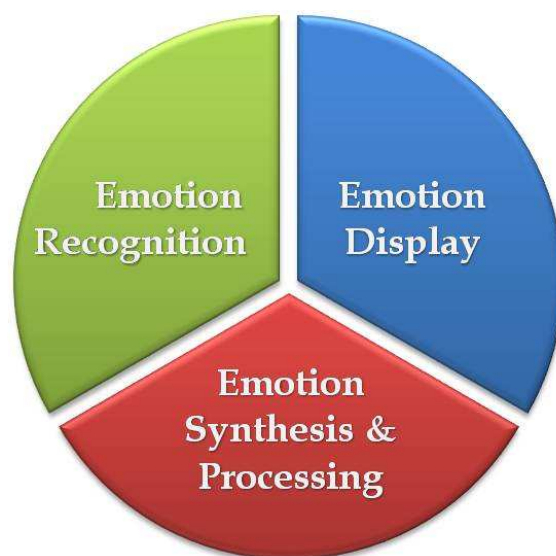


Figure 2.2: Affective computing and its three components

When analyzing these scenarios, three main tasks emerge (see figure 2.2):

1. **Emotion display:** the ability to communicate emotions is central to many of these application scenarios. In natural communications people use emotions to reinforce concepts and encourage reactions; if we want human-computer interactions, and computer-mediated communications to be as much natural, pleasant, and effective,

than we also need computers to be able to display affective states in ways that could be perceived by humans as natural. In human communications, emotional messages are transferred via several channels: facial expressions, vocal prosody, body posture, gestures, and the semantics of the words are just some of the many ways humans naturally employ to transfer emotions to their peers. There exist two industries in which these capabilities are already exploited: these are the one related to computer graphics in the cinema and video-games industries.

2. **Emotion recognition:** the ability to recognize emotions is the dual to the ability of displaying them and, as much as this latter ability, seems to be central to most of the scenarios we presented. In applications such as tele-medicine and e-learning, doctors and teachers may want to know how the users they are interacting with are feeling. For example a professor will teach differently depending on the fact that the student feels bored and un-excited or hopeful, inspired, and concentrated. Similarly, a doctor may want to keep under control the mood of his/her patients. Interacting via a computer from remote locations does not facilitate natural human emotion recognition; other computer based techniques need to be found to facilitate human natural communications.
3. **Emotion synthesis and processing:** the abilities to understand and process emotions, with the ones to synthesize, “think”, or “feel” emotions, all belong to this family. Any scenario involving advanced artificial intelligence (AI) like pervasive computing, personal AI assistants, or advanced toys (e.g. virtual pets like the SONY AIBO robot), needs this ability at a certain level of complexity. In these scenarios it is not enough to understand and display emotions, and computers also need to react with pertinence to user emotions and sometimes to initiate emotional behaviors to encourage reactions. Furthermore, as Damasio and many other researchers show, the same concepts of intelligence and rationality cannot be completely separated to the one of emotions.

## 2.4 Thesis Contributions

In this thesis we have investigated the topic of affective computing with an inquisitive computer-science approach. Our work is split in three different parts related to the three different subjects of affective computing we outlined in the previous section (i.e. *emotion display*, *emotion recognition*, and *emotion synthesis*).

The first part of this document discusses emotion display and in particular emotional facial expression generation. In this part, we have focus on the theories underlying the generation of emotions on humans and try to replicate the natural process of generation of facial expression on both a three-dimensional computer graphic avatar by Haptik and on the robotic platform iCat by Philip research labs.

The second part discusses the topic of emotion recognition. To solve this task we focus on audio and video. A set of scenarios has been defined which would allow the computer to extract information from those two modalities. In this case, we use a freely available multimodal audio-video database and apply different techniques with the only processing constraint of working in real-time or near-to-real-time with standard off-the-shelf equipment.

---

The third part of this document addresses emotion synthesis and possible links between artificial intelligence and emotions. This part represents an extension of the works initiated during my Master's studies (Paleari [2005]). ALICIA is an all-purpose BDI agent capable of evaluating surrounding event with an appraisal process similar to the one described by Scherer [1984].

Here it follows a list of the main research contributions for each of these three parts:

- **Emotion display:**

1. Generation of believable, psychologically based, facial expressions:  
in the first part of my Ph.D. we have been focusing on the generation of believable facial expressions for a 3D computer speaking face (i.e. an avatar by Haptik) and a robotic platform (i.e. the iCat robot by Philips). Facial expressions were generated taking advantage of Scherer [1984] component process theory of emotions (Paleari and Lisetti [2006a,b,c], Grizard et al. [2006], Paleari et al. [2007a]).
2. Generation of a list of questions for the psychologists:  
while generating the facial expressions we have observed that few details about the theory of emotions by Scherer were missing or unclear. We have therefore redacted a list of questions for the psychologist and we have been invited to the University of Geneva to discuss these issues with Scherer and his team (Paleari et al. [2007a]).

- **Emotion recognition:**

1. Design of SAMMI, Semantic Affect-enhanced MultiMedia Indexing:  
we present a framework designed to automatically index movies and other multimedia items via the emotions contained in the media and portrayed by the actors (Paleari et al. [2007b,a], Paleari and Huet [2008]).
2. Design of AMMAF, A Multimodal Multilayer Affect Fusion paradigm:  
AMMAF is a complex multimodal fusion system for emotion recognition solving the issue of multimodal synchronization and being capable of addressing, at the same time, different affective phenomena such as emotions, mood, and personality traits (Paleari and Lisetti [2006c]).
3. Design and development of ARAVER: Automatic Real-Time Audio-Video Emotion Recognition:  
ARAVER is a system performing automatic user-independent emotion recognition from facial expression and vocal prosody which has been designed for SAMMI (Paleari et al. [2007a], Paleari and Huet [2008], Paleari et al. [2009a]). In particular, the main contributions of this system are:
  - Automatic real-time person independent feature point detector and tracker;
  - Extensive studies on features and features vectors for audio-visual emotion recognition (Paleari and Huet [2008]).
  - Extensive studies on classifiers, pre-, and post-processing techniques for emotion recognition (Paleari and Huet [2008], Paleari et al. [2009a]);

- **Emotion synthesis:**

---

1. Improvement of VALERIE, Virtual Agent for Learning Environment Reacting and Interacting Emotionally:  
VALERIE is a proof-of-concept development of a simple artificial intelligence tutoring system which makes use of simple emotion estimation techniques and emotional feedback to improve the tutoring experience (Paleari et al. [2005], Paleari and Lisetti [2006d]).
2. Development of ALICIA affect-enhanced artificial intelligence agent:  
ALICIA is a complex general purpose artificial intelligence architecture able to react emotionally to environmental stimuli (Paleari et al. [2007a]). ALICIA's architecture is based on psychological theories of emotions of Scherer [1984] and Leventhal and Scherer [1987] but also takes deep inspiration from Gratch and Marsella [2004].

In annex to this thesis we present two other contributions which are less related to the main topics of affective computing to which we have been working on. These are:

- **Biometric People Recognition:** Typically facial expressions are considered as one of the disturbances for biometric face recognition. Similarly, the facial characteristics of different subjects are considered as an issue to overcome for emotional expression recognition. Indeed, it is much easier to understand the facial expression of a known person while it could be hard, even for a human, to perform emotion recognition from a short set of data of a unknown subject.  
In this work, we demonstrate that the very same dynamics of facial expressions that we have used for emotion recognition are also a biometric characteristic (Paleari et al. [2009c]). In other words, the way a person expresses emotions via facial expressions is unique and can, therefore, be used for the recognition of the subject. From the other side, detecting which subject is currently speaking should be helpful for the task of automatic emotion recognition. In the future, the two systems will work together, the emotion recognition module helping the person recognition one and vice versa.
- **Automatic Music Transcription:** Music, together with cinema and art, is among the purest form of emotional communications. When performing a piece of music, the artists slightly and almost imperceptibly modify tempos and pitches of the notes to communicate a certain emotion. At the same time, his/her face will show this emotion. In order to perform emotion recognition of a piece of music, two systems could be interacting: a first system analyzes the facial expression of the artist (this could be done thanks to our ARAVER emotion recognition system (see section 5 and Paleari et al. [2007a, 2009a], Paleari and Huet [2008], Paleari et al. [2009b])); a second system first detects the transcription of the played piece of music and then compares it with the baseline to extract another emotional meaning.  
In this work we present a first step toward the automatic generation of music transcript via audio-visual input (Paleari et al. [2008a,b]).

## 2.5 Outline

In the following chapters we present the topic of affective computing (section 3.1), overview some of the possible application scenarios (section 3.2), and describe the nature of the

---



emotional phenomena through some of the most relevant physio–psychological theories (section 3.3). The rest of this thesis is organized in three parts: chapter 4 describes the topic of affect display, chapter 5 regards the topic of affect recognition, and chapter 6 focuses on the topic of affect processing and synthesis.

Chapter 4 is further composed of four main chapters: section 4.1 shortly presents a general overview on the topic of emotion display; section 4.2 presents some theories of emotional expressions and some of the most relevant state–of–the–art technologies for emotional display; section 4.3 presents our research work on this topic and details the process of generation of facial expressions; finally, chapters 4.4 and 4.5 present and discuss the results obtained by the facial expression we developed.

Chapter 5 is composed of five main chapters: section 5.1 shortly presents a general overview to the topic of emotion recognition; section 5.2 overviews the most relevant state–of–the–art techniques for emotion recognition; section 5.3 discusses in detail the different parts of the real–time multimodal emotion recognition system called ARAVER, and the framework for multimodal fusion called AMMAF; section 5.4 presents the results obtained with ARAVER. finally, section 5.5 presents some perspectives for this research domain by describing SAMMI, a framework designed to index multimedia excerpts with emotions and other semantic content–based tags.

Chapter 6 is divided in four chapters: section 6.1 introduces the topic of emotion synthesis; section 6.2 overviews some theories of emotions regarding emotional synthesis and presents few existing work in this domain; section 6.3 presents the architecture of ALICIA and shows some simple scenarios we have developed for testing it; finally section 6.4 concludes the part and discusses the preliminary results obtained with ALICIA.

These three parts are followed by some general conclusions (section 7) and by different annexes. In particular, annex B presents a preliminary work on an alternative feature point detector and tracker. Annex C presents a biometric approach to people recognition which use the same techniques used for emotion recognition. Finally, annex D presents a work on automatic music transcription.

---





## Chapter 3

# Psychological Background

### 3.1 Affective Computing

Affective computing is a branch of artificial intelligence that deals with the design of systems and devices that can recognize, interpret, process, and/or display human emotions. It is an interdisciplinary field spanning computer sciences, psychology, and cognitive science.

While the origins of the field may be traced as far back as to early philosophical inquiries into emotion in the eighties (Ekman [1971], Izard [1977], Plutchik [1980], Norman [1981], Scherer [1982], Leventhal [1984], Frijda [1986], Ortony et al. [1988]), the modern branch of “*computing that relates to, arises from, or deliberately influences emotions*” originated in 1997 with Rosalind Picard [1997] “Affective Computing” book.

In the last twelve years, since the introduction of the term “affective computing”, the research community in this field has grown rapidly along the three main axis of emotion recognition (e.g. computer vision and indexing), emotion display (e.g. affective text-to-speech and 3D characters), and emotion processing/synthesis (e.g. artificial intelligence and human-computer-interactions).

At a first sight, one might question the importance of affective computing and evaluate this domain as a curious and interesting yet useless one. This is far from being the truth. In the following sections we will discuss the reasons motivating affective computing referencing works of psychologists, neurologists, and anthropologists showing the important links between emotions and many cognitive functions.

#### 3.1.1 Emotion and the Human Brain

For centuries we have been believing there was a strong separation between rationality and emotions. In particular, we believed that reasoning and decision making have better quality when they are not influenced by emotions. Indeed, according to the classical view, emotional decision implies acting irrationally and with poor judgment (Schafer et al. [1940]). Again, this is not the full truth about it (Simon [1967], Davidson et al. [2002]).

##### 3.1.1.1 Emotion and Perception

At a neurobiological level emotions and rationality reside in two different regions of the brain but these are very much linked to each other (Cytowic [1989], Hudlicka and Fellous

---

[1996]). In particular, perception is very much linked to the limbic system (the system linked to emotions) (Mayer and Salovey [1993], Niedenthal and Kitayama [1994]). Emotions have demonstrated to influence the way we hear and interpret words (Halberstadt et al. [1995]), the way we interpret what we see (Bouhuys et al. [1995]), and in general, the way we think, make plans, learn, etc. (Izard [1993], Hudlicka and Fellous [1996], Lisetti and Gmytrasiewicz [2002]).

The studies from LeDoux [1990, 1994] describe two different ways of perceiving of the human beings. The first way is “*quick and dirty*”, it is performed in the limbic system and it is, for example, the one that appraises danger when something big is suddenly approaching. This first part is the one responsible for the generation of fast psycho–physical responses such as trying to avoid the approaching object and which also generate basic “*sensory–motor*” emotions; in this case fear. The second has its origin in the brain cortex (usually linked to perception and high cognitive functions); it is more slow and precise and could, for example, recognize the object as an inflatable beach ball and appraise that the fearful response was not necessary. Goleman [2006] describes the phenomena linked with the first “*quick and dirty*” way of perceiving with the name of emotional “*perception hijacking*”.

### 3.1.1.2 Descartes Error

In 1994 the behavioral neurologist and neuroscientist Antonio Damasio demonstrated in his book “*Descartes Error*” how emotions can be fundamental for everyday human rational reasoning and decision making (Damasio [2005]). In his book Damasio present examples of patients affected by frontal–lobe disorders affecting a part of the cortex (linked to rational thinking) which communicates with the limbic system (the one in charge of emotions). Damasio’s patients appear to have normal intelligence, scoring average or above average on a variety of tests. At a first sight these people will resemble to Star Trek’s Mr. Spock character, unexpressive, very rational and otherwise very intelligent.

Damasio’s book describes how these people tend to make disastrous decisions, and have issues with relationships and social interactions, usually resulting in the loss of their jobs, friends, families, colleagues, etc. The one example which is usually brought to represent this state is the one of “*Elliot*”.

Damasio’s Elliot is an intelligent, skilled, and able–bodied man in his thirties which suffered damage to his frontal–lobe as a result of a brain tumor. Several experts have evaluated Elliot’s mental faculties and declared them intact. Their conclusion is that Elliot is at the very best lazy and at the worst a malingerer. Elliot seem a normal, cool, detached, and unperturbed person even when he is faced to potentially embarrassing discussions about personal events.

Elliot is, nevertheless, not able to maintain his job, friends, or family. On the job Elliot is not able to manage his time: he could spend all his time on a secondary task, loosing grip on the main one. For example he could spend all afternoon deliberating on the best principle of categorization for the papers he was sorting. Since all principles have advantages and disadvantages, he needs to rationally think about all the possible outcomes of all possible decisions. A certain amount of indecision is normal; an healthy person will, nevertheless, stop thinking after a while and select one principle out of the best ones acknowledging the small difference it will make on the long run. Elliot will continue evaluating the matter till one principle demonstrated to be the best.

Damasio also describes how Elliot is unable to make the link between bad or danger-

---

ous decisions and bad feelings, so he repeats making the same mistakes over and over instead of learning otherwise.

The book explains that all of these behavior are linked to something missing at the emotional level. Elliot lack of emotions severely handicaps his ability to function rationally and intelligently by impairing his ability to learn and to take fast decisions.

Nowadays artificial intelligence act pretty much as Damasio's Elliot: they have very powerful "brains" and can quite easily beat the most talented human being playing chess, as did Deep Blue with Garry Kasparov in May 1997, but fail miserably to deal with some of the simplest human everyday tasks (not considering the issues with computer vision of speech-to-text) being unable to take fast decisions in our un-constrained environment without considering "rationally" all the possible space of solutions.

The conclusion of Damasio, Picard, and many other researchers is that too few emotion (but also too much) impairs the ability to reason rationally in an "intelligent" manner. In other words we could say that there exist some form of "emotional intelligence" (Gardner [1983], Salovey and Mayer [1990]). Goleman [2006] argue that this form of "*emotional intelligence*" is more important than the usual measure known as *Intelligent Quotient* (IQ) in improving the chances of success in life.

To conclude, if too few emotion impairs the rational intelligence of humans, there is no reason why it should not also impair artificial intelligence. It seem obvious that emotions should be included into artificial intelligence if we want computers of the future to really act intelligently.

In one sentence from Marvin Minsky [1988], American cognitive scientist in the field of artificial intelligence (AI) and co-founder of MIT's AI laboratory, from the book "The Society of Minds":

*The question is not whether intelligent machines can have emotions, but whether machines can be intelligent without any emotions.*

### 3.1.2 Emotion and Communication

Human interactions are much more than a simple exchange of fundamental information through words (Merola and Poggi [2004], Besson et al. [2004]). Mehrabian and Wiener [1967], Mehrabian and Ferris [1967], Mehrabian [1971, 1972] suggest that in some form of communications as much as the 93% of the meaning of a sentence does not depend on the actual words which are exchanged but rather from paralanguage such as voice tone, facial expression, and body posture.

In recent years we are experiencing a gradual and continuous migration of human interactions from face-to-face to device-to-device communications. Indeed, in the modern era everybody relates several times a day with other people via cellphones, chats, and complex tele-presence systems. In computer-mediated communications the existing devices and the limitation of the communication links restrict the affordances available to the users in terms of physical movement (Nguyen and Canny [2007], Vertegaal et al. [2000], Sellen et al. [1992]), interaction, peripheral vision (Mark and DeFlorio [2001]), spatio-semantic integrity, and, therefore, information flow (Mark et al. [2003]).

One big limitation which is mostly not addressed by today systems is the one regarding the difficulties in transferring emotions. People have naturally tried to overcome this limitations when the emotional communication. As an example in instant messaging (IM) and in e-mails the communication channel is very restricted. When one writes

an email, he/she can only use language to express his/her feelings. Depending on how gifted he/she is as a writer the message can be more or less clear.

Quite soon IM dialogues and e-mails have been “augmented” with the use of “emoticons” (i.e. emotion–icons a.k.a. smileys) like :-) or :-( as a means of expressing emotions. Although emoticons are quite useful improving the expressivity of on–line communication, they still provide a very limited means of expressing emotion. Furthermore, it remains within the responsibility of the user to carefully prepare the affective content of the textual message (Handel and Herbsleb [2002], Isaacs et al. [2002]) and the use of this tool seem to be relegated to informal e-mails and IM.

On this theme Picard [1997] provides a suggestive question: “*How many of you have lost more than a day’s work trying to straighten out some confusion over an email note that was received with the wrong tone?*”. Picard states that usually a big part of the audience raise the hand up to this question.

Given today spread and importance of remote communication devices it seem obvious that researchers should find better ways to allow people to interact with each other. Emotions represent an important source of information in human natural communications. Therefore, the transfer of emotions should be facilitated to build more natural and pleasant remote communications. It is into the field of affective computing that such a research is held.

### 3.1.3 Emotion and Human–Computer–Interactions

In the last decade people have been spending more and more time in front of their computer for work, communication, or leisure. Computers have become part of our daily lives and are more and more accessible to people all around the globe. All this motivates the efforts that are currently done for building more and more natural human–computer–interactions (HCI).

Would not it be nice if computer could understand when you are tired of working and suggesting you make a pause? Would not it be nice if your desktop could tell you are sad and prompt a joke? Would not it be nice if your I–Pod was able to made you listen the music that best goes with your mood? Would not it be nice if your car could understand when you are sleepy and react before you have an accident?

These and many others are just few of the possible examples of human–computer–interaction where understanding the user emotional state is needed in order to improve the naturalness, the effectiveness, and the pleasantness of HCI.

It is normal, for human beings to interact emotionally, when we talk to other people but also when we interact with animals; for example we scorn our pets and we know they have understood us from their body language (Negroponte [1995]).

It has been demonstrated that, not only humans prefers to interact this way also with computer (Koda and Maes [1996]), but also that people naturally assume computers and other inanimated objects to have some form of emotional capability (Nass et al. [1994], Nass and S. [1994], Thorisson and Cassell [1996], Reeves and Nass [1998]).

Covey [2004] argue that emotional understanding and display are one human psychological need. In other words, not only people naturally interact with computers as if they were to possess some form of emotional intelligence but this behavior arise from a human psychological and insuppressible need.

Building computer with the ability to recognize and display affect will therefore not only, make the human–computer–interactions more appealing, but also more efficient

---

because we would be able to better focus on our main tasks.

This chapter defined affective computing and reviewed few of the main motivations of the existence of emotions in human being. Affective computing arises from the assumption that these motivations are valuable also for computer agents. In the next chapter we will present few possible application scenarios for affective computing which may help to understand the motivations that have just been presented.

## 3.2 Applications of Affective Computing

In the previous chapter we have described what affective computing is and why it is important for human–computer interactions, computer–mediated communications, and human–centered–computing. In this chapter we describe more extensively few prototypical affective computing scenarios. For other affective computing scenarios and examples please refer to the detailed work of Picard [1997].

### 3.2.1 e–Learning

Almost every school does not have enough time or budget to let an instructor sit down and help each single student as they struggle with a subject. This can be frustrating for students who find it hard to understand a certain topic. Providing a virtual personal assistant, Intelligent Tutoring Systems (ITS) can solve this problem. Each student can have a personal tutor referring to him to help him when he gets confused.

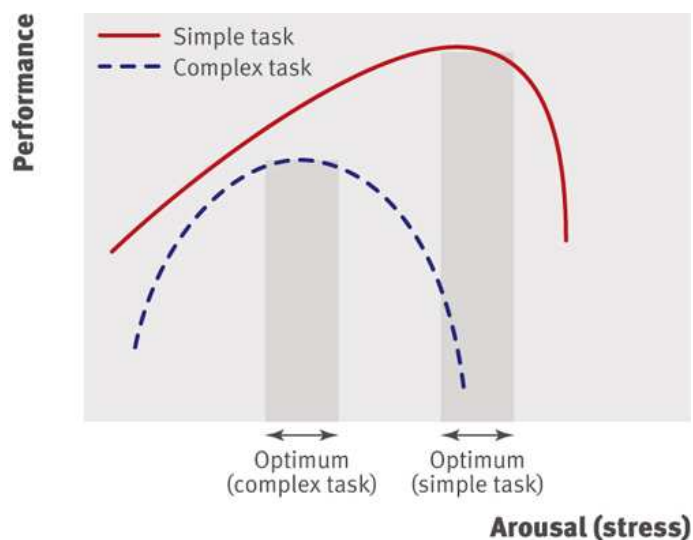


Figure 3.1: Yerkes–Dodson curves

Emotions are not always considered in ITS, but several researchers (Picard [1997], Lisetti and Gmytrasiewicz [2002]) argue that emotions are fundamental in communication and in human interactions in general. Few studies have applied emotions to learning environments (Kapoor et al. [2001], Kapoor and Picard [2005], Conati [2002], Conati and Zhou [2002]) demonstrating to increase pleasantness and effectiveness of the student learning experience.

It is well known that a carefully chosen level of stress usually improve general human performances (see figure 3.1) (Hebb [1966]) and in particular learning ones (Yerkes and Dodson [1908]). Monitoring user emotions, an ITS would be able to understand when a student is getting bored or confused and adapt its teaching style. ITS could also react to such a kind of emotional state, for example, explaining once again the subject using other words, qualitative reasoning or analogies. The same principles could be applied for different emotions. Kort et al. [2001] presented a work describing how the different emotions could be related to user learning phases as in figure 3.2.

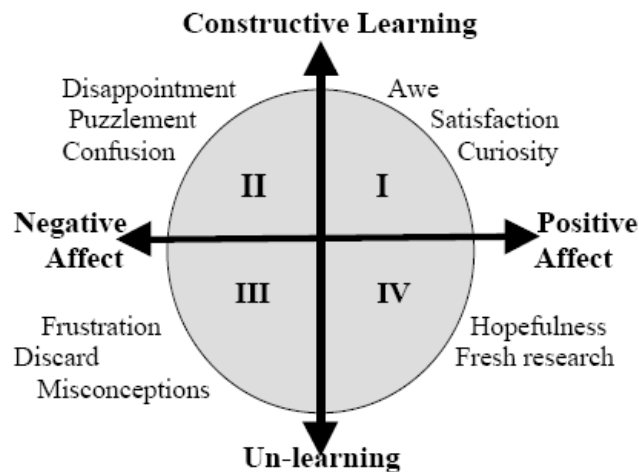


Figure 3.2: Impact of emotions on phases of user learning

If we were able to monitor the students' affective states we would also be able to assess if the student was learning or not. Therefore, the objective of an ITS should be to dynamically change style, speed, and methods in such a way that the student emotional state belong to the first quarter of figure 3.2. Anytime the detected emotion were to adrift to quarters II or IV then the ITS objective would be to take actions and bring the emotions back on the first quarter.

In tele-applications of e-learning, i.e. where the students interact with real professors, the information about students affective state could get lost due to the limitations intrinsic to computer-mediated communications. The teacher could, therefore, profit from the availability of the information about the affective state of the students by modifying is teaching style, making pauses, and generally reacting in a natural way possibly improving the effectiveness of his/her teaching.

### 3.2.2 Medicine and Tele-medicine

The history of Hunter Campbell Adams is known almost to everybody thanks to the Tom Shadyac's movie "Patch Adams" starring Robin Williams. Adams is a medical doctor who became famous for his unconventional approach to medicine. Convinced of the powerful connection between environment and wellness, he believes the health of an individual cannot be separated from the health of the family and community.

In particular he is also known to be among the founders of gelotology, e.g. the smile therapy, also known as clown therapy.



Indeed, laughter and happiness are well known to have positive effects on the human body. A positive affective state helps to combat stress and reduce pain by releasing the body's natural painkiller known as "*endorphins*" but it also has a positive effect on the cardiovascular and respiratory systems, relaxes muscles, and boosts the immune system by increasing the number of "*T-cells*" and lowering "*serum cortisol*" levels (Adams and McGuire [1986], Goodman [1992], Fry [1994], Martin [2002, 2004], Davidson et al. [2002], Lisetti and LeRouge [2004]).

We have already pointed out how positive emotions in general could help promote a positive outlook (Mayer and Salovey [1993], Niedenthal and Kitayama [1994]) helping people to cope with difficult situations. We have also showed that emotions facilitate creative and quick decision making (Damasio [2005]) helping people to prevent some unpleasant situations.

In tele-medicine, the doctor is interacting with patients at remote locations. While in face-to-face communications it is easy for the doctor to assess the emotional state of the patient and react accordingly, it might be hard to access the same information when communicating via a computer interface. Dedicated interfaces and emotion assessment software should provide the doctor with the needed information and allow him/her to react properly to negative affective states.

A particular tele-medicine scenario is the one regarding depression. Depression was 2004 third leading cause of disability in the world according to the 2008 World Health Organization's update of the "*Global Burden of Disease Report*" (Mathers et al. [1996, 2008]) and previsions are that by year 2020 depression will be second only to heart disease. Obviously, depression is very related to emotions and affect (j. Garlow and Nemeroff [2002]).

Assuming that our society aim at avoiding disabilities, that monitoring of the affective state could help preventing depression to outcome, and that not everybody can afford private psychological follows-up, then it seem obvious that affective computing for the prevention of the depression state is going to become a more and more central topic of research in the next few years.

### 3.2.3 Gaming

Gaming is soon going to be the leading business in the entertainment industry (see figure 3.3) directly employing more than 24,000 people in 2007 in just the US (Siwek [2007]). According to the NPD group data (Group [2009], W3stfa11 [2009]) the gaming industry touched more than 22 billion US dollars in retail sales in 2008 (see figures 3.4(a)).

Worldwide the estimates are more than doubled (54 billions US \$ according to Group [2009], see figure 3.4(b)).

Games are becoming more and more complex getting vivid graphics, artificially intelligent opponents, responsive action, realistic force feedback joysticks, and immersive displays. Affect plays a key role in the user experience, both in games developed purely for entertainment purposes, and in "serious" games developed for education, training, therapy or rehabilitation.

Affective computing already influences state of the art games in several relevant ways:

- game characters present more and more realistic facial expressions to convey the story;

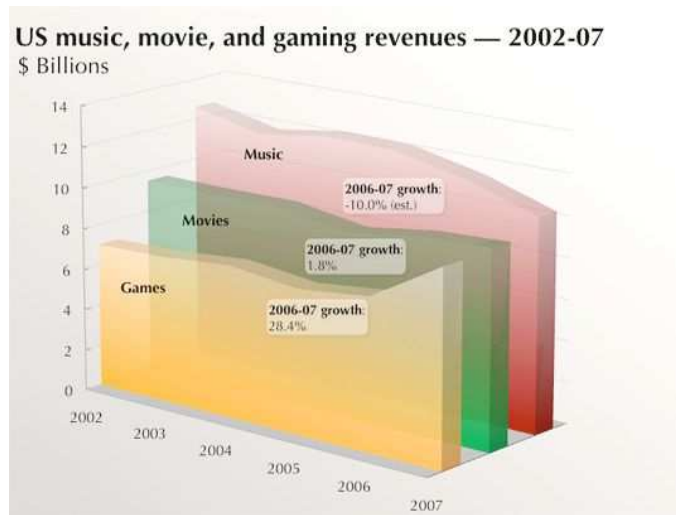


Figure 3.3: Entertainment industries comparison

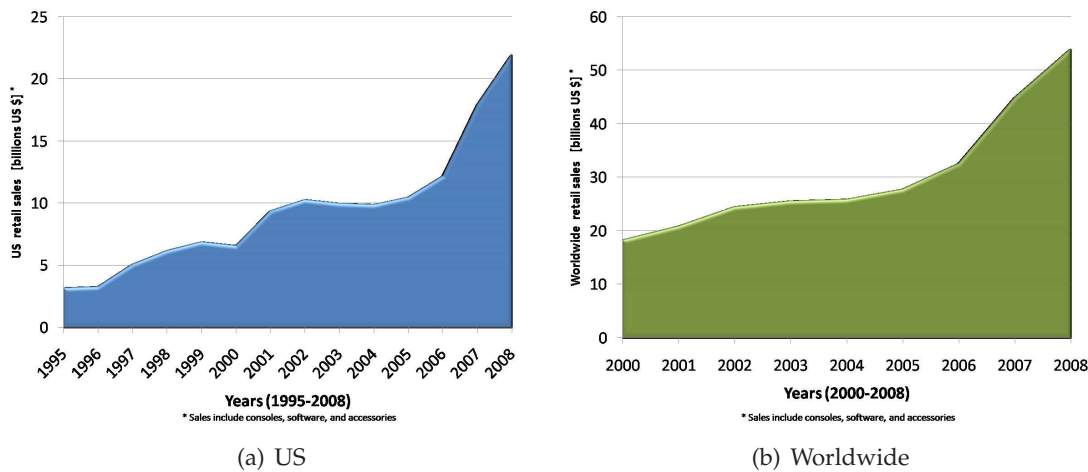


Figure 3.4: Retail games sales

- characters in the game are becoming more and more intelligent and start to simulate, synthesize, and “feel” emotions as reactions to event in the story;
- game developers try to influence the gamer emotions by “pushing” sense of calm or rhythm: beside the use of story events, musics and environments are often designed and chosen with this purpose;

Another interesting way of using emotions in the gaming environment would be to explicitly sense the gamer emotion and modify the game dynamics according to it (Jones and Sutherland [2008]). For example a computer game could react and add new opponents if the mood of the player was to become too calm approaching the bored state or, rather, decrease the difficulty of the game is the player was too aroused, stressed, and tired. Curves similar to the Yerkes–Dodson ones (see figure 3.1) can apply to this case. A certain, designed, level of arousal should be maintained, if we want to get the best feelings (i.e. the best perception over our emotions; see section 3.3) out of a game without too much stressing the player.



Another possibility, would be to design games with mood modes. An user starting the game could ask for a relaxed game or rather for a intense game session. In the first case, the computer will automatically select an easier difficulty level and dynamically adapt the challenges to reach a calm mood; in the second case, the game will adapt increasing the difficulty level and transferring more stressful emotions to the player.

### 3.2.4 Indexing and Retrieval

One exabyte correspond to  $2^{60}$  bytes  $\cong 10^{18}$  bytes. According to a study of the UC Berkeley School of Information (Charles et al. [2003]) “all words ever spoken by human beings” could be contained with 5 exabites of text. The same study assess that all the information generated in 1999 was 2 exabites; the one of 2003 5 exabites. In 2003 the US television broadcasted 31 million hours of new video.

According to Junee [2009] product manager for YouTube his company alone get about 20 hours of new videos uploaded every minute (more than 10 million hours a year). According to the same source this amount of data more than tripled in the last two years.

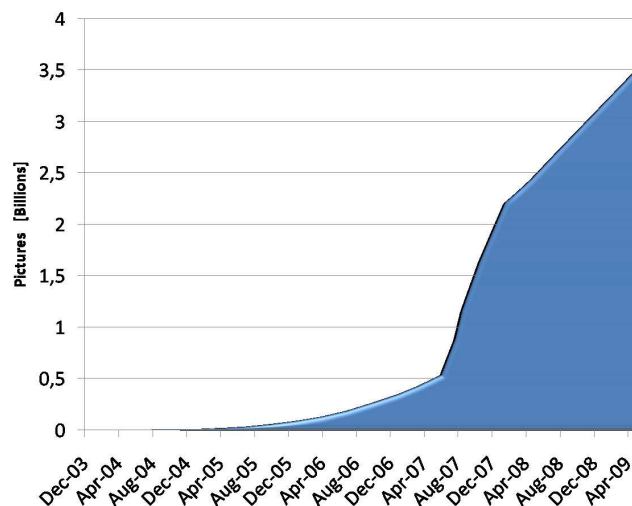


Figure 3.5: Total number of pictures uploaded on Flickr

Other services similar to YouTube (e.g. Dailymotion, Vimeo, etc.) get similar amount of video data. To these we have to sum up thousands and thousands of still images which are uploaded every minute to services like Flickr (see figure 3.5) or Picasa and all the media which are uploaded on web-pages and blogs and which are automatically crawled by the search engines.

This huge amount of user generated data is, nowadays, indexed, searched, and retrieved thanks to text labels (a.k.a. tags) which are given by the users when uploading the files or automatically assigned to the media thanks to the text surrounding them.

This system is not efficient, mostly because tags tend not to completely represent the media they are linked to and because a huge part of them are missing or misleading. State of the art research on indexing and retrieval work on “content-based” information. Current research systems extract information directly from the piece of media and do, for example, represent a song by its tempo and timbre, a pictures by its colors, and a movie a combination of all of these features.

Although content-based systems have the advantage to truly represent the content of the media researchers still struggle to define analysis patterns able to solve all issues related to the topic of indexing and retrieval. One question, which we will not tackle in this document is about the capability of these content based technologies to scale to the internet size.

The problematic that interest this document is the one commonly known as “*semantic gap*”. Users tend to use natural language to formulate queries; for example an user will search for “romantic sunset on the beach” and not for a picture where the dominant colors were red, orange, and violet. Generally speaking the semantic gap characterizes the difference between two descriptions of an object by different linguistic representations; in our case these are from the one hand the natural human language, and from the other hand content based information representing the image (e.g. colors or textures).

In all art expressions, and therefore in movies, pictures, and music, emotions are a fundamental source of information. In fact, as Ian Maitland, film director and editor, said: “*A film is simply a series of emotions strung together with a plot*” (Picard [1997]).

A prove of that is the fact that we tend to describe movie genres with affective information: for example we would say “a romantic movie”, “a funny comedy” or a “scary horror film”. Similarly we would tend to address music with affective descriptions: “a romantic lullaby”, “some energetic rock music”, or “a sad blues”.

A reason for that is that human memory tend to connect events with similar affective meanings and to remember events with strong affective meaning with more details (Cahill and McGaugh [1995], McGaugh and Cahill [2002]).

Extracting the emotional information from the content of a movie, picture, or song is, therefore, argued to be relevant for multimedia indexing and retrieval (Salway and Graham [2003], Chan and Jones [2005], Paleari and Huet [2008], Park and Lee [2008], Shan et al. [2009]). Emotions may represent one of the missing components for bridging the semantic gap. Indeed, emotions represent highly semantically abstracted features but are still computable with automatic computer processing.

Actually building retrieval systems based or helped by emotions will allow us to search for exactly the kind of media that we are in the mood for. Automatic recommendation systems could learn our emotional preferences and suggest the best music for our general tastes or even adapt music to our mood. In a not too far future, pervasive computers in our home could set the best illumination, and change the drawings to the ones best adapting to our mood also filling the environment with a soft music etc. In any of these scenarios emotions shall be used, together with other information, to index medias and control the environment.

### 3.2.5 Personal Intelligent Agent

Most of our everyday life is currently spent in front or near electronic devices such as computers, personal digital agents (PDA), or cellphones. Wouldn't it be just nice if our computers and all of our personal devices could know our preferences? Wouldn't it be nice if our phone could understand that we are in the middle of an important meeting and filters out non-important calls?

Learning one preferences and understanding surrounding events requires artificial intelligence, and often may need to be able to understand the emotions that are shown as reactions to events. Indeed, if our phone could understand that we get annoyed or embarrassed when it rings while we are in our boss office than it would also be able

---

---

to mute itself in that specific location. Similarly if computers could understand when people is getting tired, than they could suggest pauses.

Simple behaviors like this could not only improve the way we would feel about computers and our level of stress but also improve our efficiency at work. Looking back at the Yerkes–Dodson learning curves (see figure 3.1) we acknowledge that learning is more effective when students are stressed at the right level. We have already argued that similar curves can apply to computer games; similar curves can apply too to working environments. If we want workers to perform at their best we have to set a correct level of stress which shall depend on the psycho–physiological state of the worker and on the difficulty of the task (see figure 3.1).

A study from University of Melbourne showed that social network users are, in average, 9% more efficient than the others (Coker [2009]). The author of this study argues that this is because taking explicit short pauses on a regular basis allow us to better concentrate on our job during “work–time”. Dedicated software on our computers could help us to reach a certain, optimal “work–to–leisure time ratio” by suggesting to concentrate on work or on leisure. Affective computing could facilitate this kind of computer behavior as it would be favorable to adapt the ratio and the correct distribution of pauses to the current psycho–physical state of the specific worker.

### 3.2.6 Other Scenarios

In the previous sections we have seen few possible applications of affective computing. As we have said at the beginning of this chapter emotions are fundamental for human decision making, communications, memory, and many other cognitive functions. It is therefore clear that computers could take advantage of emotional abilities in many different ways and that the possible scenarios for affective computing are only limited to our imagination.

For example, affective computing could be used to help autistic people. People affect from autism often struggle to understand the emotions of the people surrounding them and therefore, to react in our natural ways. This fact often causes social troubles to to autistic people and to the people surrounding them. Personal device could help these people to partially solve these issue by recognizing the emotions and by suggesting “normal” reactions.

Another example could be the one of personal music recommendation of which we gave hints in the section 3.2.5. A system designed to recognize the user affective state could automatically learn to suggest the user’s favorite songs. The same system could further train to suggest the user’s favorite songs for a specific mood etc.

Yet another scenario could be the one usually known as the “Affective Mirror” (Picard [1997]). In this case our computers could help prepare ourself for an important meeting, interview, or show. As we said before, in human communications emotions play an important role. In some situations we would like to control the emotions that we communicate. A device, able to understand emotions could help us prepare for this kind of events by telling us how we do look like. The agent behind such a device could help us by impersonating the boss of a big company with whom we have an interview or rather our partner for a play.

To conclude, in this chapter, we have demonstrated that emotions could play an important role in everyday human–computer–interactions, computer–mediated–communications,

---

and artificial intelligence. We have motivated this fact by demonstrating the human need for emotions in communications, perception, and decision making. In many scenarios, like the one we have overviewed, we would like computers to be able to think in a way approaching the human one. In some other scenarios the affective abilities will not be very relevant; for example we probably will never want emotions to affect a calculator. Finally, there would be scenarios in which we would rather prefer computer to act completely rationally and not to feel emotions; this could be the case of the computer controlling the state of a nuclear plant.

It is, nevertheless, the responsibility of affective computing researchers to find technical solutions to allow computer to act emotionally for all those scenarios in which affective computing will be needed to improve the naturalness, the pleasantness, and the effectiveness of human–computer–interactions.

In the next chapter emotions and other affective phenomena will be discussed together with their most common representations.

### 3.3 Affective States

When referring to emotions as the key factor of affective computing we are being imprecise; we should rather say that affective states are the key component of affective computing. Indeed, emotions is just one of the possible affective states; *feelings*, *mood*, *attitudes*, *affective styles*, and *temperaments* all are, or are related to, affective states (Lisetti and Gmytrasiewicz [2002], Davidson et al. [2002]).

In this chapter we define these concepts and overview some of the models which are at the very basis of all my works.

**Emotion** could be defined as “*a relatively brief episode of coordinated brain, autonomic, and behavioral changes that facilitate a response to an external or internal event of significance for the organism*” (Davidson et al. [2002]). In other words emotions have both a physic and a psychic components and can be defined as the reaction to relevant events. When something happens the human brain always evaluate this event with respect to our needs, desires, beliefs, and intentions. In some cases the event is relevant: in these cases emotional reactions to those event are triggered. Emotional reactions will, then, affect the way we perceive, think, and act.

**Feelings** “*are the subjective representation of emotions*” (Davidson et al. [2002]). Loosely speaking, feelings refer to the way we experience the emotion from our individual point of view. In this sense emotions refer to the same entities, but while emotion refer to the entity itself (e.g. fear) feeling relate to our perception over that entity. So that, for example, we could be pleased to feel fear if that emotion was elicited by a movie scene.

**Mood** “*refers to a diffuse affective state that is often of lower intensity than emotion, but considerably longer in duration*” (Davidson et al. [2002]). In this sense moods are very similar to emotion but for two characteristics: 1) they last longer (hours to days according to Lisetti and Gmytrasiewicz [2002]) and 2) they do not focus on a specific event.

---

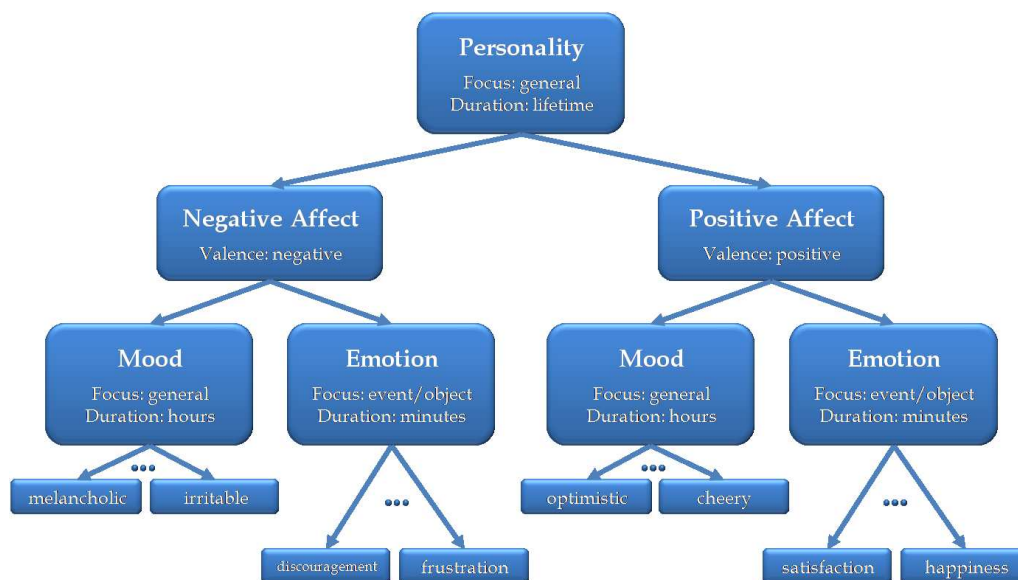


Figure 3.6: Hierarchic model of affective phenomena according to Lisetti and Gmytrasiewicz [2002]

**Attitudes** “are relatively enduring affectively colored beliefs, preferences, and predispositions toward objects or persons” (Davidson et al. [2002]). Attitudes are quite similar to emotional preferences, or as Damasio [2005] will call them “somatic markers”. Attitudes are more important than what is usually thought, indeed they allow us to make decisions rapidly and quickly express preferences.

**Affective styles** refer “to relatively stable dispositions that bias an individual toward perceiving and responding to people and object with a particular emotional quality, emotional dimension, or mood” (Davidson et al. [2002]). We could say that emotions stand to moods as attitudes stands to affective styles. Indeed, affective styles can be seen as attitudes which last longer and are directed toward big families of objects, events, or people.

**Temperament** finally “refers to particular affective styles that are apparent early in life, and thus may be determined by genetic factors” (Davidson et al. [2002]). In other words, temperament and personality are even more generally focused affective states. One person could be for example generally crusty or unsociable and generally react more negatively than others to people and events.

Natural communications often make use of the words “emotions” and “emotional” in their spread meaning of affective states. In the applications of affective computing the word emotions is usually used in its narrower meaning. It can nevertheless, be useful to keep in mind the definitions of all the different affective states and the interactions that different states can have with each other.

### 3.3.1 What is an emotion?

After more than 100 years (Darwin [1872]) of study of emotions and emotional phenomena (i.e. affective states), researchers cannot agree on a single and precise definition.

Many questions about emotions are still topical (Lazarus [1991], Lazarus and Lazarus [1996], Ekman and Scherer [1984]). Emotions are studied in several different domains: psychologists, physiologists, biologists, anthropologists, sociologists, philosophers and even ethologists study emotions (Scherer and Ekman [1984]).

Albeit differences still exist, researchers seem to agree on few characteristics of emotions and other affective states:

- emotions are *reactions to events* (both internal or external);
- emotions have both a *psychological* and a *physiological* components;
- emotions have *communicative, social, biological, and cognitive* functions;
- some characteristics of emotions are common to *every culture*;
- some characteristics of emotions can be found in *other species*;

### 3.3.2 Structured Model of Emotions

Over the years researchers have tried to classify emotions and to find models able to represent them. Three main families can be defined. These are:

- discrete categories;
- dimensional models;
- componential model.

In the next sections we will overview these three families in more detail.

#### 3.3.2.1 Discrete Categories

An instinctive approach to affect classifications is to construct a list of emotional families that disjointly categories basic emotions. The most common categories are the one often referred to as the *big six* or the *six universal* emotions. These are *anger, disgust, fear, happiness/joy, sadness, and surprise* to which researchers often add a *neutral* state. These classes are the one listed by Ekman, Friesen, and Ellsworth in their works (Ekman et al. [1972, 1982], Ekman [1992]).

But many other classification schemes have been adopted during the years by the different researchers (see table 3.1 from Ortony and Turner [1990]).

Parrott [2001] takes a different approach and defines a taxonomy of emotions categories starting from the basic emotions (i.e. love, joy, surprise, anger, sadness, and fear) and from these 6 derives 25 secondary categories and therefore the same number of sets of tertiary emotions (see table 3.2) thus allowing for a more deep compression and classification of the emotions.

The underlying principle remain nevertheless the same; some labels are assigned to the emotions. The actual meaning of each one of these labels is, nevertheless, culturally dependent and in some way subjective. In other words not only different researchers in table 3.1 use different labels, but also they have similar yet different meaning for each one of these labels.

---



Emotion categories	Reference Paper
Acceptance, anger, anticipation, disgust, joy, fear, sadness, surprise	Plutchik [1980]
Anger, aversion, courage, dejection, desire, despair, fear, hate, hope, love, sadness	Arnold [1960]
Anger, disgust, fear, joy, sadness, surprise	Ekman et al. [1982]
Desire, happiness, interest, surprise, wonder, sorrow	Frijda [1986]
Rage and terror, anxiety, joy	Gray [1985]
Anger, contempt, disgust, distress, fear, guilt, interest, joy, shame, surprise	Izard [1977]
Fear, grief, love, rage	James [1884]
Anger, disgust, elation, fear, subjection, tender-emotion, wonder	McDougall [1926]
Pain, pleasure	Mowrer [1960]
Anger, disgust, anxiety, happiness, sadness	Oatley and Johnson-Laird [1987]
Expectancy, fear, rage, panic	Panksepp [1982]
Anger, interest, contempt, disgust, distress, fear, joy, shame, surprise	Tomkins [1984]
Fear, love, rage	Watson [1930]
Happiness, sadness	Weiner and Graham [1984]

Table 3.1: Basic emotion categories defined over the years

### 3.3.2.2 Dimensional Emotion Descriptions

Dimensional models of emotions do not assign labels to the different emotions but coordinates into a  $n$ -dimensional space. This space is designed in such a way that different coordinates define and identify uniquely all the possible different affective states.

The most used dimensional space is the Valence - Arousal (VA) (a.k.a. Pleasure - Activation) space, derived from the Pleasure, Arousal, and Dominance (PAD) space by Russell and Mehrabian [1977] (see figures 3.7 and 3.8). This latter three dimensional approach breaks down emotions along three dimensions: *pleasure*, *arousal*, and *dominance*. In figure 3.8 we show the PAD space and give examples of possible placement of four standard emotions (i.e. anger, fear, happiness, and sadness).

The *pleasure* dimension, often referred to as pleasantness, valence, or evaluation, indicate how much an emotion is positive or negative. For example someone sad has evaluated surrounding events as very negative. On the contrary, someone feeling joy would have appraised the environment as positive for his well being.

The *arousal* dimension, a.k.a. activation indicates how relevant the surrounding events are and therefore how strong the emotion is. In this case, someone feeling excited will have an emotion represented by a bigger arousal coordinate and someone feeling bored will experience a much less relevant emotion.

Finally the *dominance*, sometimes named power or control, reflects the degree of control the subject has over the current situation. This dimension is often used to distinguish between the two emotion of anger and fear (see figure 3.8) as they both refer to negative and arousing emotions with the difference that when we experience anger we believe we could change, or could have changed, the output of the current events and when we experience fear we do not trust our power to be enough to do this.

In figure 3.7 and 3.8 we respectively show the valence-arousal and the pleasure-arousal-dominance spaces together with few example about how some emotional cat-

Primary	Secondary	Tertiary
Love	Affection	Adoration, affection, love, fondness, liking, attraction, caring, tenderness, compassion, sentimentality
	Lust	Arousal, desire, lust, passion, infatuation
	Longing	Longing
Joy	Cheerfulness	Amusement, bliss, cheerfulness, gaiety, glee, jolliness, joviality, joy, delight, enjoyment, gladness, happiness, jubilation, elation, satisfaction, ecstasy, euphoria
	Zest	Enthusiasm, zeal, zest, excitement, thrill, exhilaration
	Contentment	Contentment, pleasure
	Pride	Pride, triumph
	Optimism	Eagerness, hope, optimism
	Enthrallment	Enthrallment, rapture
	Relief	Relief
Surprise	Surprise	Amazement, surprise, astonishment
Anger	Irritation	Aggravation, irritation, agitation, annoyance, grouchiness, grumpiness
	Exasperation	Exasperation, frustration
	Rage	Anger, rage, outrage, fury, wrath, hostility, ferocity, bitterness, hate, loathing, scorn, spite, vengefulness, dislike, resentment
	Disgust	Disgust, revulsion, contempt
	Envy	Envy, jealousy
	Torment	Torment
Sadness	Suffering	Agony, suffering, hurt, anguish
	Sadness	Depression, despair, hopelessness, gloom, glumness, sadness, unhappiness, grief, sorrow, woe, misery, melancholy
	Disappointment	Dismay, disappointment, displeasure
	Shame	Guilt, shame, regret, remorse
	Neglect	Alienation, isolation, neglect, loneliness, rejection, defeat, homesickness, dejection, insecurity, embarrassment, humiliation, insult
	Sympathy	Pity, sympathy
Fear	Horror	Alarm, shock, fear, fright, horror, terror, panic, hysteria, mortification
	Nervousness	Anxiety, nervousness, tenseness, uneasiness, apprehension, worry, distress, dread

Table 3.2: Emotion taxonomy by Parrott [2001]



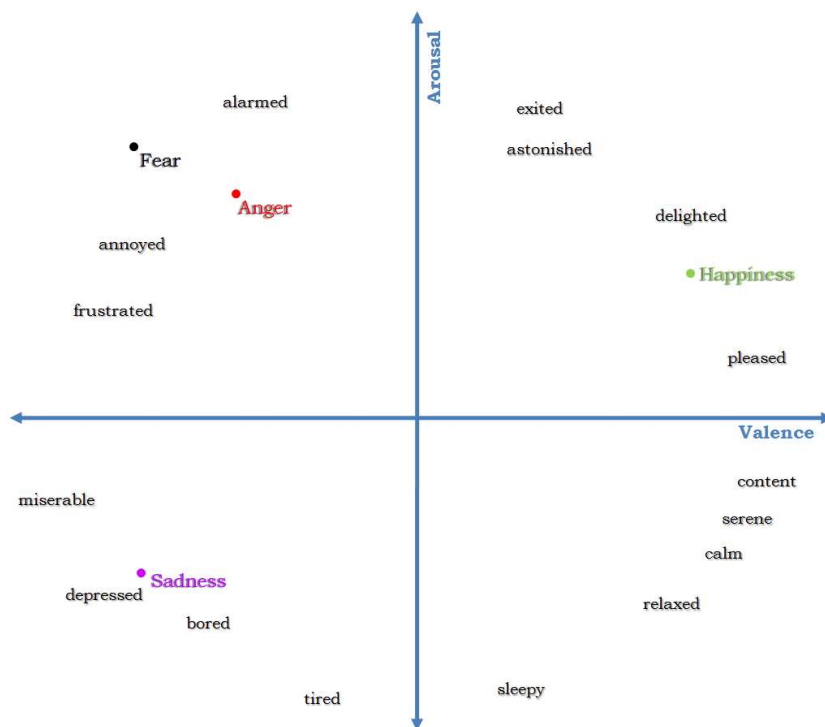


Figure 3.7: Valence–Arousal model of emotions

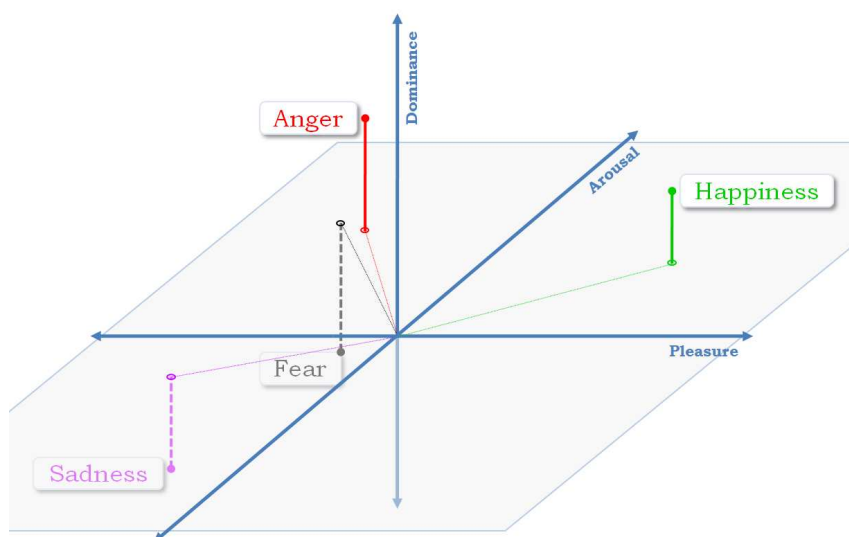


Figure 3.8: PAD model of emotions adapted from Russell and Mehrabian [1977]

egories from section 3.3.2.1 could be placed inside the spaces.

### 3.3.2.3 Componential Models

A third approach consists in representing the emotions through the *components* that brought the evaluation of the surrounding events to that emotion. If the discrete emotions can be useful for their simplicity of use and the dimensional model can be useful for computer applications because they can represent emotions in a continuous space, these third componential models are very interesting because they can not only represent the emotions but also the process of appraisal of the emotional states. Indeed, these representation always pass through the concept of *emotional appraisal*. According to theories of appraisal, emotions arises via a process of evaluation of the events (internal or external) surrounding the subject. Different emotions are associated with different appraisal patterns and vice versa.

Few models have been developed which are currently used by researchers. The most relevant are the one by Frijda [1986], the OCC model by Ortony et al. [1988], the Component Process Theory (CPT) by Scherer [1984], and the derived one by Lisetti and Gmytrasiewicz [2002].

In a certain sense dimensional model descriptions could be considered similar to these models but for two main characteristics:

- dimensional models have usually 2 or 3 dimensions while here we can have as much as 16 different components;
- in dimensional models infinite emotions are represented by n-dimensional vectors, in componential models the maximum number of emotions is computable and it is, usually, equal to  $3^{n\_dimensions}$  (i.e. for each dimension/component there are only 3 possible values: "positive", "negative", and a third "none/both/not-applicable" state).

The main advantage of the componential models is that the components are linked to the appraisal process and therefore are generally explicitly computable with logic and/or mathematical operations while the dimensions are not.

**Frijda** In his masterpiece "*The Emotions*", Nico H. Frijda [1986] describes a simple componential model which can be used to discriminate among 31 emotional states thanks to 26 different components.

The components of this model are: *positive character, negative character, desire, interest, positive valence, negative valence, presence, absence, certainty, uncertainty, change, open, closed, intentionality of the other, intentionality of the self, controllability, non controllability, modifiability, finality, object, event, focality, globality, strangeness, familiarity, and value.*

In this example we see that some components negate each other (e.g. positive vs. negative character or positive vs. negative valence), further reducing the maximum number of different representable emotions.

**Ortony, Clore, and Collins** The Ortony, Clore, and Collins (OCC) model (Ortony et al. [1988]) is a computational model of emotions. The OCC model is well known for it has been used by researchers for building up artificial intelligence (AI) systems with emotional capabilities. Through the OCC 22 different emotion categories are defined. Figure

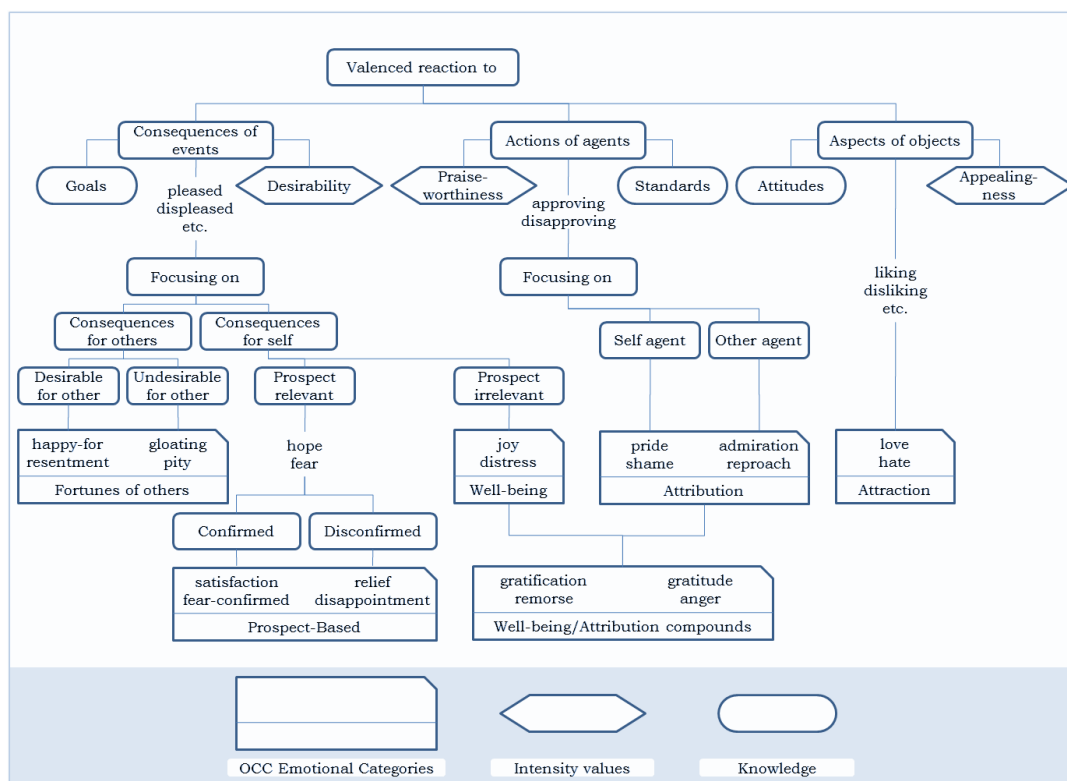


Figure 3.9: OCC computational model of emotions Ortony et al. [1988]

3.9 shows how, according to Ortony Clore and Collins, emotions can be computed by an agent.

The process of computing the resulting emotion given an event can, in the OCC model, resumed in five main steps:

- *Classification*: the agent evaluates an event and finds the affected emotional category;
- *Quantification*: the agent finds the intensities of the affected emotional categories;
- *Interaction*: the agent finds the new emotion given the current (old) one;
- *Mapping*: the agent maps the emotion to a subset of emotion (optional);
- *Expression*: the agent express the emotion through behavior and/or facial expression.

During the first phase of classification of the emotion, some of the characteristics of emotions described before are taken in account. Than, during the other phases, the agent would compute the intensity of the emotion and adjust all the values to better behave with respect to social and internal non explicit rules.

Please note that Christoph Bartneck [2002] argues that consistency of the agent can be seen as a personality trait. In his idea it is possible to set OCC parameters to set the personality of the agent. Therefore, computing emotions through the OCC model would automatically take into some account personality traits. In this case the model would not

have an explicit model of the personality of the agent but nevertheless the OCC model would be a computational model that, in some way, presents an “embedded” personality. This should be done by refining parameters in an iterative way until the desired behavior are verified.

**Scherer** Scherer’s *Component Process Theory* (CPT) Scherer [1984, 1987] describes emotions as arising from a process of evaluation of the surrounding events with respect to their significance for the survival and well-being of the organism. The nature of this appraisal is related to a sequential evaluation of each event with regards to some parameters called SECs or *Sequential Evaluation Checks*.

*Sequential Evaluation Checks* (SECs) are chosen to represent the minimum set of dimensions necessary to differentiate emotions and they are organized in four classes or in terms of four appraisal objectives:

1. **Relevance SECs:** How relevant is the event to the agent? This can incorporate attention focusing mechanisms. Included are notions of:
    - **Novelty:** representing whether the event was sudden, familiar, and expected or not;
    - **Intrinsic Pleasantness:** basic pleasure or pain check and prior to a positive goal/need conduciveness check;
    - **Goal Relevance:** establishes the relevance, pertinence, or importance of a stimulus or situation for the momentary hierarchy of goals/needs.
  2. **Implications SECs:** What are the consequences of this event and how does it affect my “well-being” and immediate to long-term goals? A central appraisal objective which includes:
    - **Causal Attribution:** attribution of responsibility to another agent, and if “intelligent”, then inference undertaken relating to their intent or motive;
    - **Outcome Probability:** represents the likelihood of the event (e.g. if the event was likely or not to happen);
    - **Discrepancy from Expectations:** represents whether the agent was expecting the outcome of the event or not;
    - **Goal-Need Conduciveness:** represents whether the event helps the agent reach its goals and needs or not;
    - **Urgency:** represents how much time the agent has to react to the outcomes of the event.
  3. **Coping Potential SECs:** How well can we cope with or adjust to these consequences? Determines which types of responses to an event are available to the agent and their consequences
    - **Control:** represents how much the event and the situation are controllable or they rather follow a randomic behavior;
    - **Power:** if control is possible, coping potential depends on the power one has to be able to exert control or to recruit others to help. It influences, for example, whether to feel anger or fear, whether to fight or flight;
-

- **Adjustment:** can the agent adjust, adapt to or live with the consequences of an event after intervention, even if the outcome of the event was not good? This may be associated with revisiting ones desires or goals to a certain extent.
4. **Normative significance SECs:** The last appraisal objective represents the significance of events with respect to self-concepts as well as social-norms and values?
- **Internal Standards:** describes how much the event conform to a set of internal rules like ideals and morals;
  - **Social Norms:** describes how much the event conforms to a set of socially accepted rules.

	ENJ\HAP	ELA\JOY	DISP\DISG	CON\SCO	SAD\DEJ
<i>Relevance</i>					
<b>Novelty</b>					
Suddenness	low	high\med	open	open	low
Familiarity	open	open	low	open	low
Predictability	medium	low	low	open	open
<b>Intrinsic Pleasantness</b>	high	open	very low	open	open
<i>Implication</i>					
<b>Conduciveness</b>	high	very high	open	open	obstruct
<i>Coping Potential</i>					
<b>Control</b>	open	open	open	high	very low
<b>Power</b>	open	open	open	low	very low
	DESPAIR	ANX\WOR	FEAR	IRR\COA	RAG\HOA
<i>Relevance</i>					
<b>Novelty</b>					
Suddenness	high	low	high	low	high
Familiarity	very low	open	low	open	low
Predictability	low	open	low	medium	low
<b>Intrinsic Pleasantness</b>	open	open	low	open	open
<i>Implication</i>					
<b>Conduciveness</b>	obstruct	obstruct	obstruct	obstruct	obstruct
<i>Coping Potential</i>					
<b>Control</b>	very low	open	open	high	high
<b>Power</b>	very low	low	very low	medium	high
	BOR\IND	SHAME	GUILT	PRIDE	
<i>Relevance</i>					
<b>Novelty</b>					
Suddenness	very low	low	open	open	
Familiarity	high	open	open	open	
Predictability	very high	open	open	open	
<b>Intrinsic Pleasantness</b>	open	open	open	open	
<i>Implication</i>					
<b>Conduciveness</b>	open	open	high	high	
<i>Coping Potential</i>					
<b>Control</b>	medium	open	open	open	
<b>Power</b>	medium	open	open	open	

Table 3.3: Predicted appraisal patterns for some emotions

In table 3.3 we report the predicted appraisal patterns for some of the most common emotional categories.

Scherer [2001] discusses the SEC approach within the context of three levels of emotional processing, as also suggested by Leventhal [1984]. In section 3.1.1 we have described how LeDoux [1990, 1994] describe two way of perceiving. To these two Leventhal [1984] and Scherer [2001] are adding a third “*schematic*” level in which perception is basically compared with acquired schema. According to Leventhal and Scherer in some

	SENSORY MOTOR	SCHEMATIC	CONCEPTUAL
NOVELTY	SUDDENNESS i.e. a Gun Shot	FAMILIARITY check <i>Working Memory (WM)</i> & schemata i.e. a second shot - the agent shots?	EXPECTATIONS check <i>Expected Memory (EM)</i> i.e. the agent asked someone to shot
INTRINSIC PLEASANTNESS	check a Belief DB with pleasantness for human preferences i.e. warmth, light, or sweet	check a DB with learned preferences i.e. chocolate, coke	check recalled preferences that are not yet learned i.e. the second time riding horses
GOAL RELEVANCE	check life threatening goals i.e. survival, food & procreation	check learned desires or learned main sub-goals i.e. chocolate or money	check each possible desire and sub-goal i.e. get this table filled
CAUSAL ATTRIBUTION	-	check into schemata and WM the agent which started the action i.e. light lits up -> switch -> agent	check into WM and LTM which agent could have started the action
OUTCOME PROBABILITY	-	-	computes the probability that the event will occur, has occurred or is occurring i.e. is my uncle winning a lottery?
DISCREPANCY FROM EXPECTATIONS	-	check schemata and WM for situation leading to the actual state. i.e. expecting switches to lit on lights	check EM for the event to see if it was expected i.e. expect something to happen
GOAL-NEED CONDUCTIVENESS	-	check schemata and WM for possibilities for reaching goals i.e. schematic reasoning	check memories for possible future goal state that are been approaching (avoiding) i.e. reasoning
URGENCY	it is set in computing Novelty suddenness i.e. a gun shot is novel and it is urgent	estimate time according to schemata i.e. playing basket what is written on the notice board	estimate remaining time i.e. reasoning about remaining time
CONTROL	-	check from schemata whether the situation is controllable i.e. a basket game is controllable	check whether the situation is controllable i.e. reasoning about the situation
POWER	check energy available to the agent i.e. the agent is sick	search for schemata controlling the situation i.e. take meds when sick	check how many coping strategies the agent has i.e. buy meds, ask a doctor or mum
ADJUSTEMENT	-	-	check whether the agent could adjust to the consequences of the event
INTERNAL STANDARDS	-	check whether the event is represented in one-self schemata	check one-self moral ideals i.e. not to kill, not to eat meat
SOCIAL NORMS	basic human standards i.e. empathy	checks whether the event is represented in social schemata i.e. politeness	check social moral ideals i.e. laws and social rules

Table 3.4: SECs interpretations and the levels of cognition. Derived from Scherer [2001]

cases we react to the events that we perceive not as an automatic reaction nor as a cognitive reaction but rather as a learned schema. For example when driving, if we see a red traffic light, we do not have to think about the driving rule set to start breaking. We have stopped in front of a traffic light so many times that we have created a kind of shortcut in our brain linking the perception of the red traffic light to the action of breaking.

The three level of emotional processing referenced by Scherer are therefore:

- *Sensory-Motor Level*: Checking occurs through innate feature detection and reflex systems based on specific stimulus patterns. Generally it involves genetically determined reflex behaviors and the generation of primary emotions in response to basic stimulus features. For example if something big and black approaches, then the reaction of moving back and the elicited emotion of fear will both belong to this level.
- *Schematic Level*: The learned automatic non-deliberative rapid response to specific stimulus patterns largely based on social learning processes. For example the small talk sentence Good afternoon when meeting someone is a typical behavioral or schematic reaction as it is the emotions arising from the victory of a sport team.
- *Conceptual Level*: Checking is based on conscious reflective (deliberative) processing of evaluation criteria provided through propositional memory storage mechanisms. Planning, thinking and anticipating events and reactions are typical conceptual level actions. Emotions arise from cognitive processes, as the reproach for a non moral action or anxiety for the result of an important exam.

According to Scherer's theory every check can be performed at all the level with different levels of abstraction and precision. However it is our opinion that some of the checks cannot be computed, for their intrinsic natures, at all level; for example it seems to us that checks like causal attribution, outcome probability and discrepancy from expectation cannot be evaluated without some form of reasoning (conceptual or schematic).

In table 3.4 we list the checks with their interpretation according to the level of cognition listed before.

Thanks to the CPT it is possible to define many different emotions. Scherer [2001] gives 13 different examples of emotional categories but in theory this model could represent all the different emotional categories that we can feel (see table 3.3 for the SECs prediction for some of the most common emotional categories).

The component process theory is particularly interesting for computer scientists for three main reasons. Firstly, the CPT of emotions is among the most complete psychological models of emotions and is, nevertheless, very clear. Secondly, the very nature of the CPT facilitate the translation to computer algorithms for artificial intelligence. Thirdly, Scherer is actively collaborating with computer scientists willing to transfer his theory into practice. For example Wehrle and Scherer [2001], Kaiser and Wehrle [2001] developed two computer science study on respectively the recognition and the display of facial expressions which are strictly related to Scherer's work. It is therefore, theoretically possible to build emotion display and emotion recognition modules using the Scherer's CPT of emotions.

**Lisetti and Gmytrasiewitch** The model from Lisetti and Gmytrasiewitch (Lisetti and Gmytrasiewicz [2002]) is the last one that we will present in this part of the thesis. According to this model, emotions can be modeled via sixteen different characteristics which are partially inspired by Scherer's CPT (Scherer [1984]) but also from the works of Frijda [1986], Frijda and Swagerman [1987], Ortony et al. [1988], Roseman et al. [1996].

The representation of Scherer's parameters has been modified to make it better fit a computer representation and expanded to also represent other affective states than emotions. In particular, the two parameters *focality* and *duration* help defining if the affective state described is an emotion or mood. In the first case (emotion) focality will be defined and duration will be short (seconds), in the latter (mood) no focality will be specified and the duration will be longer (hours/days). The parameters that the two authors list are:

- Facial expression: the facial expression associated to the emotion.
- Valence: information about the pleasantness of the emotion.
- Intensity/urgency: the intensity of the emotion felt.
- Duration: the time span affected by the emotion.
- Focality: can be an object, an event or global. Together with duration can give us the information about the kind of affective phenomena.
- Agency: who is responsible for the emotion?
- Novelty: if the emotion was predictable or not
- Intentionality: if the triggering event was caused by someone

- Controllability: how much the agent can cope with the triggering event?
- Modifiability: referred to time necessary to cope with the event
- Certainty: referred to anticipation effect to come
- Legitimacy; is the emotion felt legitimate?
- External norm: is the triggering event, acceptable for the others?
- Internal standard: is the triggering event, acceptable for oneself?
- Action tendency: identify the most appropriate coping strategies
- Causal chain: identify the cause of the triggering event

In this part of the thesis the definition and motivations for affective computing have been given together with some simple application scenarios. Then, we have presented the various affective phenomena and given a simple taxonomy. In this chapter we have defined some concepts related to the psychological theories of emotions and described the main models of emotions which can be used in computer science.

In the next parts we will discuss the topic of emotion display and present the processing involved with the generation of believable facial expressions.

---



## Chapter 4

# Emotional Display

### 4.1 Introduction

We have discussed in part 2 the motivations for affective computing which include the influence emotions held on most human cognitive functions such as decision making, perception, memory, and communications. All of these reasons, coupled with the world-wide spread of computing devices motivate the extensive study of affective computing. Affective computing is the field of study that investigate how computers can understand, simulate, and display emotions during human–computer–interactions and how these capabilities could be exploited to improve the quality, the pleasantness, and, therefore, also the effectiveness of these kind of interactions.

We have seen that there exist three different families of affective capabilities: affective display, affect recognition, and affect synthesis. This part is commit to the first one of these three abilities namely **affect display**.

Among the three affective computing parts, affect display is the only one which is already largely exploited by the industry. It is enough to think about the cinema and video–games industries to have an idea of this fact.

Emotional facial expressions are also largely used whenever people is interacting with virtual characters, which is more and more common in e–commerce web–sites. The same can be said for emotional vocal prosody. It is not by chance that Loquendo [2009], the world leader in speech related technologies and in particular in text–to–speech, advertise its products saying that “*Loquendo’s voices are expressive, clear, natural and fluent: they have been enriched with a repertoire of “expressive cues” that allow for highly emotional pronunciation*”.

E–learning (virtual teachers), tele–medicine (virtual nurses, remote medic assistance), home and office assistance (virtual secretaries) and game environment are only a few of the existing application domains for which affective display capabilities are studied and used by industrials.

### 4.2 Relevant Work

In this chapter we will discuss the state of the art about display of emotions by computers. Human have several different ways for transferring emotional information during communications:

---

- **facial expressions** are the most evident mean of displaying emotions. Indeed, if we do not consider the facial movement linked to speech production every movement on the human face seem to have evolved in the direction of allowing us to display our emotions through facial expressions.
- **vocal prosody** regards the way we change our voice tone while speaking. Vocal prosody represents a second obvious source of emotional information in human communications.
- **head, body posture, and gaze** are also simple means of expressing emotional information. For example lowering the gaze and the head could be interpreted by itself as a negative valenced, non-dominant emotion such as sadness or shame.
- **gestures** can also being used by humans to express emotions. In this case the transfer is usually more voluntary but some non-volitional gestures can be interpreted by experts as hints with emotional meaning.
- **word semantics** can obviously been linked to emotional meaning. For example one could simply say he is happy or non-volitionally use more enthusiastic words than usual.

In the next few sections we will overview one psychological theory about the generation of multimodal emotional expressions (facial, prosodic, and responses from the autonomous nervous systems a.k.a. ANS) and some techniques and technologies involved with facial expression generation (3D avatars and robots) and prosodic expression generation.

#### 4.2.1 Scherer's Component Process Theory for Emotional Expressions

Scherer's CPT allows to model the process of appraisal of emotions. According to Scherer appraisal passes through a process of multilevel sequential checking of the surrounding events with respect to the agent beliefs, needs, and norms (for more details about this theory please refer to section 3.3.2.3 and Scherer [1984, 1987, 2001]).

Scherer [1987] also details some of the physical components of emotions.

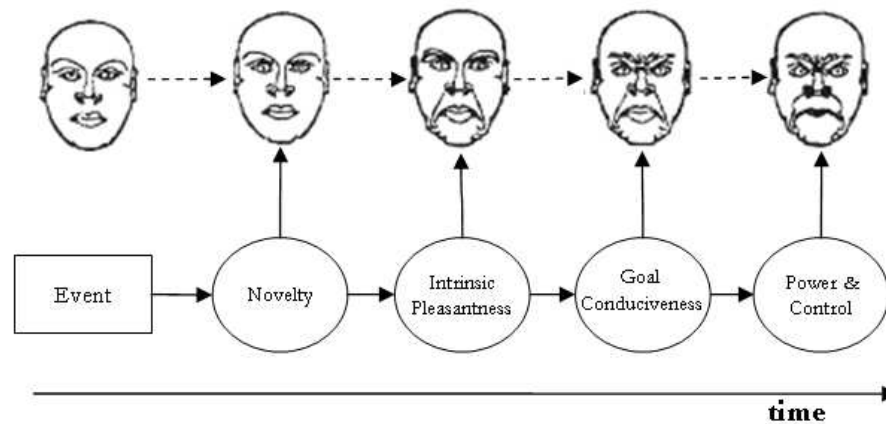


Figure 4.1: The facial expression process for *anger* according to Scherer's CPT

The CPT approach, exemplified in figure 4.1 for the *anger* facial expressions, is to characterize each SEC which the additive modification which take place on the expressions (facial, prosodic, ANS, etc.). In Scherer [2001] these predictions are given in terms of combinations of Ekman [1971] action units (AUs, see section 4.2.2.2). In a similar way Scherer [2001] also details the expected modifications of the speech prosody (in term of *pitch, energy, low frequency energy, duration, ...*) or ANS signals (*heart rate, skin conductivity, salivation, pupil diameter, ...*).

One interesting idea is the one that CPT could be used by computers both for *generating* emotional expressions and for *recognizing* emotional states. In figure 4.2 we exemplify this idea for the case of emotional facial expressions.

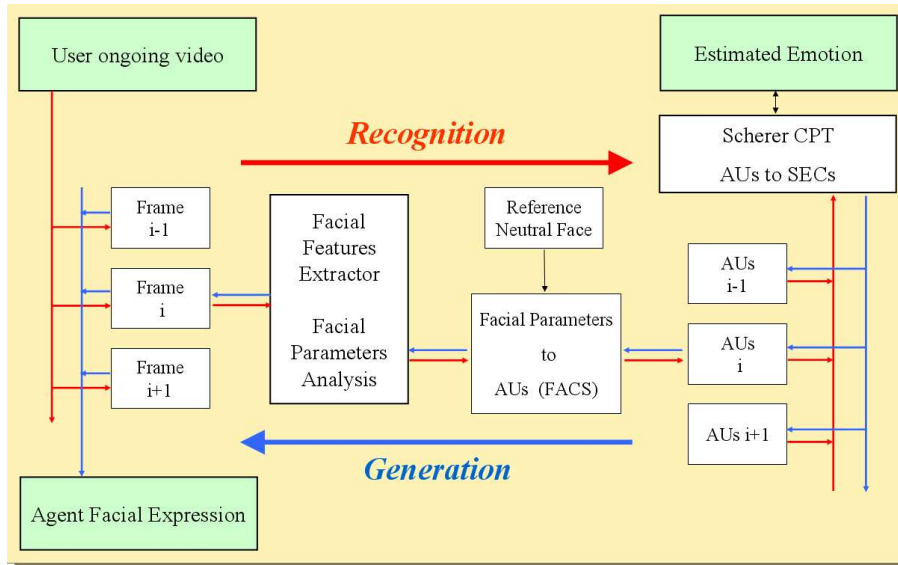


Figure 4.2: One algorithmic approach to emotional facial expression recognition and display based on CPT

Given a video of a real subject, a computer does automatically extract some facial movement measures (e.g. feature points movements or facial motion measures) for the recognition of emotions. Secondly, it compares these measures to the neutral facial state and deduces which muscles are activated. Finally, comparing the activated action units with the predictions given by Scherer's CPT, the computer can estimate the activated SECs, and therefore have a representation of the displayed emotion.

For the display we can, given an emotional state appraisal, convert the various SECs into muscular movements predictions and display these movements to the agent face to generate believable facial expressions. Please note that if the emotional appraisal was not available, then we would have to convert the emotional category to the matching SECs sequence (see table 3.3 for the SECs prediction for some of the most common emotional categories).

It is important to note that we are not discussing about replicating or imitating the facial expressions which are recognized on the human face but about the possibility of replicating the very process of appraisal which brought to the creation of these facial expressions.

The principal advantage of such an approach is that if an agent's cognitive architecture uses the same SECs representation for the reasoning that are used for recognition and

generation of emotive expressions then, in the whole system, emotions will not have to be linked to labels (e.g. happiness, sadness or fear) and it would be possible to maintain all the emotional nuances that might otherwise be lost.

Few researchers before us have tried to use Scherer's theories to develop facial expression generation and recognition systems.

Wehrle, Kaiser, Schmidt, and Scherer (Wehrle et al. [2000], Wehrle and Scherer [2001], Kaiser and Wehrle [2001]) have created a tool to calculate and draw facial expressions based on Scherer's psychological theory. FACE (Facial Animation Composing Face) (a.k.a. Geneva Appraisal Theory Environment GATE) is the tool they designed.

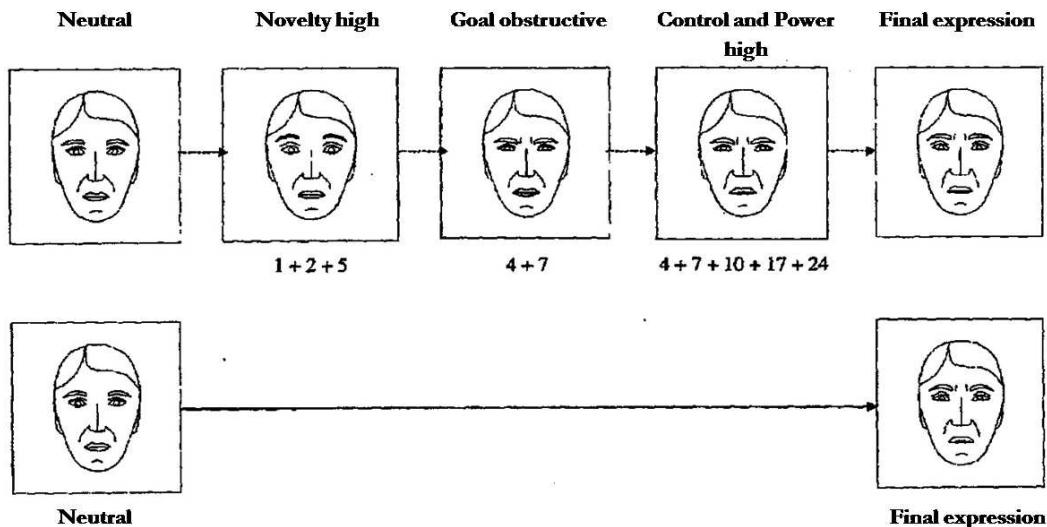


Figure 4.3: FACE example for the emotion anger

FACE permits to control the dynamics of expressions, to simulate head movements and to represent AUs via a sketch of a human face (see figure 4.3). In their study they have evaluated how humans do recognize Ekman et al. [1982], Ekman [1992] universal expressions: happy, fear, anger, disgust, sadness, surprise. They compared the results with the one obtained for Ekman expression's and with some photos. The recognition score of Wehrle's expressions is comparable to the one obtained with pictures showing Ekman expressions except for fear, disgust and surprise but the graphics was very low as it can be seen in figure 4.3

## 4.2.2 Emotional Facial Expressions

Emotional facial expressions are the most noticeable form of emotional communication. In the part 3 while talking about the motivations for affective computing we have demonstrated that emotions are fundamental for human natural communications (Mehrabian and Wiener [1967], Mehrabian and Ferris [1967], Mehrabian [1971, 1972]).

Thanks to the researches by Ekman (Ekman et al. [1982], Ekman [1992]) we also know that there exist some emotions which are represented in similar way all around the globe. Furthermore, we have just overviewed in section 4.2.1 a model allowing to describe the process of creation of facial expressions in humans. In this section we will overview some models for "pseudo-mathematically" representing facial expressions.

#### 4.2.2.1 Facial Expression Models

#### 4.2.2.2 Action Unit and FACS

**Duchenne's Researches** Guillaume Benjamin Armand Duchenne was French neurologist, who was first to describe several nervous and muscular disorders and, in developing medical treatment for them, created electrodiagnosis and electrotherapy. He applied electrodes for recording the path that electricity took in a contracting muscle's fibers. Duchenne investigated every major superficial muscle with his development and application of surface electrodes, which were used to measure abnormal and normal muscle action.

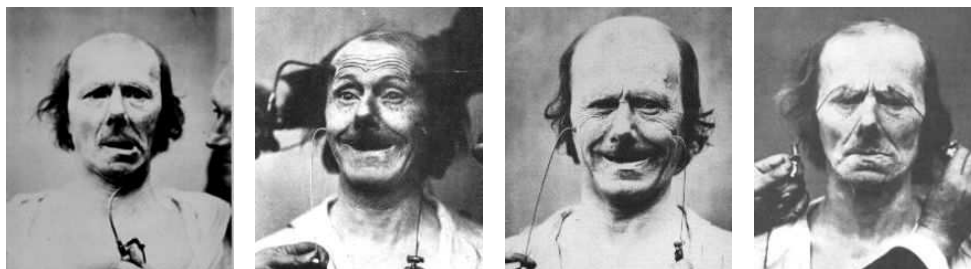


Figure 4.4: Pictures of Duchenne De Boulogne [1862]'s experiments on the "Old Man"

In the 1850 *"Bulletin de l'Academie"*, Duchenne made his first public reference to his long study of the relationship between muscular contraction (facial) and expressed emotion, via a process he delineated as 'faradism', in which an electrical stimulus is applied directly to or through the skin by rheopore.

In *"The Mechanisms of Human Facial Expression"* (Duchenne De Boulogne [1862]), first published in French in 1862, his fascinating photographs (see figure 4.4) and insightful commentary provided generations of researchers with foundations for experimentation in the perception and communication of human facial affect. Duchenne's principal photographic subject, the "Old Man", was afflicted with almost total facial anesthesia. This circumstance made him an ideal subject for Duchenne's investigations, because the stimulating electrodes he used were certainly somewhat uncomfortable, if not actually painful.

Guillaume Duchenne mapped 100 facial muscles in 1862. In the course of that work, he had something to say about smiling. He pointed out that false, or even half-hearted, smiles involved only muscles of the mouth. But "the sweet emotions of the soul," he said, activate the pars lateralis muscle around the eyes.

**Ekman's FACS** Several works have tried to model facial expressions according to some parameters, like movements or position of points and combination of these parameters. The most used model is the *Facial Action Coding System* by Ekman and Friesen.

The Facial Action Coding System (FACS) has been defined by Ekman and Friesen in 1971 and perfected during the following years (Ekman [1971, 1984], Ekman and Friesen [1978], Ekman et al. [1972, 2002]). The FACS describes facial expressions as the combination of facial basic movements called Action Units (AUs). According to this theory there are 46 different AUs, 35 of which corresponds to muscular actions (e.g. AU1: inner brow raiser corresponding to *Frontalis - pars medialis* muscle in figure 4.5) and 11 to

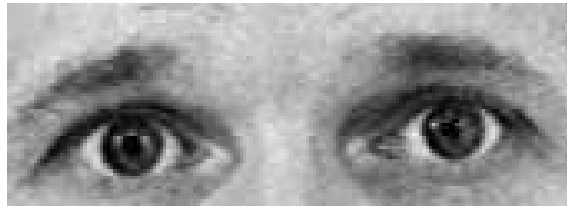


Figure 4.5: AU1: inner brow raiser (*Frontalis - pars medialis*)



Figure 4.6: AU18: lip puckerer (*Incisivii labii superioris* and *Incisivii labii inferioris*)

more complex movements (e.g. AU 18: lip puckerer corresponding to the *Incisivii labii superioris* and *Incisivii labii inferioris* muscles as in figure 4.6).



Figure 4.7: Action Units Example (AU10, AU15, AU17, combination of the three)

Even if AUs can be referred to more muscles, according to Ekman and Friesen, they represent minimal actions and cannot be divided into smaller or simpler actions. In other words a person cannot activate a part of the group of muscles implied in one AU. For similar reasons two or more AUs can refer to the same muscles that can therefore be activated in different group and different ways to show different AUs.

Combining the different action units, it is possible to represent every possible human facial expression. Ekman et al. [1982] also defines 6 basic emotions which have similar representation all around the world. Those emotions are commonly known as universal emotions and they are: happiness, fear, sadness, disgust, anger and surprise (see figure 4.8).

#### 4.2.2.3 MPEG-4 FAPs

The MPEG-4 standard defines a set of facial animation parameters (FAP) in order to represent facial expressions. Facial action parameters essentially indicate translations of the corresponding feature points (FPs) with respect to their position in the neutral face. Feature points represent key-points of a human face, like the corner of the mouth or the tip of the nose. These points can be found illustrated in figure 4.9.



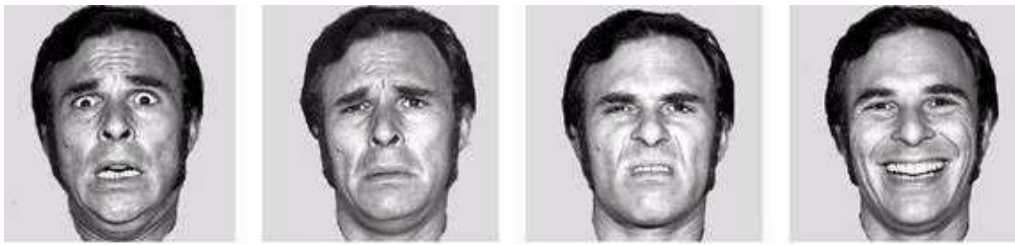


Figure 4.8: Four facial expressions from Paul Ekman: a) fear; b) sadness; c) disgust; d) happiness

MPEG-4 defines two high-level FAPs, visemes and expressions, and 66 low-level FAPs. Visemes and expressions are predefined and complex set of low-level FAPs.

Visemes are used to do lip synchronization with the speech and basically define the facial expressions (mostly regarding the mouth) related to each phoneme (the unit of speech signal).

Expressions represent a mixture of two out of the six basic expressions (Joy, Sadness, Anger, Fear, Disgust and Surprise).

Low level FAPs values are defined by six face animation parameter units (FAPUs). FAPUs are the distances between the major facial feature points on the model in its neutral state as shown in figure 4.10.

Furthermore the standard defines a facial animation table (FAT). FATs are an improvement of facial action parameter and specify the actual set of vertexes of the 3D model to be affected for each FAP. This is a way to gain even more control on the modeling.

Loosely speaking MPEG-4 have 3 level of control: high level FAPs, that are equivalent to the expressions of the agent, the low level FAPs that are more or less equivalent to Ekman AUs and finally the FAT which defines better how the FAP have to be represented in a specified model and specify, therefore, how the AUs are designed in the given face model.

Albeit the FACS system has been designed to represent human facial movements and the MPEG-4 FAP standard to control virtual 3D faces nothing actually prevent from converting the one representation to the other. Indeed, all FACS action units can be converted to MPEG-4 representation while all believable MPEG-4 facial movements can be coded in term of AUs.

If one wants to apply these models to robotic platforms some considerations need to be taken in account. In particular, robotic platforms generally are not capable of performing all human facial movements.

In the next few sections we will overview some of the most relevant technologies for the display of emotional expressions. Starting with next chapter we will describe the conversion of a psychological human theory of facial expression to the processing phases which can be used to animate both a virtual character and a simple expressive robot. We will see then, how the expressive limitations of an avatar and a robot can affect the quality of the resulting facial expressions.

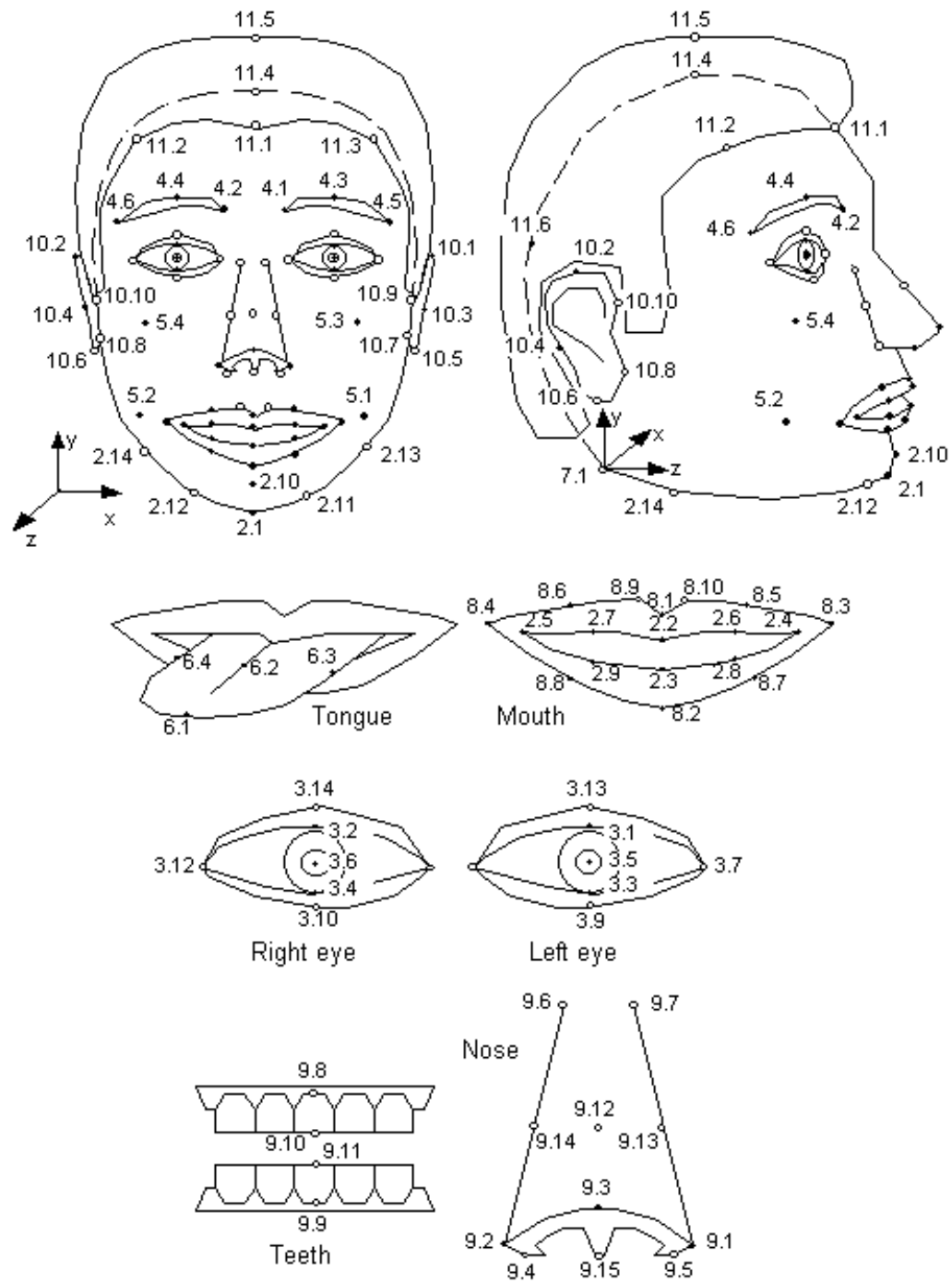


Figure 4.9: Feature Point (FP) in the neutral face



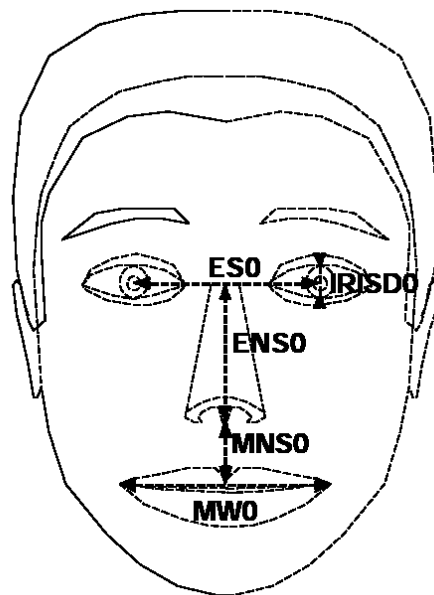


Figure 4.10: Facial Animation Parameter Units (FAPU)

#### 4.2.2.4 3D Graphics Animation Engines

There exist a number of animation engines which are able to display emotional facial expressions (see Bates [1994], Cassell [2000], Prendinger and Ishizuka [2004]). In this section we briefly overview some of them, which we consider the most relevant to our work<sup>1</sup> and which better exemplify the capabilities of current technologies. The interested reader could find other examples by referring to Prendinger and Ishizuka [2004], Cassell [2000], Pelachaud et al. [2007], Prendinger et al. [2008], Ruttkay et al. [2009].

#### 4.2.2.5 GRETA

Greta is a complete agent architecture developed for the MagiCster project by Pelachaud, De Rosi, De Bernardis, Poggi and others (de Rosi et al. [2003], Poggi et al. [2005]). The aim of the project was to create an embodied agent capable of interacting in a believable way with users. In particular the three main research subjects regard animation, conversation and information delivery.

While the main focus of this research is Greta's mind above its 3D graphic, great effort was put into the generation of believable animation and good synchronization between animation and speech. Furthermore, GRETA also takes advantage of gestures, gaze, and pose to help communications and to express emotional information.

The communicative meanings are associated with dialog moves by a markup language called APML: Affective Presentation Markup Language (see de Carolis et al. [2004]). APML allows researchers to tag the agent's dialogs with meaning level information such as the function of the communication and the emotions felt and displayed. It is interesting to notice that APML is such that "body" and "mind" can be two complete separated modules; changing the body of the agent (and its expressive capabilities) does not affect

<sup>1</sup>In particular, our focus will be in overviewing the facial animation engines which includes complex emotional facial expressions and gives direct mean of controlling them.

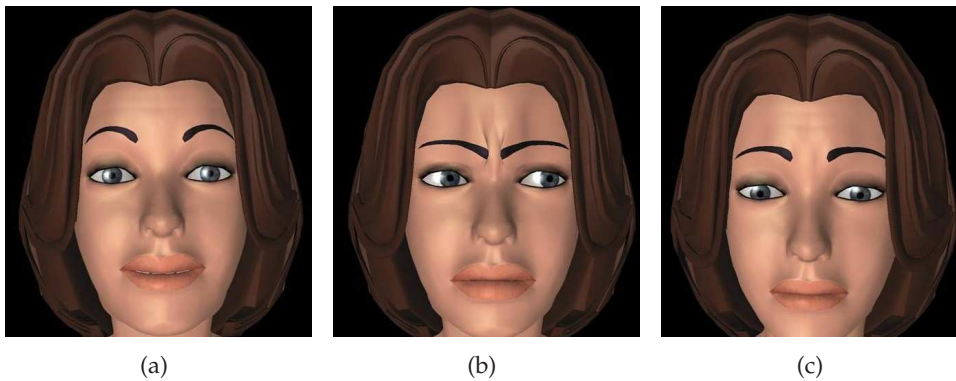


Figure 4.11: Three expressions of the Greta agent: a) Joy, b) Anger, & c) Sadness

the production of the APML code. During a conversation the mind decide what to communicate and describe it with APML; the body read, parse APML and finally renders the message with all the communicative channels it can employ automatically combining the available facial expression channels (gaze direction, eyebrow shape, head direction, movement etc.), vocal prosody, and body gestures. The 3-D face model respects the MPEG-4 standards.

Some of GRETA's communicative acts are directly linked to output channels (head direction, gaze etc.) via a belief network (BN). Figure 4.11 illustrates some facial expressions of the Greta agent.

Greta agent uses a facial description language above the facial action parameter (FAP) level to enrich the expressions that are associated with a given communicative function. The language includes facial basis (FB) that are basic facial movement, such as raised eyebrow, and facial displays (FD) that are made from one ore more FBs using operators '+' and '\*'.

#### 4.2.2.6 X-Face

Xface is freeware by Balci [2004], Balci et al. [2007]. The character is controlled through APML language (see 4.2.2.5). The agent can be controlled in a believable way thanks to some simple controls but more complex level of control are available.

In particular, some emotions are already available but it is possible to implement others through a FDP/FAPU editor which is included in the the XfaceClient module.

This agent is particularly interesting for our purposes since it gives full control over the displayed facial expressions of a believable character, it includes a set of tools for the design of new facial expressions and can be controlled in an intelligent way through the APML language. Nevertheless, when we made our first experiments in this domain this technology presented some major limitations.

Firstly, the APML to FAP converter was third-part and it was, therefore, not available to end users. Secondly, the software presented small problems with lip synchronization. The lip synchronization problem appeared because speech synthesis was not included in the software. That version of Xface used a wav file as speech input and, therefore, lacked of the information about the pronounced phonemes. The agent mouth opens and closes with respect to the energy of the audio signal only. This behavior is far from being sufficient when interacting with 3D virtual characters.



Figure 4.12: An avatar developed with X-Face

This second limitation is also limiting the interest in using it for building agents since it limits its possibilities to interact freely with the user. Every sentence that the agent want to say need to be recorded off-line possibly with emotional intentions already embedded.

#### 4.2.2.7 Galatea

Galatea is a sophisticated toolkit for dialog based human computer interaction. It includes software for speech recognition and a speech synthesis. The 3D face is done by mapping a simple 2D photo on a 3D face model.

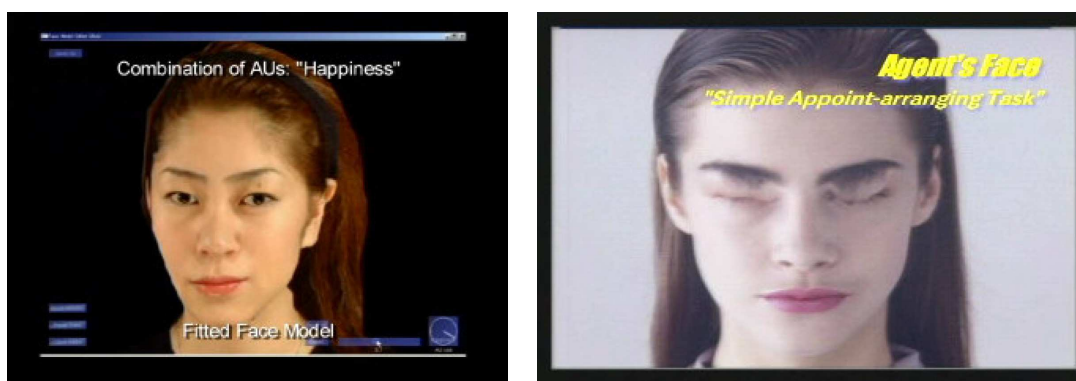


Figure 4.13: An avatar developed with Galatea

One advantage of Galatea is the customizability of the emotions: the control is fine and complete; users can decide to move a single group of muscles or deal with complete facial expression. Because the 3D model is extrapolated by a photo, still images of the face look more believable than other toolkits but Galatea can only clone existing characters .

The illusion of photo realism disappears in the video streams because a number of flaws in the modeling. Firstly, the eyes are immovable, and this gives the character an

unbelievable, machine-like appearance. Secondly, teeth are mapped from a simpler texture than the character photo and they do not fit in an harmonious way with the general appearance of the Galatea character. Thirdly, when the avatar is blinking his eyes the texture/photo is stretched to close the eyes. This deforms the textures above the eyes making the appearance unbelievable (see figure 4.13).

To conclude, there are a number of interesting points that would encourage the use of this software, but the problems mentioned above decrease the level of believability and should be solved to take full advantage of Galatea's features.

#### 4.2.2.8 Haptek

Haptek avatars (shown in figure 4.14) have been developed to represent believable human faces (Haptek [2006]). Haptek tools are commercial and avatars can easily be inserted into applications or web pages. Haptek animation is based on a dedicated technology, similar to MPEG-4 FAP (Facial Action Parameters).



Figure 4.14: Cherry, an avatar developed with Haptek. Lisetti et al. [2002]

There are different levels of control over Haptek avatars: from the control over expressions, morph and position of the avatar to the control of basic facial movements.

Basic control of the avatar comes from *Haptek hypertextt* technology. Through hypertext one can easily control text to speech, avatar position and launch the so called Haptek *switches*. Switches allow a control similar to the high level FAPs control over MPEG-4 FAPs (i.e. expressions and visemes); through Hypertext one can actually control general behavior and facial expressions of the avatar.

Switches are collections of *states* which represent still expressions of the avatar in term of combinations of *facial parameters* defined by Haptek.

Haptek facial parameters are similar to Ekman AUs or MPEG-4 low level FAPs and are divided into 15 viseme parameters, 23 parameters regarding eyes, 6 regarding head movement, 15 more regarding the mouth, 6 regarding the nose, and 2 regarding the checks. Few more parameters control sub expressions, and morphs (the character can actually shape to anthropomorphized animals and objects).

Through switches one can control the evolution of states over time as well as the softness of the passages from one state to another, i.e. the evolution of the avatar expression.

---

---

Therefore, switches are more or less equivalent to high level MPEG-4 FAPs while Haptex parameters are more or less equivalent to low level FAPs. It is not possible, according to what we know about Haptex technology, to have a level of control similar to the one allowed in MPEG-4 by FATs. FAT like information is embedded in the code and cannot, to our knowledge, be modified.

#### 4.2.2.9 Robotic Platforms

Traditionally, autonomous robots are designed to operate as independently and remotely as possible from humans, often performing tasks in hazardous and hostile environments. However, a new range of application domains (domestic, entertainment, health care, etc.) are driving the development of robots that can interact and cooperate with people, and play a part in their daily lives.

Humanoid robots are arguably well suited to this. Sharing a similar morphology, they can communicate in a manner that supports the natural communication modalities of humans. Examples include facial expression, body posture, gesture, gaze direction, and voice.

In this section we will briefly overview some of the existing technologies which can be used for human-robot-interactions and which support emotional facial expressions as a communicative mean.

#### 4.2.2.10 AIBO

AIBO is a robot developed by SONY starting in the early 90' (SONY Entertainment Robots [2006]). AIBO is designed for human-robot-interactions and has the shape of a dog, the humans' best friend (see figure 4.15). AIBO's target application is user entertainment but it could also be used to automatically snap pictures of the home environment for security applications.



Figure 4.15: SONY AIBO

AIBO features, touch sensors, face recognition, emotional behavior, and artificial intelligence which allow it to learn from its environment. SONY states that *"AIBO has instincts to move around, look for its toys, satisfy its curiosity, play and communicate with its owner. AIBO's personality develops by interacting with people. AIBO develops according to its*

---

*different experiences. The more you interact with it, the more it learns*". Starting from the ERS-110 generation back in 1999 SONY decided to integrate the ability to express emotions in the AIBO robots (see figure 4.16 and Sony Entertainment Robots [2005]).



Figure 4.16: AIBO ERS-7 Illume-Face expressions

#### 4.2.2.11 Papero

The PaPeRo by NEC (Corporation [2005]) has been researched and developed since 2001 with the intention of being a partner for human-robot-interactions. For this reason, it has various basic functions for the purpose of interacting with people.



Figure 4.17: NEC's Papero

In particular PaPeRo features: facial recognition technology to find and identify people in its environment, speech recognition technology to interact with humans, human-like behaviors, emotional reactions, spontaneous suggestions, and autonomous actions to facilitate interactions with people. PaPeRo present different personalities such as the leadership, the lazy, or the knowledgeable ones. PaPeRo is designed to engage the user and help him with everyday tasks; for example, it could be used to know the weather forecast, reading news, keep the user's calendar, or play with users.

#### 4.2.2.12 Kismet

Kismet is a humanoid robot developed at MIT by Breazeal [2004]. Kismet has been designed to engage humans in social and natural face-to-face communications. His intelligence is inspired by infant social development and psychology.

Kismet can enter into natural and intuitive social interaction with human caregivers and learn from them, being reminiscent of parent-infant exchanges. To do this, Kismet



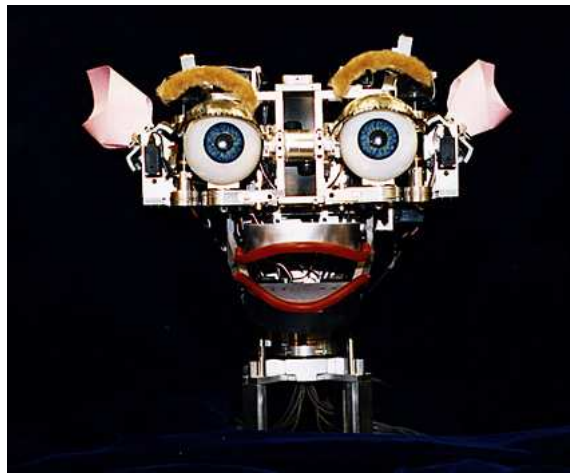


Figure 4.18: MIT's Kismet

perceives a variety of natural social cues from visual and auditory channels, and delivers social signals through gaze direction, facial expression, body posture, and vocal babbles.

The robot has been designed to support several social cues and skills that play a role in socially situated learning with a human instructor.

#### 4.2.2.13 Philips iCat

iCat is a robot designed by Philip Research to be integrated in home environments (van Breemen [2005]). It look like a cat to facilitate the social integration in the domestic environment (see figure 4.19).

iCat has been created for social robotics, human–robot interaction and collaboration, joint–attention, gaming, and ambient–intelligence. One typical task for the iCat would be to be connected to an in-home network, to connect to the internet, and to control the different home devices (ubiquitous computing).

Some simple applications have been designed to test the platform: for example iCat has been interfaced to a DVD recorder being able to connect to the internet suggesting TV programs and to record them and it has also been used as a player in simple gaming scenarios (Saini et al. [2005], Bartneck et al. [2004], van Breemen [2004a]). In one of our experiments two different iCat personalities were compared while the iCat was autonomously interacting with users during a checkers game session.

iCat has thirteen motors for facial expressions and body control (figure 4.19). These servos allow to control these parts : eyebrows, eyelids, eyes, lips, body and head. Four touch sensors are located in feet and ears . Multicolor LEDs can be used to express different states such as: listening, sleeping, or waking up. A webcam is hidden in his nose; speakers and microphones are integrated in its paws.

Albeit iCat has expressive capabilities, compared to humans or even to 3D embodied conversational agents (ECA) they are very limited. Philips research has created some default facial expressions for iCat using principles of animations defined for Disney's characters (see figure 4.20 and van Breemen [2004b]).

In this part of the thesis we will see how the iCat can be controlled for the display of facial expressions generated with the use of Scherer [1984] CPT of emotions. We will





Figure 4.19: Philips iCat

apply the same theory to the generation of facial expressions for an avatar constructed with Haptik technology.

In this section we have overviewed some of the embodiment platform which can be used to display emotional facial expressions. Next section will briefly overview engines which can be used to express emotions through vocal prosody.

### 4.2.3 Emotional Speech

We have said that speech is the second most obvious source of emotional information in human communications. In human–computer and human–machine interactions vocal prosody and emotional vocal expressions have been extensively studied because the more and more spread use of computer–generated voices (train stations, airports, automatic phone answering machines, etc.) demonstrated that non–emotional, “flat” voices are not only unnatural, but also often annoying if listened for more than few seconds. In this section we will briefly describe the speech production phenomenon and overview few existing technologies which are commonly used for the generation of emotional speech.

#### 4.2.3.1 Speech Models

There exist several systems to model and characterize speech production and speech samples. It is not the intention of this document to go into details about speech representations. In the next few sections we will, nevertheless, overview some of the most well known voice representation. The interested reader is invited to make reference to the

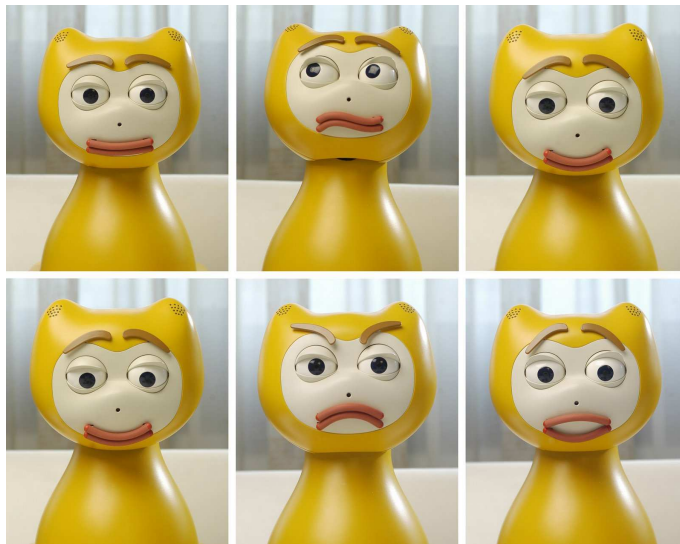


Figure 4.20: Some of Philips iCat default facial expressions (neutral, disgusted, happy, elated, angry, and surprised)

works of Scherer [2003], Scherer and Bänziger [2004], Paulmann and Kotz [2008], Clavel and Richard [2009]

Before explaining how human speech can be modeled it is important to know how it is produced by the different configurations of tongue, lips, jaw, and other speech organs (see figure 4.21). Sounds are variations in air pressure detectable by the human ear.

In human speech there exist two big families of sounds: the voiced ones (e.g. vowel sounds) and the unvoiced sounds (e.g. f, t, r, and s sounds). The former ones are initially produced by a vibration of the vocal cords (which represent the fundamental frequency of the sound) and modified in their spectrum<sup>2</sup> by the shape of the vocal tract (see figure 4.22). The latter ones are produced as a noisy sound which can be either assimilated to white noise as in the fricative sounds f or s or as a plosive sound as in p or t. In reality spectrum are much more complex as no sound is perfectly voiced and no sound is perfectly unvoiced (see figure 4.23 for a typical speech spectrum)

There exist various mathematical characteristics to model the phenomena involved in speech productions. Here we list some of the most used features and try to briefly explain them:

- **Prosodic Features:**

- **Pitch:** the pitch represent the fundamental frequency of the voiced sounds (see figure 4.22(a)).
- **Jitter:** the jitter is a measure of the oscillation of the pitch around its central frequency.
- **Formants:** formants represent the frequency shaping the spectrum of the vocal sound (e.g. F1 and F2 in figure 4.22(b)).

---

<sup>2</sup>The spectrum represents the distribution of the energy on the frequency scale.

---

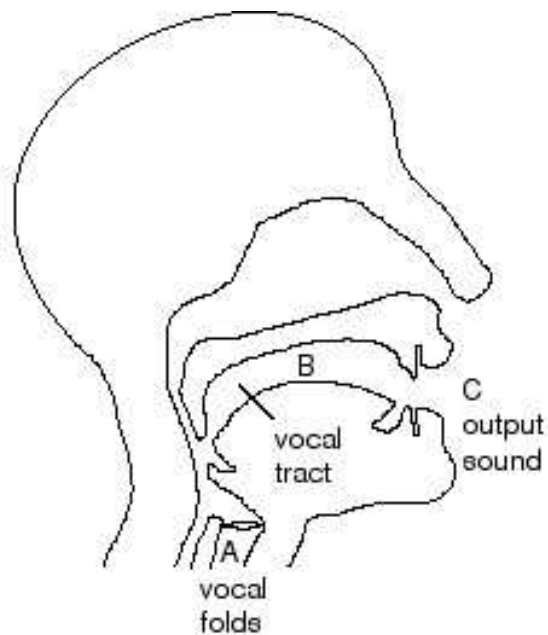


Figure 4.21: Model of the vocal tract

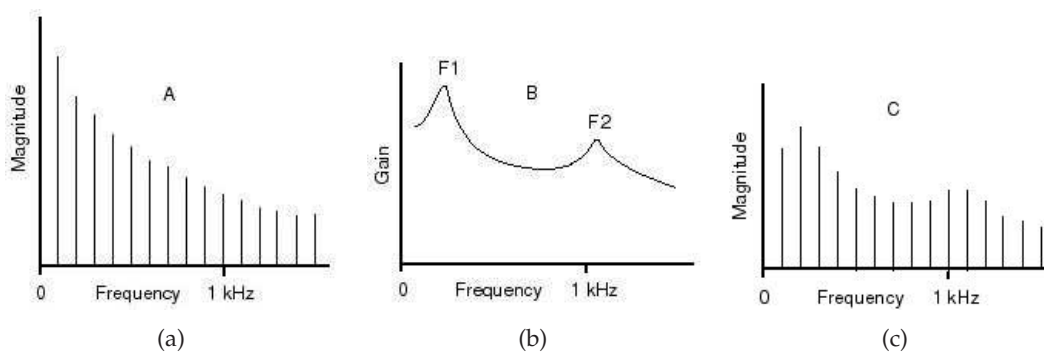


Figure 4.22: Speech Production for voiced sounds: a) periodic signal, b) vocal tract filter / spectral envelope, & c) resulting signal

- **Spectral Envelope:** the spectral envelope represent the shape of the filter characterizing the vocal tract (i.e. B in figure 4.21; figure 4.22(b)).

- **Energy Features:**

- **Energy:** energy represented the total energy of the signal (i.e. his loudness).
- **Shimmer:** the shimmer represent a measure of the oscillation of the energy of the signal around is central value.
- **Harmonicity:** harmonicity (a.k.a. Harmonic-to-Noise Ratio (HNR)) represent the ratio between the energy of the harmonic part of the signal and the energy of the “noisy” part of the signal.

- **Rhythm Related Features:**

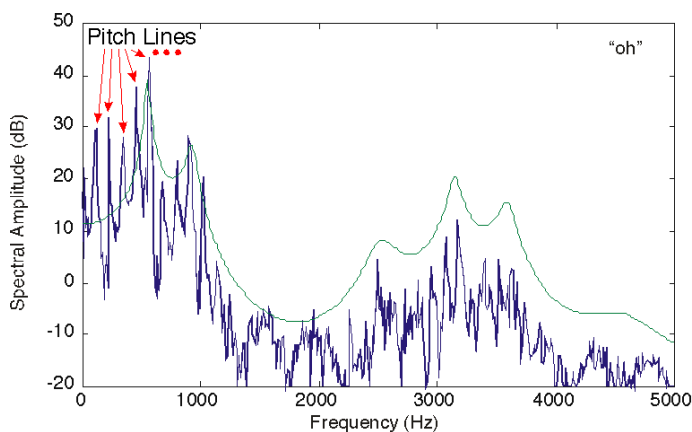


Figure 4.23: Typical speech spectrum: in red the fundamental frequency (pitch) lines (see figure 4.22(a)), in green the estimated spectral envelope(see figure 4.22(b))

- **Speech Rate:** speech rate is a measure of the number of speech units which are vocalized in the time unit.
- **Pause Structure:** with pausing structure it is intended the statistical analysis of the pauses lengths and distribution.
- **Linear Predictive Coding:** linear predictive coding (LPC) analyzes the speech signal by *estimating the formants*, removing their effects from the speech signal, and estimating the intensity and frequency of the pitch<sup>3</sup>. The numbers which describe the intensity and frequency of the pitch, the formants, and the residue signal are called LPC coefficients can be used to represent, and synthesize the voice.
- **Mel-Frequency Cepstral Coefficients:** mel-frequency cepstral coefficients (MFCC) represent the signal using a non-linear *Mel spectral scale*. To extract these coefficients is necessary to analyze the *cepstrum* (the spectrum of the spectrum) of the signal. The representation of the signal by the MFCC is commonly thought to be more efficient than the LPC one.

Please see Scherer [2003] and Burkhardt [2009] for a review of existing techniques for representing and modeling emotional speech.

### 4.2.3.2 Emotional Speech Engines

#### 4.2.3.3 Festival

Festival is a general research framework for building speech synthesis systems initially developed by Black and Taylor [1997]. Festival is able to synthesize emotional speech Hofer et al. [2005]. This is the system used by the virtual agent Greta (see section 4.2.2.5).

Festival is a framework; different emotional voice can, therefore, being developed according to the actual implementation but usually they are limited to fundamental emotional categories. A rule-based approach is used to determine which signal parameters should be changed to communicate the intended emotion.

<sup>3</sup>LPC considers the signal to be stationary on small enough amount of time.

#### 4.2.3.4 Loquendo

Loquendo is a commercially available TTS software (Zovato et al. [2004]). The software allows to change emotional intention via a rule-based approach. The MBROLA data is used (as it is for the other two systems) to improve the quality of the synthesis.

#### 4.2.3.5 Mary

Mary is a text-to-speech (TTS) synthesis system for German, English and Tibetan developed at the German Institute for Artificial Intelligence (DFKI) by Schröder and Trouvain [2003].

Four parts of the TTS system can be distinguished: 1. the preprocessing or text normalization; 2. the natural language processing, doing linguistic analysis and annotation; 3. the calculation of acoustic parameters, which translates the linguistically annotated symbolic structure into a table containing only physically relevant parameters; 4. and the synthesis, transforming the parameter table into an audio file.

Mary can display emotional intention in speech in term of continuous emotion dimensions (i.e. activation, evaluation and power) as in section 3.3.2.2.

### 4.3 Building Believable Facial Expressions

We have overviewed in the previous chapters some of the technologies which can be used for the display of facial expressions. We have seen how Scherer's component process theory of emotions allow to generate dynamic facial expressions from appraisal patterns.

To better clarify the process of generation of the facial expressions let us give the example of the discrete emotion "happiness". For this emotion we have (table 3.3) *low* novelty which translates (table 4.1) in alternatively AUs 1, 2 & 5 or AUs 4, 7, 26; for the intensities all AUs will have the value 'b' that on Ekman's scale from 'a' to 'e' is a *low* value. Intrinsic pleasantness has high intensity: that translates into AUs 5, 26 & 38 or AUs 12, 25; the intensity is high and therefore 'd'. We replicate the same process for all the different components and one of the possible sequences of AUs (and intensities) we obtain is the following:

- Novelty: AU1b, AU2b, AU5b
- Intrinsic Pleasantness: AU12d, AU25d
- Conduciveness: AU12c, AU25c

As it can be already seen from this simple example it is not straightforward to automatically convert a SEC sequence to a complete dynamic facial expression. We have identified four main steps constituting this process:

1. convert each SEC to AUs;
  2. convert AUs to platform specific parameters;
  3. find appropriate intensities;
  4. exploit the temporal and intra-SEC correlation adapting AUs intensities.
-

SEC	AUs
Novelty	1, 2 & 5
	4, 7, 26 & 38
Pleasantness	5, 26 & 38
	12 & 25
Unpleasantness	4, 7, 9, 10, 15, 17, 24 & 39
	16, 19, 25 & 26
Goal-Need Conduciveness (discrepant)	4, 7, 17 & 23
Coping Potential & No Control	15, 25, 26, 41 & 43 (if tears 1 & 4)
Coping Potential, Control & High Power	4 & 5
	7, 23 & 25
	23, 24 & 38
Coping Potential, Control & Low Power	1, 2, 5, 20, 26 & 38

Table 4.1: conversion from SECs to AUs predictions (derived from Scherer [2001])

In the following section 4.3.1 we will detail the process of generation of facial expressions for *Cherry*, the avatar from Haptেক. Then, in section 4.3.2 we will detail the process involving the Philips iCat that, in our lab, we called *Cleo*. Please note that this work has been developed with the collaboration of Amandine Grizard, who has been a valuable Ph.D. colleague of ours for about a year.

### 4.3.1 Cherry, the Haptেক Avatar

#### 4.3.1.1 Step 1 - Converting SECs to AUs

For this step we used the tables given in Scherer [2001] (see table 4.1).

#### 4.3.1.2 Step 2 - Converting AUs to Haptেক parameters

The second step is about converting AUs to Haptেক parameters. Although Haptেক parameters are quite similar to MPEG-4 FAPs and Ekman AUs, there is no predefined one-to-one mapping among those different representations. We have therefore, created a conversion table (see Table 4.2) and a software designed to create switches representing single AUs at different intensities.

One issue arises when trying to convert AUs regarding mouth movements in Haptেক parameters. Most of the available Haptেক parameters controlling the mouth area are in fact designed as visemes<sup>4</sup> and are therefore much more complex facial movements than AUs. This consideration leads to the conclusion that AUs representing the mouth area are very difficult to model (we would need some Haptেক sub-parameters). We designed the parameter combinations regarding difficult AUs to have the most similar facial movements with regards to the original ones. To evaluate the quality of the conversion please refer to table 4.3<sup>5</sup> and to appendix A.

<sup>4</sup>visemes are designed to represent the lip synchronization information, and describe therefore the mouth expression shown when pronouncing a phonemes (i.e. vocal sounds like th, a, etc.)

<sup>5</sup>Human images from <http://www.cs.cmu.edu/afs/cs/project/face/www/facs.htm>

AU ID	Haptek parameter	$I_{max}$	Haptek parameter	$I_{max}$	Haptek parameter	$I_{max}$	Haptek parameter	$I_{max}$
1	MidBrowUD	-1.3						
2	RBrowUD	-1	LBrowUD	-1				
4	MidBrowUD	1	LBrowUD	1	RBrowUD	1		
5	trust	-1						
6	smile3	0.5	lipcornerL3ty	-0.7	lipcornerR3ty	-0.7		
7	distrust	1						
9	nostrilL3ty	0.6	nostrilR3ty	0.6	nostrilL3tx	0.1	nostrilR3tx	-0.1
10	mouth2ty	0.2	ey	0.2	kiss	0.2		
11	nostrilL3tx	0.4	nostrilR3tx	-0.4	nostrilL3ty	0.4	nostrilR3ty	0.4
12	smile3	0.6	lipcornerL3ty	-0.8	lipcornerR3ty	-0.8		
13	smile3	0.6	lipcornerL3ty	-0.8	lipcornerR3ty	-0.8		
14	lipcornerL3ty	-0.3	lipcornerR3ty	-0.3				
15	lipcornerL3ty	-1.2	lipcornerR3ty	-1.2	mouth2ty	0.1		
16	th	0.4	ch	0.25				
17	lipcornerL3ty	-0.4	lipcornerR3ty	-0.4	mouth2ty	0.2		
18	kiss	1						
20	smile3	0.6	lipcornerL3ty	-0.9	lipcornerR3ty	-0.9		
22	ch	1						
23	kiss	0.75	b	0.65				
24	b	0.9	kiss	0.25				
25	ey	0.8						
26	aa	1.1						
27	aa	1.3	ey	1.2				
28	b	1.1						
38	nostrilL3tx	0.6	nostrilR3tx	-0.6	nostrilL3ty	0.3	nostrilR3ty	0.3
39	nostrilL3tx	-0.6	nostrilR3tx	0.6	nostrilL3ty	-0.3	nostrilR3ty	-0.3
41	trust	1						
42	blink	1						
43	blink	1	smile3	0.3				
44	l_lidL	-1	R_lidL	1	trust	0.8		
45	blink	1						
46R	winkR	1						
46L	winkL	1						
51	HeadTwist	-1	NeckTwist	-0.4				
52	HeadTwist	1	NeckTwist	0.4				
53	HeadForward	-0.2	HeadForward	0.2				
54	HeadForward	0.2	HeadForward	-0.2				
55	HeadSideBend	0.6	NeckSideBend	0.6				
56	HeadSideBend	-0.6	NeckSideBend	-0.6				
57	HeadForward	1	HeadForward	-1				
58	HeadForward	-1	HeadForward	1				
61	LEyeBallLR	1	REyeBallLR	1				
62	LEyeBallLR	-1	REyeBallLR	-1				
63	LEyeBallUD	-1.4	REyeBallUD	-1.4				
64	LEyeBallUD	1.4	REyeBallUD	1.4				

Table 4.2: AUs to Haptek parameter conversion



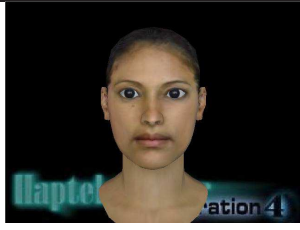



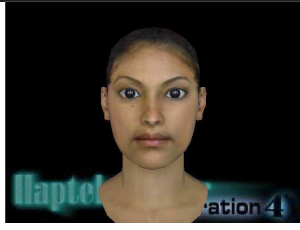


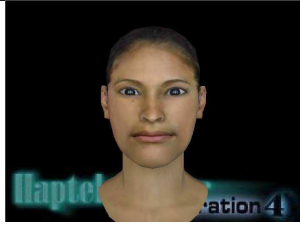





AU	Hapttek (whole face)	Hapttek	Human
			
Neutral Face			
1			
Inner Brow Raiser - Frontalis, pars medialis			
2			
Outer Brow Raiser - Frontalis, pars lateralis			
20			
Lip Stretcher - Risorius et platysma			
22			
Lip Funneler - Orbicularis oris			

Table 4.3: Hapttek showing some AUs compared to actors<sup>2</sup>

### 4.3.1.3 Step 3 - Finding the right intensities for the AUs

Based on Scherer [1982, 1985, 1987] theory of emotions we have converted SECs intensities (see table 3.4) to the Ekman scale from 'a' (min. intensity) to 'e' (max. intensity). The main problem is that Scherer gives one intensity for each SEC while we would need a different intensity for each AU. At first we have assigned the same intensity to all AUs belonging to the same SEC, then we have relaxed this constrain when necessary.

### 4.3.1.4 Step 4 - Exploiting temporal information

A problem in designing facial expressions with CPT is that Scherer's theory does not exploit any intra-SEC or temporal correlation. In other words SECs are considered as completely independent one from the other but predictions are not. According to Scherer theory SEC sub-expressions should be additive: the AUs involved in the *novelty* SEC should be activated when *novelty* appraisal occurs and last over the entire expression; the AUs involved in the *pleasantness* check are activated with the appraisal and their effect should sum to the AUs related with novelty.

This approach presents two main limitations which may be made even more evident when this theory is applied to a virtual agent such as Cherry:

- two SEC predictions in the same expressions may involve the activation of the same AUs (e.g. both novelty and pleasantness can activate AUs 4 and 7);
- two different AUs may involve the activation of muscles belonging to the same region (e.g. AU 1 - Inner brow raiser, AU 2 - Outer brow raiser and AU 4 - brow lowerer all involve muscles in the same region)

If in human beings, concurrent activations of the same muscles is simply solved, in virtual agents two (contrasting or not) activations of the same regions are not always correctly handled by the animation engines. When designing facial expressions on virtual characters we need, therefore, to pay particular attentions to these kind of issues.

Furthermore, because there exist multiple predictions for the same SEC, when we try to build an expression there also exist different possible evolution of the facial expression as it can be observed in figure 4.24.

According to Scherer all evolutions are equally likely. We have asked four subjects (three Ph.D. candidates and a professor, 2 men and 2 women, aged  $29.6 \pm 9.5$  years) to evaluate the different facial expressions and simply rate them in term of believability (not believable, somewhat possible, believable). It resulted that some pattern were not believable.

A last problem is that Scherer does not give, to our knowledge, any information about the timings for each single sub-expression. In other words, we do not know how much shall every sub-expression last, if they should last all the same time, etc (see figure 4.25).

We referred to acted video from the emotion database "Interval" to extract hints about this missing information. These videos show one subject at a time displaying one out of five facial expressions (namely: happiness, fear, sadness, disgust, and anger) on a blue background. Two experts analyzed these videos frame-by-frame to extract information about the muscles activated during the process of generation of the facial expressions. We used the same videos to find solutions to the choice of the best pattern and to the problem of concurrent activation of the same facial muscles. From this analysis we have concluded that:

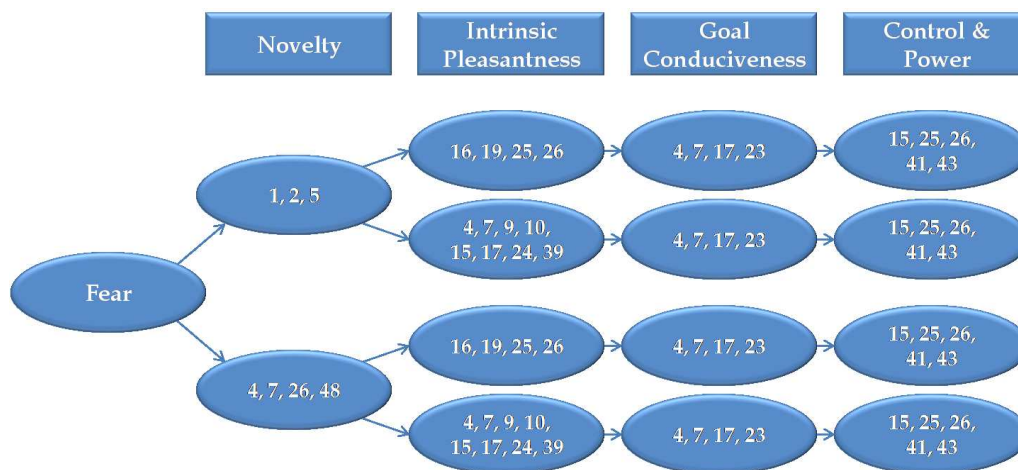


Figure 4.24: Multiple Prediction for a novel, unpleasant, non goal conducive, and non-controllable event (i.e. scary - fear)

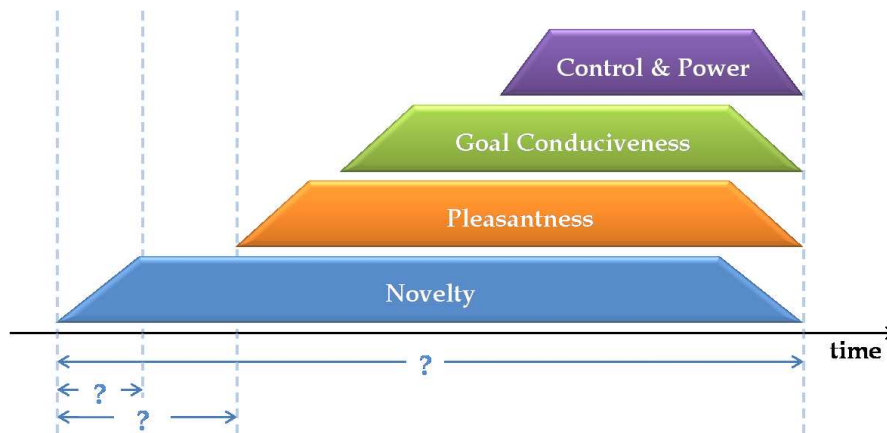


Figure 4.25: Theoretic evolution of SEC related sub-expressions from Scherer [2001] (missing information about the timing of the sub-expressions)

- When possible it is suitable to select contiguous SEC predictions to minimize the muscular effort (i.e. predictions which have the highest possible number of AU in common).
- Depending on the displayed emotion, on the subject, and other factors that we were not able to identify (e.g. day to day dependence) the SEC timing shall be adjusted<sup>6</sup>.
- Generally different SEC have different timings.
- To obtain similar result to these acted emotional expressions AUs involved in one SEC should gently fade away possibly leading to final facial expressions in which the AUs activated for the novelty or pleasantness check are no more perceivable (see figure 4.26).

<sup>6</sup>For example when displaying fear some subject apparently activated abruptly all AUs almost at the same time. In this case the novelty check was almost undetectable. The same check would instead be much more evident in some display of the emotion happiness.

- To further smooth the facial expressions attention should be given to AUs advocating the activation of the same facial muscles.

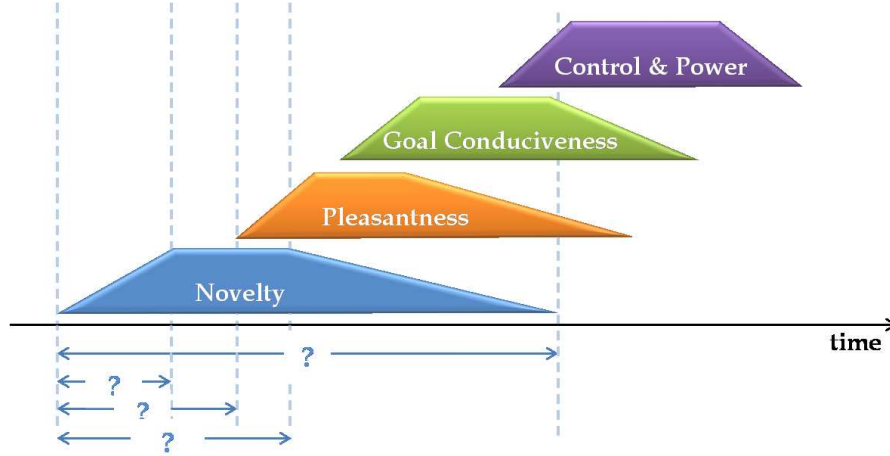


Figure 4.26: Extrapolated evolution of SEC related sub-expressions

#### 4.3.1.5 Implementation details

We have developed two pieces of code to easily generate and test the facial expressions.

The *Cherry Expression Generator* (CEG) in figure 4.27 is a piece of C++ MFC code which can be used to generate Hapttek switches. Switches generated with the CEG are composed of 5 different states: one for each involved SEC (novelty, pleasantness, goal-conduciveness, and the set of control & power) plus a concluding one which can be used to further smooth the facial expression. For each state a user can define 8 pairs of AU and intensity and a ending time. The nature of the Hapttek technology is such that it is very difficult to express two states at the same time (not like in figures 4.25 and 4.26). The desired fading effect (see figure 4.26) has therefore to be obtained by manually inserting AUs of the previous SEC in the current one.

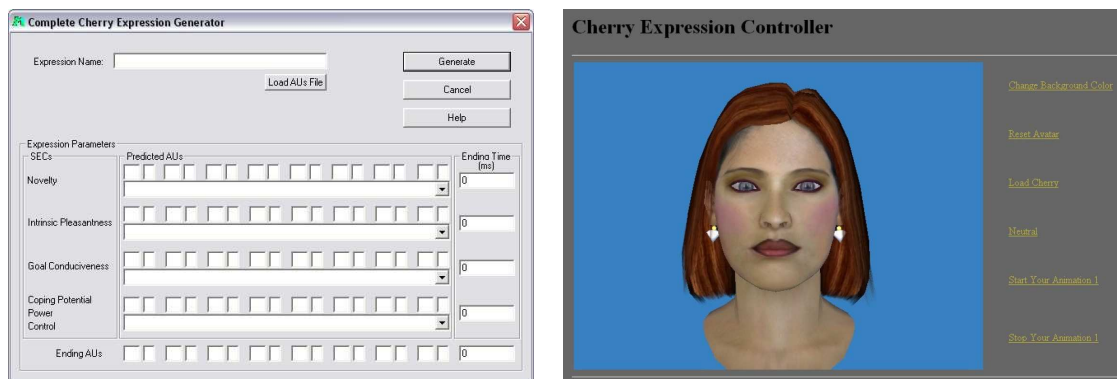


Figure 4.27: Cherry Expression Generator (CEG) and Controller (CEC) interfaces

The *Cherry Expression Controller* is a simple HTML page including the agent and the Hapttek hypertext necessary to launch switch. An user shall manually change the HTML code by inserting the name of the newly created switch.

---

The resulting transitions for five emotions *fear*, *sadness*, *disgust*, *anger*, and *happiness* are shown in figures 4.28-4.32

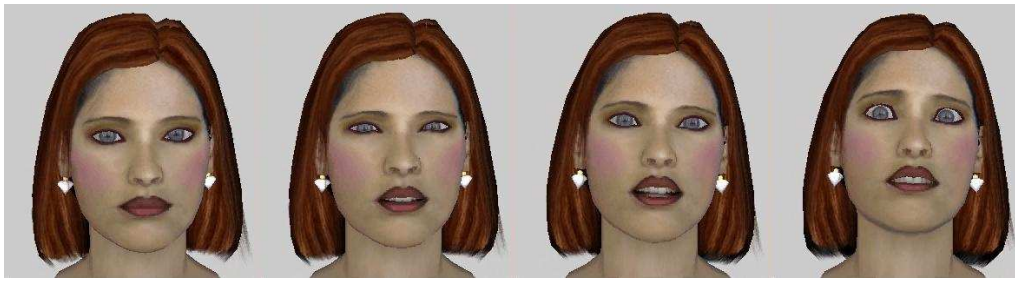


Figure 4.28: One possible evolution of the expression Fear using CPT



Figure 4.29: One possible evolution of the expression Sadness using CPT



Figure 4.30: One possible evolution of the expression Disgust using CPT

In this section we have showed how to build emotional facial expressions within Scherer's CPT using an avatar by Haptik. It is important to know that all the facial movement which we have built for the facial expressions goes to add upon a layer of small face and head movement which are randomly set and which give liveness to the character. In this scenario, these random facial movement also take into account small asymmetries of the facial expressions which can also be noticed in figure 4.8.

In section 4.3.2 we will describe a similar processing applied to the iCat Robot. Finally, in the next chapter 4.4 we will evaluate the quality of the designed expressions with some simple user studies and comment the results.

---



### 4.3.2 Cleo, the Philips Robot

Cleo is built upon a robotic platform and, as such, it has much more limited expressive capabilities than an avatar. Nevertheless, with respect to humans and avatars it has one expressive channel more: this is represented by the multicolor LED. In this section, we will describe how the four phases introduced at the beginning of this chapter can be adapted for the generation of facial expressions for this second platform.

#### 4.3.2.1 Step 1 - Converting SECs to AUs

The first phase of generating facial expressions for the iCat is always the conversion between SECs and AUs. Here again the conversion is done thanks to table 4.1.

#### 4.3.2.2 Step 2 - Converting AU to iCat servo controls

The second step consists in representing Ekman [1971] AU on the iCat robotic platform. If a human has around 100 different muscles to represent his facial expressions on iCat we are limited to 12 servos (one for each eyebrow, four for the two eyes, two for the eyelids, and four for the mouth). Because of this limitation the same servo (and therefore movement) has, for example, to be used to express AU2 (outer brow raise) and AU4 (brow lowered).

Table 4.4 illustrates the extrapolated Action Units we did implement on the iCat. For each AU we attributed arbitrary values to servos to map the movement with the five Ekman AU intensities: very low (a), low (b), medium (c), high (d), very high (e).

#### 4.3.2.3 Step 3 - Finding the right intensities for the AUs

Based on Scherer [1982, 1985, 1987] theory we have converted SECs intensities (see table 3.4) to the Ekman scale from 'a' (min. intensity) to 'e' (max. intensity). As we have seen in the previous section, the main problem is that Scherer gives one intensity for each SEC while we would need a different intensity for each AU. To find a solution to this issue we have, at first, assigned the same intensity to all AUs belonging to the same SEC and then relaxed this constraint if necessary.

#### 4.3.2.4 Step 4 - Exploiting temporal information

Starting from the same observations made for the Haptik avatar (see section 4.3.1.4 we have developed the facial expressions trying to fuse the different SEC and to make AU gently fade. Furthermore, when possible we have selected contiguous SEC predictions to have maximize the number of common AUs. Because the iCat is more similar to a cartoon character than to a human we thought about using the same cartoon animation principles which inspire Disney's cartoon and the iCat original facial expressions (van Breemen [2004b]). We have, for example, exaggerated the facial expressions by increasing the intensities of the AUs and added red LED light to the ears of the iCat to express *anger*. Red is a color often associated to the anger and it reflect the increased blood flow to head and chest (Scherer [2001]) typical of the "control and high power" check. Similarly, we have decided to turn on and off the LED in a pseudo-randomic manner to represent the confusion elicited by the *fear* emotion and to use harmonious light illumination pattern that may remind the ones of a winning slot-machine for the emotion happiness.

---

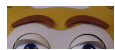
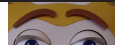
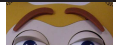
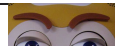
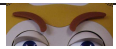
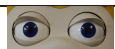
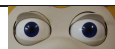
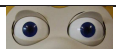







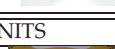
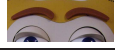
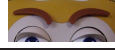
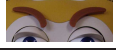
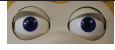
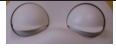





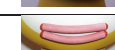

Action Units	FACS Name	Neutral Example	Medium Example	Very High Example
POSSIBLE ACTION UNITS				
AU1	Inner Brow Raise			
AU2	Outer Brow raise			
AU5	Upper Lid Raiser			
AU10	Upper Lip Raiser			
AU12	Lip Corner Puller			
AU15	Lip Corner Depressor			
AU16	Lower Lip Depressor			
AU25	Lips Part			
AU43	Eyes Closed			
EXTRAPOLATED ACTION UNITS				
AU4	Brow Lowered			
AU7	Lid Tightened			
AU20	Lip Stretcher			
AU22	Lip Funneler			
AU23	Lip Tightener			
AU24	Lip Pressor			
AU26	Jaw Drop			
AU41	Lid Droop			

Table 4.4: Extrapolated Action Units

#### 4.3.2.5 Implementation details

With this processing, we have created nine emotional facial expressions. These are: happiness, disgust, contempt, sadness, fear, anger, indifference, pride and shame.

To develop these expressions we have used the iCat Animation Editor by Philips (see figure 4.33). In this editor we can set independently the position and the evolution of each iCat servo. As we did for the virtual agent Cherry, also with Cleo we set a layer of small movement which gave the iCat more liveness. With the iCat, this was done by setting on a different parallel channel a cyclic, pseudo-random, animation.

In figures 4.34-4.42 we show the apex of the resulting facial expressions. Few expressions were more complex to model. In particular, some expressions have very similar (if not identical) sequential evaluation patterns<sup>7</sup> resulting in very similar facial expressions. Without more information about the context in which the agent is immerse these expressions are practically non-distinguishable. To tackle this issue, whenever possible we have also used other modalities such as head pose, gaze, and lights.

In the next chapter we will see if the results confirm this observation.

<sup>7</sup>We observe that not all SEC in Scherer's CPT have emotional expression predictions.



## 4.4 Results

### 4.4.1 Open Questions for Research in Psychology

In the previous chapter we have detailed the phases needed for the development of believable facial expressions within Scherer [1984] component process theory. In doing this, we have observed that some open research questions arise. We have, while possible discussed them in the previous sections. In this section we intend to put them all together.

- Are all SEC related sub-expressions equally important? If they are not should some of them last longer or be represented with more intense AUs?
- How should different sub-expressions fuse? Can we model the timings (durations, onsets, offsets) in some way (see figures 4.25 and 4.26)?
- Are AU related to the same sub-expression equally important? Should some AU last longer or have more intense representation?
- How should we chose among the different possible predictions (see figure 4.24)?
- How could we consider the effect of other affective phenomena such as mood, preferences, or personality into the development of facial expressions?

In this work we have decided to tackle most of these issues by manually analyzing some videos of actors performing our emotions through facial expressions. While the best solution would be to obtain answers to these questions by the component process theory, this approach revealed to be a valuable one. Another possible solution will be to build a machine learning system that would automatically extracts timing informations by analyzing statistics of the activated AUs in videos of people expressing the relative emotions. While this second system would have the drawback of needing a dedicated database and use some potentially complex algorithms, the results may be very interesting for both computer scientists trying to build facial expressions and psychologists such as Scherer trying to model the phenomena involved with emotional expressions.

### 4.4.2 User Study

In the previous section we have overviewed the set of open psychological research questions arisen from our work. While these questions represent an interesting research result themselves, they do not represent the main objective of this work which is to build believable facial expressions. In this section we detail the results of some user studies we have conducted to evaluate the quality of the facial expressions we have developed with the processing explained in chapter 4.3. The results are split in two halves; we will first discuss the quality of the facial expressions generated for the HapteK avatar and then detail the results obtained with the iCat expressions.

#### 4.4.2.1 Cherry, the HapteK Avatar

The facial expression generated for the avatar Cherry have been evaluated through a user study which has been conducted with 16 candidates on the 18-30 age range from both sex. The user study was on the form of a closed form questionnaire. The subjects were first asked to recognize the expressed emotions by choosing for six randomly

---

In \ Out	Hap	Dis	Sad	Fea	Ang
Happiness	100%	0%	0%	0%	0%
Disgust	0%	63%	0%	0%	0%
Sadness	0%	0%	100%	0%	0%
Fear	0%	6%	0%	94%	0%
Anger	0%	13%	0%	0%	88%

Table 4.5: Emotional recognition rate for Cherry

sorted videos an emotion over 5 different possibilities and then to express their opinion in term of believability and exaggeration, by choosing a mark between 0 and 4 (not believable/exaggerated to very believable/exaggerated), about the expressions shown by one actor, an avatar tuned by standard Haptex expressions and another tuned with the designed parameters.

The recognition scores of this user study can be seen in table 4.5 while the average values of believability and exaggeration are grouped in the graph in figure 4.43.

The results shows that the expressions are well recognizable. Albeit the task was quite simple (recognize an emotion choosing among five) the result presented in table 4.5 is exceptionally good with an average recognition rate of approximately 94%.

For what regards believability and exaggerations it can be seen that the results are in line with the one obtained with Haptex player default expressions while both sets of facial expressions are less believable but also less exaggerated than the expressions showed by the actors in our video database.

We believe that these results are very satisfying confirming the possibility of port the expressions described in Scherer's component process theory to an avatar platform.

#### 4.4.2.2 Cleo, the Philips Robot

The experiment we conducted to evaluate the facial expressions generated for Cleo had exactly the same modalities of the one we conducted to evaluate Cherry's expressions. Sixteen participants, four women and twelve men, between twenty and thirty years old were involved and had to recognize the expressions we generated in a closed form questionnaire. For this, each of the nine expressions was shown twice starting from the neutral position and participants could choose the expression they recognized between a list containing the nine expressions.

Table 4.6 shows results obtained. In general, participants recognized happiness, anger, fear and indifference better than others expressions. Disgust expression is confused with contempt expression. Pride is recognized as pride with 38% and as fear with 31%. Some expressions such as contempt, pride, indifference and shame are very difficult to recognize without context. Some of these expressions could be better recognized if we were able to fine tune the gaze of the iCat to directly watch user eyes.

In average the expressions are recognized in 51% of the cases and therefore 4.6 times better than random. If we consider the improved complexity, then this result is comparable with the one obtained with the Haptex avatar (indeed 94% is just about 4.7 times better than random). We are therefore pretty satisfied with this result which confirm the possibility of porting the process of generation of facial expressions defined by Scherer onto this particular robotic platform.

In a second time subjects were asked to evaluate the believability and the exaggera-



Figure 4.31: One possible evolution of the expression Anger using CPT



Figure 4.32: One possible evolution of the expression Happiness using CPT

In \ Out	Hap	Dis	Con	Sad	Pri	Fea	Ang	Ind	Sha	None
Happiness	75%	13%	0%	0%	0%	0%	0%	0%	6%	6%
Disgust	0%	25%	56%	0%	6%	0%	0%	6%	0%	6%
Contempt	6%	0%	19%	0%	63%	6%	0%	0%	0%	6%
Sadness	0%	19%	0%	56%	0%	0%	0%	0%	25%	0%
Pride	13%	0%	0%	0%	38%	31%	0%	6%	0%	13%
Fear	0%	0%	0%	0%	0%	69%	0%	6%	13%	13%
Anger	0%	0%	6%	0%	0%	0%	88%	0%	0%	6%
Indifference	0%	0%	0%	6%	19%	0%	0%	63%	13%	0%
Shame	0%	0%	0%	31%	0%	0%	6%	19%	31%	13%

Table 4.6: Emotional recognition rate for iCat

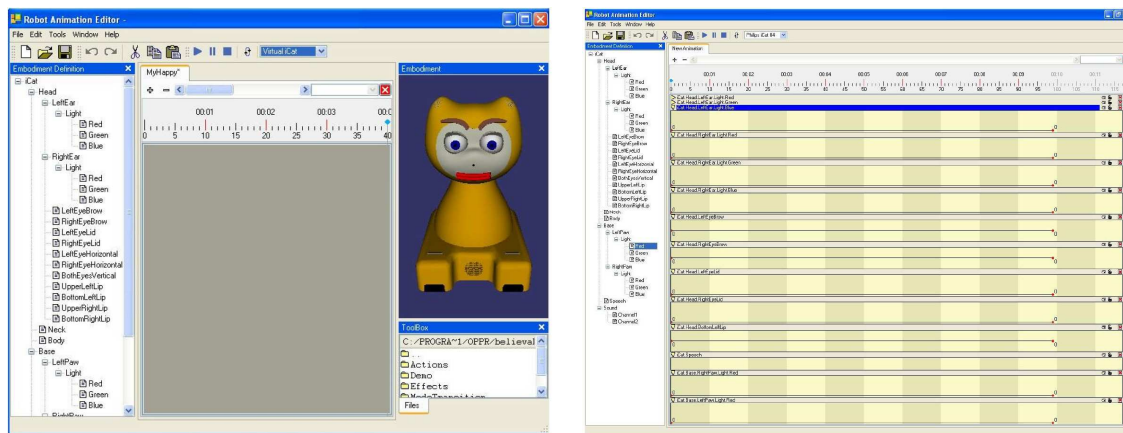


Figure 4.33: Philips iCat Animation Editor



Figure 4.34: Happiness Example



Figure 4.35: Disgust Example

tion of our new expressions by answering to a questionnaire. Participants had to evaluate the intensity of believability and exaggeration with a scale that going from 0 for not believable (or exaggerated) to 5 for very much believable (or exaggerated). Because not all of the expressions we designed were available among the expressions defined by Philips and in our database of human videos we compare the results obtained by the five expressions anger, happiness, sadness, fear, and disgust (i.e. the same expression that were evaluated on the avatar).

The obtained results, represented in figure 4.44, have shown an amelioration of the results obtained with our expressions with respect to the one developed by Philips.

One might be surprised to notice that the believability of the expression generated with the avatar by Haptik has been evaluated as worse than the believability of the expressions displayed by the iCat robot. This result might, nevertheless, have been expected for one main reason: people does expect much less from a cat robotic face than they do from a high fidelity 3D computer human-like face. At the same time robots, for their nature, influence more easily our reality because of their physical presence.

## 4.5 Concluding Remarks

In this part of the thesis we have discussed the topic of affective display. After having introduced the topic we have overviewed Scherer's Component Process Theory of Emotions and showed how emotional expressions (facial, prosodic, ANS, and others) can be



Figure 4.36: Contempt Example



Figure 4.37: Sadness Example



Figure 4.38: Fear Example



Figure 4.39: Anger Example



Figure 4.40: Indifference Example



Figure 4.41: Shame Example



Figure 4.42: Pride Example

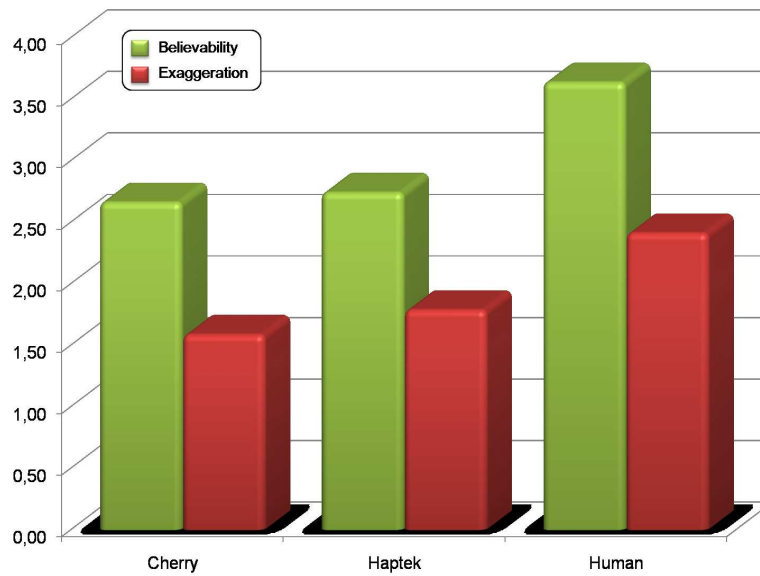


Figure 4.43: Believability and exaggeration for Cherry expressions

linked to the process of appraisal of the surrounding events which generate emotions.

In the following chapter we have discussed the most common representations for facial and vocal prosody expressions, together with some existing technologies. We have continued presenting the process of generation of facial expressions for two different platforms: Cherry, an avatar built with the Haptik technology and Cleo, a cat robotic platform from Philips. Finally, we have evaluated the developed expressions with two user studies whose objective was, on the one hand to evaluate the possibility of recognize these expressions and on the other hand to evaluate the believability and exaggeration of the developed facial expressions.

The results confirm that Scherer's CPT can be successfully used for the generation of facial expressions on both an avatar (the expressions are recognized in 94% of the cases) and a robot (in this case the expressions are recognized in more than 51% of the cases).

In the next part we will discuss the topic of automatic **affect recognition**.

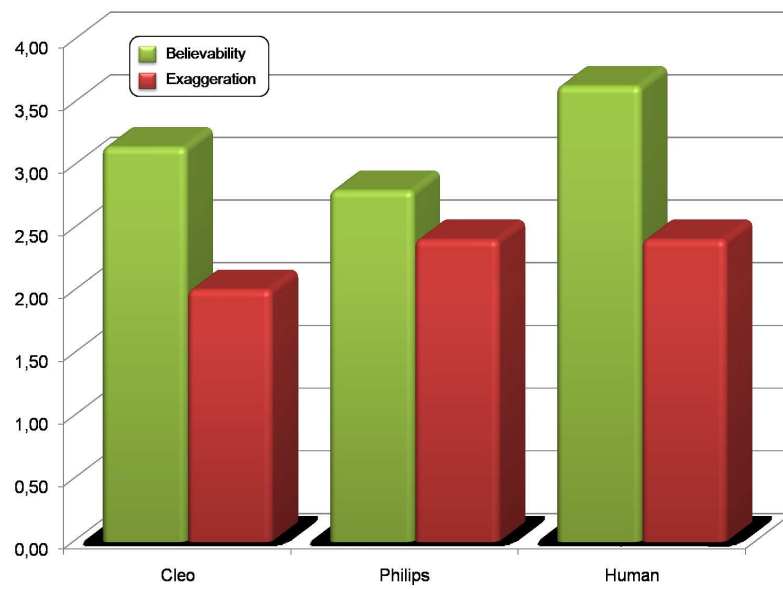


Figure 4.44: Emotions believability



## Chapter 5

# Emotional Recognition

### 5.1 Introduction

In part 2 we have discussed the motivations of affective computing and given few applications examples. We have seen that, in most of these possible applications scenarios, affect display, automatic affect recognition, and affect synthesis play a fundamental role allowing for believable behavior and natural interactions.

In part 4 we have discussed the topic of affect display and we have described the process of generation of believable facial expressions, demonstrating that Scherer's CPT of emotions can be ported to different computer-based and robotic platforms. In this part of the thesis we discuss the topic of affect and **emotion recognition**. Emotion recognition is probably the key component of most of the affective computing scenario of chapter 3.2.

For example, in e-learning applications the computer needs to understand the student's affective state to react accordingly and maximize the learning performances. Although affective reactions are likely to be more effective, an interactive tutoring system could also react to the student's affective state in a "cold" rational manner and obtain good results.

Similarly, a personal electronic assistant would need to recognize and process both the context and its user's affective state to properly help him/she filtering out calls, setting appointments, or suggesting pauses without need for emotional behaviors.

In this part we firstly, discuss some of the relevant work on this domain and then we present our research results. In particular we present:

- SAMMI: a Semantic Affect-enhanced MultiMedia Indexing paradigm which uses emotions and other content based semantic information to tag media;
- ARAVER: Automatic Real-Time Audio-Video Emotion Recognition system that we developed to automatically recognize people's emotions through their vocal prosody and facial expressions;
- AMMAF: an Automatic Multimodal and Multilayer Affect Fusion paradigm that use multimodal information to extract reliable user emotional estimates.

### 5.2 Relevant Work

In this chapter we present the most relevant state of the art works about emotion recognition. Before discussing this topic we need, nevertheless, to make a distinction because

---

two main families of emotion recognition can be drawn according that the emotion estimates refer to people or objects. Indeed, while it is possible to extract emotional estimates for many different medias (or rather estimates about the emotions conveyed by different medias) by analyzing, for example, the colors and composition of a picture or the tempo and timbre of a song, this is not the focus of our current research.

Albeit few systems tried to exploit few different modalities (e.g. gestures), there are three main areas where people's emotion recognition systems have been developing. These are:

1. emotion recognition from still images and video of people *facial expressions*
2. emotion recognition from audio and *speech*
3. emotion recognition from the *Autonomous Nervous System* (ANS) signals

In the following sections we will review some of the most relevant works about emotion recognition focusing on the first two modalities. It is not the intent of this chapter to cover all the extensive literature in these domains. For more details about existing automatic emotion recognition system please refer to Pantic and Rothkrantz [2000b, 2003], Sebe et al. [2005b,c], Jaimes and Sebe [2007], Zeng et al. [2007, 2009].

### 5.2.1 Emotion Recognition via Facial Expressions

The face provides our primary and preeminent mean of communicating emotions Picard [1997], Scherer [2001]. Research has extensively studied this phenomenon starting in 1862 with Duchenne's book. Recent works in computer-based automatic facial expressions Pantic and Rothkrantz [2000a], Sebe et al. [2002] can exceed 90% average recognition scores. Ideally, an emotion recognition system working on facial expressions should:

1. Work in real-time
  2. Automatically detect the position of the face in the image;
  3. Automatically extract emotion related features;
  4. Process subjects independently of their sex, age, ethnicity, or outlook;
  5. Deal with lightening variation and difficult illumination conditions;
  6. Deal with partially occluded faces and distractions such as glasses, facial hairs, etc.;
  7. Deal with rigid head motions;
  8. Deal with unilateral facial changes;
  9. Do not require any special make-up or markers;
  10. Deal with inaccurate facial expression data;
  11. Automatically classify facial expressions;
  12. Distinguish an unlimited number of emotions;
  13. Assign quantified interpretation labels;
-

14. Assign multiple interpretation labels;
15. Classify action units (AUs);
16. Quantify action unit activation;

While most of the existing system succeed in automatically process images extracting emotional features from the subjects and classifying emotions, most system still fail in handling difficult lightening conditions, head pose changes, and in meeting the real-time needs.

In table 5.1 summarizes the features of existing systems for facial affect recognition system with respect to the properties of the ideal facial expression emotion recognition system listed above.

---

Reference	Properties																Test Base	Result
	1	2	3	4	5	6	7	8	10	11	12	13	14	15	16			
Analysis of Still Images																		
Cottrell and Metcalfe [1990]	U	X	U	★	X	X	U	★	U	★	8	X	X	X	X	40i, 5s	40%	
Kearney and McKenzie [1993]	X	X	X	★	X	X	U	★	U	★	8	X	X	X	X	17i, 1s	91%	
Padgett and Cottrell [1996]	X	X	X	★	X	X	U	★	U	★	7	X	X	X	X	84i, 1s	86%	
Huang and Huang [1997]	X	★	★	1	X	X	X	★	U	★	6	X	X	X	X	90i, 15s	85%	
Kobayashi and Hara [1997]	★	★	★	1	U	X	X	★	U	★	6	X	X	X	X	90i, 15s	85%	
Edwards et al. [1998a]	X	X	★	U	★	X	★	★	U	★	7	X	X	X	X	200i, 25s	74%	
Hong et al. [1998]	★	★	★	U	X	X	★	★	★	★	7	X	X	X	X	175i, 25s	81%	
Zhang et al. [1998]	X	X	X	★	X	X	U	★	U	★	7	★	★	X	X	213i, 9s	90%	
Colmenarez et al. [1999]	U	★	★	T	U	X	★	U	★	★	6	X	X	U	U	5970i, 18s	86%	
Lyons et al. [1999]	X	X	X	★	X	X	U	★	U	★	7	X	X	X	X	193i, 9s	92%	
Pantic [2001]	X	★	★	3	X	★	★	★	★	★	6	X	★	31	X	196i, 8s	97%	
Cohen et al. [2003]	X	X	★	U	U	U	X	★	U	★	7	X	X	U	U	12600i, 5s	65%	
Sebe et al. [2002]	X	X	★	U	U	U	X	U	U	★	7	X	X	X	X	12600i, 5s	64%	
Fasel et al. [2004]	X	★	★	T	U	U	X	X	X	★	★	X	X	★	X	503i, 210s	95%	
Bartlett et al. [2006]	★	★	★	T	X	U	X	X	X	★	★	X	X	20	X	4257i, 119s	>90%	
Whitehill and Omlin [2006]	U	★	★	T	X	X	X	X	X	U	U	X	X	11	X	580i, 210s	92%	
Analysis of Videos and Image Sequences																		
Mase [1991]	X	X	★	U	X	X	X	U	X	★	4	X	X	X	X	30v, 1s	80%	
Black and Taylor [1997]	X	X	★	U	★	X	★	★	X	★	6	★	X	U	X	70v, 40s	88%	
Essa and Pentland [1997]	★	★	★	★	★	★	U	★	★	★	4	X	X	2	X	30v, 8s	98%	
Kimura and Yachida [1997]	U	★	★	U	★	X	X	★	★	★	3	X	X	X	X	6v, 1s	X	
Otsuka and Ohya [1998]	X	U	★	U	U	X	★	X	X	★	6	X	X	X	X	120v, 2s	U	
Wang et al. [1998]	X	X	★	U	X	X	X	★	U	★	3	★	X	X	X	29v, 8s	95%	
Ji et al. [2006]	★	★	★	U	U	U	X	U	U	★	2	X	X	X	X	320v, 8s	0.95	
Yeasin et al. [2006]	X	★	★	★	★	U	U	U	★	★	6	★	X	X	X	488v, 97s	91%	
Valstar et al. [2007]	X	★	★	T	U	U	★	U	U	★	2	X	X	X	X	202v, 52s	94%	
Paleari et al. [2009a]	★	★	★	★	X	★	X	★	X	★	6	X	X	X	X	1320v, 44s	43-97%	

Legend: ★ = yes, X = no, U = unknown/missing entry, T = handles subjects in the train base,  
i = images, v = videos, s = subjects

Table 5.1: Properties of state of the art approaches to facial expression emotion recognition (adapted from Pantic and Rothkrantz [2003])

---

We can distinguish two main stream in the current research of automatic emotion recognition: the stream of systems that employ geometric features such as the shapes of the facial components (i.e. eyes, eyebrows, mouth, etc.) and the locations of facial salient points (e.g. the corner of the mouth) and the second stream composed of systems based on appearance features mainly representing facial texture (including furrows and wrinkles) and employing techniques such as Gabor wavelets or eigenvalues. Typical examples of systems based on facial features are those of Pantic and Rothkrantz [2000a], who used 25 features as distances and angles from predefined feature points, Sebe et al. [2002], who considers 12 facial motion measures, or Essa and Pentland [1997] who uses either a muscle activation model derived from the optical flow of user's video or 2D motion energy maps. Typical examples of methods based on appearance-features are those of Bartlett et al. [2006] who use Gabor wavelets, Whitehill and Omlin [2006] who use Haar features, and Fasel et al. [2004] who uses the latent semantic statistics of the gray-level intensities. Chang et al. [2004] uses the concept of manifolds, lower space representations of the images reminding eigenvalues.

A second possible distinction can be made among methods using static information (which work essentially on still images) and methods explicitly taking advantage of the dynamic information available from the analysis of videos.

Recently, some system have been focusing on the recognition of action units rather than the emotions because AUs are a relatively objective and complete description of the facial expressions. Simple mapping such as EMFACS or FACS-AID by Ekman et al. [2002], or more complex one such as the one described by Scherer [2001], can be used to convert the action unit activations to facial emotional expressions. This is, for example, the case of the works by Bartlett et al. [2006], Whitehill and Omlin [2006], Fasel et al. [2004], and Essa and Pentland [1997]. This approach presents, nevertheless, a limitation linked to the scarce limitation of publicly available databases tagged with AU information and the difficulties related with the construction of such a database.

### 5.2.2 Emotion Recognition via Speech Signal

The auditory aspect of communications has been extensively studied for the retrieval of emotional information. Indeed, as we detailed in part 2, many human-machine interactions application scenarios could profit of the emotional information and voice is probably the primary human communication channel.

A simple example of an existing system which could profit from the availability of emotional information is the one of an automatic phone answering machine: if the user manifest cues of frustration in his voice then we would probably like to take action and modifying this situation before the user could become angry; for example the machine could try to switch the user to a human operator). Another reason why understanding emotions in speech signals is important is linked to the fact that the accuracy of speech recognition, which can exceed 90% for neutrally spoken words, tends to drop to 40% 50% in the presence of emotional speech (Steeneken and Hansen [1999]). A similar behavior has been also reported in the case of speaker verification systems (Scherer et al. [1998]).

Any automatic emotion recognition system consists of two main steps: the extraction of the emotion-related features and the classification phase. In the case of the voice signal the feature used are mainly the one described in section 4.2.3 or the semantics of the words being pronounced.

Pantic and Rothkrantz [2003] lists the 16 characteristics of existing vocal emotion

---

recognition systems. These are:

1. Can non professionally spoken input samples be handled?
2. Is the performance independent of variability in subjects, their sex and age?
3. Are the auditory features extracted automatically?
4. Are the pitch-related variables utilized?
5. Is the vocal energy (intensity) utilized?
6. Is the speech rate utilized?
7. Are pitch contours utilized?
8. Are phonetic features utilized?
9. Are some other auditory features utilized?
10. Can inaccurate input data be handled?
11. Is the extracted vocal expression information interpreted automatically?
12. How many interpretation categories (labels) have been defined?
13. Are the interpretation labels scored in a context-sensitive manner (application-, user-, task-profiled manner)?
14. Can multiple interpretation labels be scored at the same time?
15. Are the interpretation labels quantified?
16. Is the input processed in real time?

Table 5.2 reports the the main state of the art efforts in this domain according to these characteristics.

---

Reference	Properties																Test base	Result
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16		
Banse and Scherer [1996]	X	★	★	★	★	★	X	X	★	X	★	14	X	X	X	U	2 se, 12 s	54%
Dellaert et al. [1996]	U	X	★	★	X	★	★	X	X	X	★	4	X	X	X	U	50 se, 5 s	80%
Tosa and Nakatsu [1996]	★	T	★	★	★	X	★	★	★	X	★	8	X	X	X	X	100 wo, 10 s	60%
Amir and Ron [1998]	★	X	★	★	★	★	★	X	★	X	★	5	X	★	★	★	ss, 2 s	U
Li and Zhao [1998]	★	X	★	★	★	X	★	X	X	X	★	6	X	X	X	X	20 se, 5 s	62%
Nakatsu et al. [1999]	★	★	★	★	★	X	X	★	★	X	★	8	X	★	★	X	100 wo, 100 s	50%
Fellenz et al. [2000]	U	U	★	★	★	X	★	X	★	★	★	4	X	X	X	X	U	U
Huber et al. [2000]	★	★	★	X	X	X	X	★	★	X	★	2	X	X	X	X	ss, 39 s	66%
Kang et al. [2000]	X	X	★	★	★	X	X	X	X	X	★	6	X	X	X	X	5 wo, 8 s	85%
Petrushin [2000]	★	T	★	★	★	★	X	X	★	X	★	5	X	X	X	X	4 se, 30 s	66%
Polzin [2000]	X	T	★	★	★	X	X	X	★	X	★	5	X	X	X	X	50 se, U	73%
Zhao et al. [2000]	X	X	★	★	★	X	X	X	★	X	★	4	X	X	X	X	4 se, 5 s	90%
Nwe et al. [2001]	U	T	★	X	★	X	X	X	X	X	★	6	X	X	X	X	90 se, 2 s	66%
Sato et al. [2001]	U	T	★	★	X	X	★	X	X	X	★	4	X	X	X	X	1 wo, 13 s	74%
Noble [2003]	U	★	★	★	★	★	★	X	★	X	★	13	★	X	X	X	>2000 se, 7 s	24%
Morrison and Wang [2005]	★	★	★	★	★	★	X	★	★	★	★	2	★	X	X	★	391 se, 11 s	80%

**Legend:** ★ = yes, X = no, U = unknown, T = handles subjects in the train base,  
se = sentences, wo = words, ss = spontaneous speech, s = subjects

Table 5.2: Properties of state of the art approaches to vocal prosody emotion recognition (adapted from Pantic and Rothkrantz [2003])



### 5.2.3 Physiological Signals and Alternative Emotion Recognition Systems

The third more common media for emotion recognition are physiological signals. Heart rate and skin conductivity are the two signals which are mostly considered in this kind of system Picard et al. [2001], Lisetti and Nasoz [2004], Villon [2007].

Few works focus on texts for recognizing emotions. Miyamori et al. [2005] focuses on blog texts and detects excitement through an analysis of the content of the comment left by users for indexing and summarizing videos. Salway and Graham [2003], applies similar processing to the transcriptions of the film descriptors used for visually impaired people.

Wu and Huang [1999] presents a survey on systems extracting the emotional information from people gestures.

Few systems try to attach an emotional meaning to a music piece. The typical example of this paradigm is the work of Kuo et al. [2005] which extrapolate the emotional meaning from film music and use this information for music recommendations. An interesting study by Kim et al. [2005] recognizes emotions linked to still images by looking at texture and color information with accuracy approaching 85%.

### 5.2.4 Multimodal Emotion Recognition Systems

Even though this approach is very promising, very few works have exploited the intrinsically multimodal nature of emotions by using two or more modalities to increase recognition scores. Most (Chen et al. [1998], Busso et al. [2004], Sebe et al. [2005a]) couple audio vocal prosody and video facial expression information and claim a substantial improve in performances. Conati [2002] instead, prefers to couple a system simulating the emotional appraisal of the user based on Ortony et al. [1988] and a system based on physiological signals having as a result a more reliable emotion estimate. Only very few systems try to employ different modalities such as the gestures in combination with speech and or facial expressions. This is the case of the works from Castellano et al. [2007] who improve the monomodal average score by using multimodality and also compare feature-level fusion with decision-level fusion for these three modalities concluding that, in their experiments, feature-level fusion slightly outperforms decision-level fusion.

In this section we have presented some of the most relevant works on automatic emotion recognition. In the next chapter we describe SAMMI, an indexing and retrieval system based on content based semantic information which also includes emotions for tagging media excerpts, and we give few simple examples of the possible interactions between emotions and other content-based information.

## 5.3 ARAVER: Automatic Real-Time Audio-Video Emotion Recognition

In our approach, emotion recognition is performed by fusing information coming from both the visual and audio modalities. In particular, we are targeting the identification of the six “universal” emotions listed by Ekman and Friesen Ekman and Friesen [1986] (i.e. anger, disgust, fear, happiness, sadness, and fear).

According to the study of Ekman and Friesen these six emotions are characterized by the fact of being displayed via the same facial expression regardless of sex, ethnicity,

---

age, and culture. In their studies Ekman and Friesen visited many different cultures and visioned videos of two preliterate New Guinea cultures taken in the late 50s when these peoples had their first contacts with “the outside world” (Ekman [1984]). Ekman and Friesen found out that, although different people looked very different in their dress and other aspect of their behavior, their facial expressions were completely familiar. The conclusion is that the evolution must have brought human beings to communicate the emotion in the same way because some facial expressions have components other than the communicative ones. As an example, when we experience fear we spread our eyes open; it is likely that this happens because we need to enlarge our field of view.

As several researchers did before us Zeng et al. [2009], Noble [2003], Datcu and Rothkrantz [2008], we implicitly make the assumption that these findings are true for emotional prosodic expression too. In other words, we assume that people express the same emotions varying the way they speak in about the same way regardless of sex, ethnicity, age, and culture. The fact that several researchers successfully performed recognition on these emotions confirms the validity of this hypothesis.

The idea of using more than one modality arises from two main observations:

1. when one, or the other, modality is not available (e.g. the subject is silent or hidden from the camera) the system could still return an emotional estimation thanks to the other,
2. when both modalities are available, the diversity and complementarity of the information, should couple with an improvement on the general performances of the system.

We based our system on a machine learning approach. In this case we can identify four main phases:

1. **Feature extraction:** data is automatically analyzed and some features are extracted which represent the relevant data (i.e. the one related to emotions) in a compressed way.
2. **Feature fusion and vector generation:** the features are put together in a vector and shaped to conform the input of the machine learning system.
3. **Machine learning:** some kind of classifier is used to extract estimation about the emotions.
4. **Multimodal fusion:** the outputs coming from different modalities are fused together to increase the precision and reliability of the estimations.

Each one of these four phases can be further divided into several sub-phases.

In the next few sections we will describe in detail the processing involved in these phases.

### 5.3.1 Databases

As we have said, using more than one modality might be the key for recognizing emotions in a reliable, stable, and precise manner. Machine learning approaches base their performances on pattern learned from ground-truth data: in this case multimodal audio and video data. With ground-truth data it is intended data (video or other media) which

---

was manually indexed by experts to represent the concept(s) we are searching in it; in our case this is videos with people expressing emotions indexed so that we know for sure which frames represent which emotion(s).

In our case we are looking for a database presenting few specific characteristics:

- The ground–truth data should present the highest possible number of emotions. In particular we search for a database presenting samples of the six universal emotions and neutral samples.
- The ground–truth data should use multiple representations for the emotions. For example ground–truth could couple discrete categories (see section 3.3.2.1) and a dimensional or componential representation of emotions (see sections 3.3.2.2 and 3.3.2.3). When applicable we may also want experts to tag the data with the most common relevant feature values, like AUs for the facial expressions or pitch, formants, and phonemes for the audio. Indeed, because of the complexity of the whole system we may want to be able to evaluate the quality of each step of the processing.
- The emotions should be presented in the most complete way and make use of the highest possible number of modalities. Ideally we would like to have multiple views, multiple audio signals, and physiological (ANS) signals.
- The highest possible number of different subjects should be involved. If we want to build a system really capable of working with subjects of different ages, sexes, ethnicity, etc. we need ground–truth to cover extensively every sex, ethnicity, age, etc.
- The ground–truth database size should be the biggest possible. For each emotion, age, sex, ethnicity, lighting condition, etc. we would like to have the biggest possible amount of ground–truth data.
- Different constraint scenarios should be covered: we would like to test our system to both constrained and un–constrained data.
- The representation of the emotions should be the most natural possible. Ideally we would like real (as opposed to acted) emotions.

Unfortunately, to our knowledge such a kind of ground–truth database does not exist. Indeed, some of the difficulties involved with the collection of emotional data may not be easily overcome and the cost related with the construction of such a database not easily covered. For example, the availability of “real” emotional data may be questionable, especially if we want to put the subject into constrained, noise free, environments and possibly while monitoring his/her physiological signals with complex devices.

In the next section we will describe the eNTERFACE’05 database, a publicly available audio visual emotion database built for the network of excellence eNTERFACE (Martin et al. [2006]).

### 5.3.2 eNTERFACE’05

For our experiments we have selected the eNTERFACE’05 database Martin et al. [2006] (see figure 5.1). This database is composed of over 1300 emotionally tagged videos por-

---

traying non-native English speaker displaying a single emotion while verbalizing a semantically relevant English sentence.

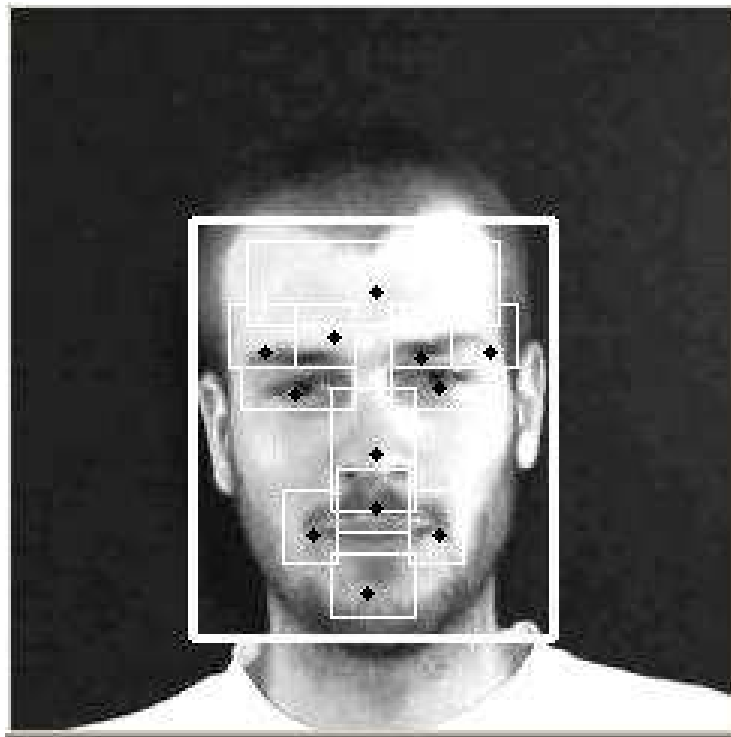


Figure 5.1: Anthropometric 2D model

The 6 universal emotions from Ekman and Friesen Ekman and Friesen [1986] are portrayed, namely anger, disgust, fear, happiness, sadness, and surprise. Videos have a duration ranging from 1.2 to 6.7 seconds ( $2.8 \pm 0.8$  sec). This database is publicly available on the internet but carries few drawbacks due to the low quality of the video compression and actor performances:

- The quality of the video encoding is poor: the 720x576 pixels interlaced videos are interlaced and encoded using DivX 5.05 codec at around 230 Kbps and 24 bits color information. Blocking effect is quite visible on the frames.
- Subjects are not trained actors resulting in a potentially mediocre quality of the emotional expressions.
- Subjects were asked to utter sentences in English even though this was not, in most cases, their natural language; this may result in a low quality of the prosodic emotional modulation.
- Not all of the subject learned their sentences by heart resulting in a non negligible percentages of videos starting with the subjects looking down to read their sentences.
- The reference paper Martin et al. [2006] acknowledges that some videos (around 7.5%) do not represent, in a satisfactory way, the desired emotional expressions. In

theory, these videos were rejected but this is apparently not the case in the database which is available for download.

- For each video shot only one emotional tag is given. Ideally we would have liked each and every frame to be indexed with a, possibly quantified, emotion. Furthermore, in the database there is no neutral sample.

This kind of drawbacks introduce some difficulties but, in some cases, it also allows us to develop algorithms which should be robust in realistic scenarios.

### 5.3.3 Facial Expression Feature Extraction

We have developed a system performing real-time, user independent, emotional facial expression recognition from still pictures and video sequences. In order to satisfy the computational time constraints required for real-time operation, we employ Tomasi Lucas–Kanade’s algorithm Tomasi and Kanade [1991] to track characteristic face points as opposed to more complex active appearance models Batur and Hayes [2005], Cohen et al. [2003], Valenti et al. [2007], Datcu and Rothkrantz [2008].

Our system is based on an approach at the edges of feature point tracking and region of interest movement tracking. In the following sections we detail the process adopted for extracting and tracking the positions of 12 fiducial facial points on the human face.

#### 5.3.4 Step 1: Face Detection

As a first step the system detects the position of the face (see figure 5.3(a)) in the video using the Viola and Jones [2001] face detector (see Tiddeman [2007] for a simple demo of the Viola–Jones face detector). The Viola–Jones face detectors makes use of a boosted approach; a cascade of many different simple classifiers are computed on Haar-like features. In this kind of approach images are represented by simple brightness templates like the ones in figure 5.2.

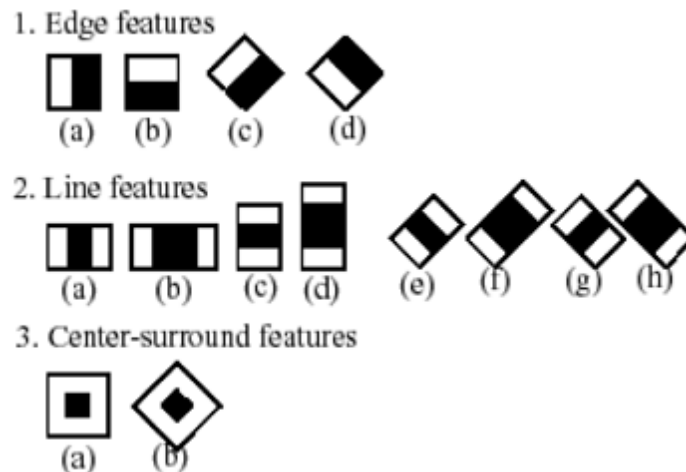


Figure 5.2: Sample Haar filters

---

We employ three detectors, one for the whole face, a second for the eyes region, and a third for the mouth.

The Viola–Jones detector that was implemented in OpenCV tend not to be very stable (i.e. the size and position of the bounding box may differ considerably) and reliable (i.e. only about 79% of the faces are detected) along different consecutive frames of most eNTERFACE'05 videos. To reduce the error we take a variable number of consecutive frames bigger than three but limited to 11 such that 60% or more of the detection are compatible (i.e. 2 out of 3, 3 out of 5, 5 out of 7, etc.). We defined that two frames are compatible if their relative areas do not change by more than 10% and if the distance between their center point is smaller than 10 pixels.

Not having a ground truth for the estimation of the face we had to evaluate the precision of this algorithm manually with a scale of evaluation going from *wrong estimation* to *correct estimation* passing from *not precise estimation*. Thanks to this algorithm we can correctly estimate the size and the position of the face in more than 96% of the cases. In few cases (about the 4%) the position or the size of the face is not precisely found; in these cases, the error was, nevertheless, so small that the effect shall not be relevant for the following phases.

The estimation of the position of the eyes and mouth regions is more problematic. Also applying the algorithm described before the percentage of the detected eyes and mouth is so small and the errors so big that in some of the cases we were not able to correctly detect the position of those two regions. In these cases we just supposed the subject to face the camera (i.e. rotations around the x and y axis around  $0^\circ$ ) with his/her head posing vertically (i.e. rotation around the z axis connecting the camera with the subject equal to  $0^\circ$ ).

Thanks to the information about the position and size of these three bounding box we estimate position, size, and pose of the face in the video frame.

This algorithm is robust to small head pan, tilt, and rotations (i.e. pose angles with respect to the camera smaller than  $25^\circ$ ,  $30^\circ$  on all three dimensions). For bigger rotations (which are not present in the eNTERFACE'05 database anyway) the Viola–Jones classifier did not work properly anymore.

A possible solution to this problem could be found with the use of Viola–Jones classifier trained with other head poses or with the use of other face detectors systems based, for example, on skin color or other features. In the case of big rotations, some ROI may also be hidden. An intelligent system could make use of the redundancy given by the face symmetry to estimate the hidden part of the face.

### 5.3.5 Step 2: ROI definition

The second step consist in defining the regions of interest (ROI) for which we want to extract the feature points we want to track. For doing so we have tried two approaches.

Both approaches are based on a 2D anthropometric model of the human face (see figure 5.1). In the first approach this model is computed from the data regarding the position, size, and rotations of the face that we have estimated with the use of the Viola–Jones face detector. The second possible approach consist in trying to detect the actual pupil position (see figure 5.3(b)) and from this information re–estimate size, position, and orientation of the head to more precisely construct and apply the 2D model.

One interesting thing about the first approach is that we can easily change the 2D anthropometric model with a regular grid and get as output the so called motion flow in

---



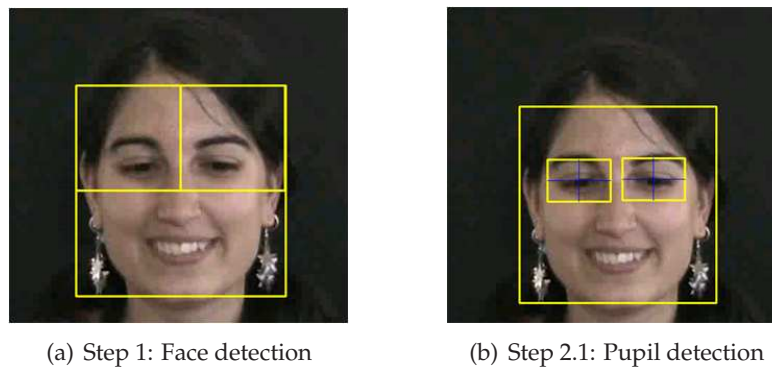


Figure 5.3: Face and pupil detection

a fairly efficient way.

The second approach passes, as it has been said, through a phase of pupil detection. For the pupil detection we have further tried two techniques.

**The first technique for pupil detection** consist of computing, for the upper-left and the upper-right quarters of the detected face, the horizontal and vertical projections (similarly to what it is done in Vukadinovic and Pantic [2005]).

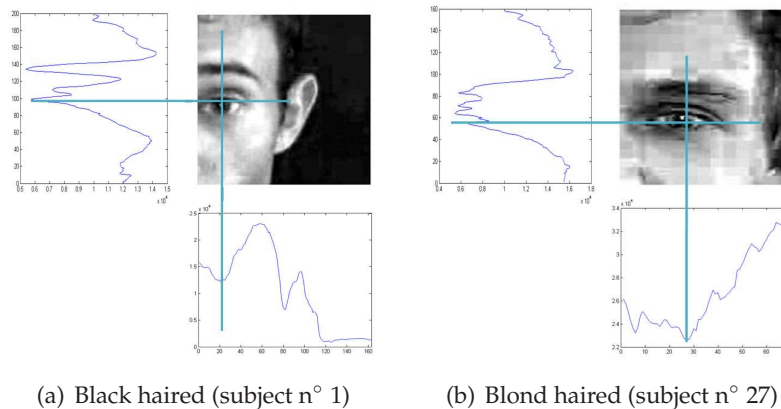


Figure 5.4: Two example of pupil detection using vertical and horizontal projections

The results depend very much on illumination and physical characteristics of the subject depicted. In particular it is often more difficult to find the pupil position of blond, or fair haired, subjects under normal illumination. In figure 5.4 we show two examples of pupil detection using vertical and horizontal projections.

As it can be seen from the second example (figure 5.4(b)), on the blond subjects the detection is more ambiguous. The projections are more noisy because of the contrast adjustment and many minima appear in the eye region of both the horizontal and vertical projections.

Because of these reasons it was not possible to find a simple algorithm allowing to determine the correct peaks on the vertical and horizontal projections for all subjects and illumination condition.

**The second approach for pupil detection** consist in extracting the position of the single eyes thanks to other two Haar-cascade classifiers employing the Viola-Jones tech-



nique overviewed in section 5.3.4. With this approach we have better results (see figure 5.5); nevertheless, in almost 14% of the cases the eye pupil are still not correctly identified.

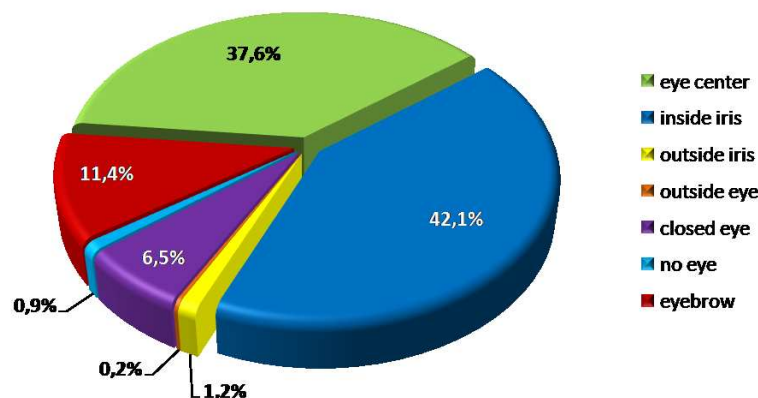


Figure 5.5: Pupil detection performances using boosted Haar-cascade classifiers

The most common error happens when the Viola-Jones classifier identifies the eyebrow as if it was an eye (see figure 5.6(c)). This happens because the classifiers have been trained to also recognize closed eyes as such and can be facilitated by a non perfect estimation of the face position. Indeed if the face position is estimated as being upper than it actually is then the eyebrow position may be more easily recognized as being the normal one for an eye.

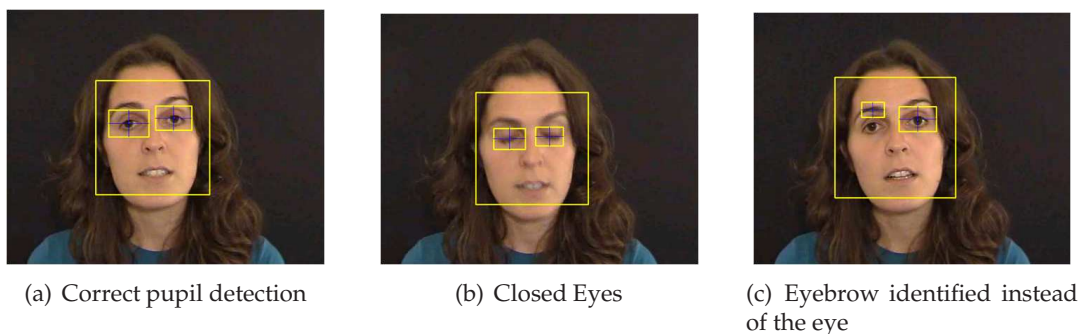


Figure 5.6: Examples of pupil detections with boosted Haar-cascade classifiers

Although the results for this second pupil detection algorithm are quite accurate for the vast majority of the frames the errors involved in the remaining part of the frames are such that using this system will not be acceptable for the following computation steps.

**A 2D anthropometric model** of the human face has been computed following some of the guidelines from the Sohail and Bhattacharya [2007] work. The resulting model (which is superposed to the face in figures 5.1) defines 12 region of interest corresponding to the following regions of the face (see also figure 5.7(a)):

1. right mouth corner

2. left mouth corner
3. nose
4. right eye
5. left eye
6. forehead
7. mouth bottom / chin
8. external right eyebrow
9. internal right eyebrow
10. internal left eyebrow
11. external left eyebrow
12. upper lip / mouth top

### 5.3.6 Step 3: Feature Point Extraction

For each one of the 12 ROI defined in the former section 5.3.5 we search points with strong gradient on the neighborhood region with the Shi and Tomasi [1994] version of the Harris and Stephens [1988] features. To track these points we employ the Tomasi and Kanade [1991] pyramid version of the Lucas and Kanade [1981] algorithm for optical flow estimation. The algorithm evaluates the differences of two consecutive frames by assuming the flow to be constant in a local neighborhood around the central pixel under consideration at any given time.

Since points may vibrate around their actual position because of compression blocking effect and/or small errors of the Lucas–Kanade algorithm, instead of tracking one point (the central point) for each ROI we search for as many as 50 different points and for each frame we compute the center of mass as shown in figure 5.7(a). Points with strong gradient tend to be the one at the borders of two different regions (e.g. eyebrow–to–skin, eyes–to–skin, mouth–skin–to–cheek–skin, etc.).

As a result of this process, we have 24 features per frame, corresponding to 12 pairs of the feature points (FP)  $x(i)$  and  $y(i)$  coordinates representing the average movement of points belonging to the regions of interest defined above.

In appendix B we present a preliminary work about an alternative solution to the detection of feature points which we have developed together with two EURECOM Master’s students, namely Silvia Rabellino and Sofia Altare. This second method takes deep inspiration from Sohail and Bhattacharya [2007] work.

**Step 3.2: Distance Features** The set of 24 coordinate signals represent a first feature set. Nevertheless there are many different way to represent this data.

As a first step toward the generation of an intelligent, compact, and meaningful feature set we have attempted to extract some other features, from these 24, in a similar way to the one adopted by MPEG-4 Face Definition Parameters (FDPs), Face Animation Parameters (FAPs), and to the work of Valenti et al. [2007].

---

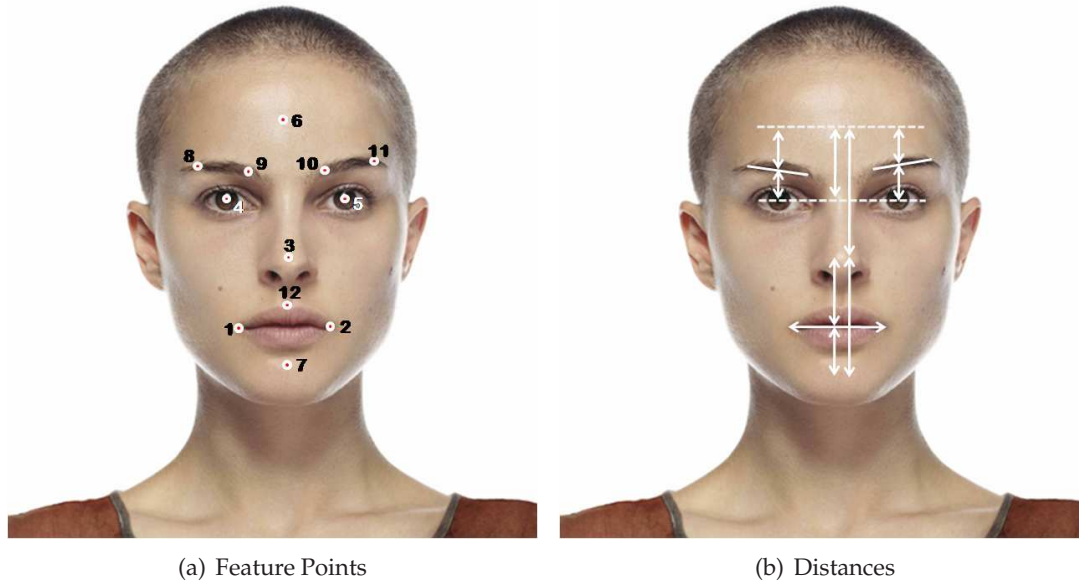


Figure 5.7: Video Features

For each subject we have computed the mean values of the 24 coordinates ( $(x_{mean}(i)$  and  $y_{mean}(i)$  with  $i = [1, 24]$ ). Secondly we have computed 14 signals ( $f(i)$  with  $i = [1, 14]$ ) by performing the following operations:

1. mouth corner distance  $f(1) = (x(2) - x(1))$
2. chin distance to mouth  $f(2) = ((y(1) + y(2))/2 - y(7))$
3. nose distance to mouth  $f(3) = (y(3) - (y(1) + y(2))/2)$
4. nose distance to chin  $f(4) = (y(3) - y(7))$
5. left eye to eyebrow distance  $f(5) = ((y(10) + y(11))/2 - y(5))$
6. right eye to eyebrow distance  $f(6) = ((y(8) + y(9))/2 - y(4))$
7. left eyebrow alignment  $f(7) = (y(10) - y(11))$
8. right eyebrow alignment  $f(8) = (y(8) - y(9))$
9. left eyebrow to forehead distance  $f(9) = (y(6) - (y(10) + y(11))/2)$
10. right eyebrow to forehead distance  $f(10) = (y(6) - (y(8) + y(9))/2)$
11. forehead to eye line distance  $f(11) = (y(6) - (y(4) + y(5))/2)$
12. head x displacement  $f(12) = ((x(3) + x(6))/2)$
13. head y displacement  $f(13) = ((y(3) + y(6))/2)$
14. size factor  $\propto$  head z displacement  $f(14) = (y(6) - y(3))$

Finally, we have computed the 14 distances signals ( $distances(i)$  with  $i = [1, 14]$ ) by dividing the found value by the mean distance between two points  $f_{mean}(i)$  (e.g.  $f(1)_{mean} = (x(2)_{mean} - x(1)_{mean})$ ). In other words, the signals  $distances(i)$  represent the ratio between the current distance and the average distance for that particular subject. It is interesting to point out that all appearance influence for the subject shall have been purged out of the features.

In this way, we have compressed the information which was carried by the 24 signals into 14 relevant movements while removing the appearance information of the subjects. We expect most of these distances to carry more emotional information than single point coordinates and, therefore, to work better for automatic emotion recognition.

It is important to notice that all the processes described here have low computational requirements and can easily be used for real time processing. Furthermore, the process could easily be written for parallel computing and speed up even more by taking advantage of the computational power of retail graphical processing units (GPU) and of libraries such as Intel's CUDA Halfhill [2008].

### 5.3.7 Prosodic Expression Feature Extraction

Our system for speech emotion recognition, takes deep inspiration from the work of Noble [2003]. We used PRAAT (see figure 5.8), a C++ toolkit written by P. Boersma and D. Weenink to record, process, and save audio signals and parameters (Boersma and Weenink [2008]), to extract some audio features. These are:

- the fundamental frequency or pitch ( $f_0$ )
- the energy of the signal ( $E$ )
- the first three formant ( $f_1, f_2, f_3$ )
- the harmonicity of the signal ( $HNR$ )
- the first ten linear predictive coding coefficients ( $LPC_1$  to  $LPC_{10}$ )
- the first ten mel-frequency cepstral coefficients ( $MFCC_1$  to  $MFCC_{10}$ )

Speech rate and pausing structure are two other features which are thought to carry emotional information but are harder to extract reliably and are not usually employed Noble [2003], Datcu and Rothkrantz [2008]. We have overviewed in section 4.2.3 what those features are and we will not cover this subject again.

This process results in a set of 26 different features. In most of the cases we downsample these signals at 25 Hz (i.e. 25 value per second) to be able to synchronize them with the video features.

It is interesting to notice that also the processing time of the audio analysis is compatible with real-time constrains.

In this section we have overviewed the procedures needed to extract 24 + 14 video and 26 audio features. Next section will discuss feature vector generation.

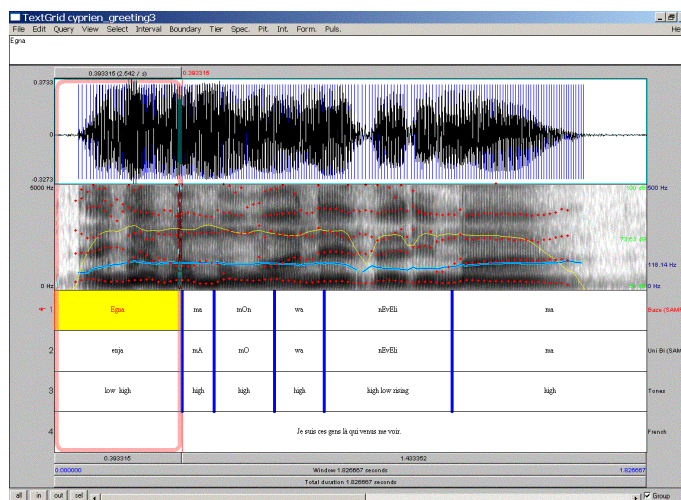


Figure 5.8: Boersma and Weenink [2008] PRAAT Interface

### 5.3.8 Feature Vector Construction

Given a set of features there are many different ways of using them to represent the original media.

We have said that some recognition system works on still images. It is possible, using the coordinate features and the vocal features to use the frame-by-frame features to classify the emotion<sup>1</sup>.

Nevertheless, We have seen that emotions and emotional expressions are a dynamic phenomena Scherer [2001] in section 4.2.1. It is therefore in the dynamics of the features that we shall expect the most emotional information to be carried. In this section we will overview few different ways that we have employed to represent the dynamics of the features signals in time. In the next sections we will detail the results we obtained with these approaches.

**Concatenation** The first and simplest way to use the features to represent the dynamics of the feature signals is to concatenate different frames of features. Nevertheless, we have detailed the processing we employ to extract 24 (or 14) visual features and 26 vocal features; it is clear that if we want to characterize the feature signals for one second using this method we have to use a classification method accepting as much as  $(24 + 26)features * 25framespersecond = 1250features$  as input. Although such systems exist this is probably not the most convenient way to represent the dynamics of the feature signals.

In this thesis, we have only performed few test with this kind of dynamics representations. The results were not satisfying, nor particularly meaningful, and, therefore, we will not detail them.

**Derivatives** A second approach to dynamics representation is the one of representing the signal through its derivatives ( $\Delta$ ,  $\Delta\Delta$ , etc.). This approach is often used for the speech

<sup>1</sup>We could also use the distances features we have defined in section 5.3.6 but it is important to notice that, since the features are normalized by their mean value, then they do not really represent a static information.

processing but easily carry the drawback of only representing the signal locally on a small time span. A possible way of overcome this limitation is to compute the derivatives on a sub-sampled, filtered signal. If we sub-sample a signal and then compute the first derivative then the resulting number represents the average first derivative on a longer time span.

**Statistical Analysis** A third, more sophisticated, way to represent the feature signal dynamics is to represent the signal on a certain time-window through the statistics of the signal itself. This is the approach taken, for example, by Noble [2003]. In our case, for each time-window we represent the signal with twelve statistical features. These are:

1. mean value *mean*
2. standard deviation *std*
3. variance *var*
4. minimum value *min*
5. minimum position in the time-window *m\_pos*
6. maximum value *MAX*
7. maximum position in the time-window *M\_pos*
8. 0.05 quantile *quantile<sub>05</sub>*
9. 0.25 quantile *quantile<sub>25</sub>*
10. 0.50 quantile *quantile<sub>50</sub>*
11. 0.75 quantile *quantile<sub>75</sub>*
12. 0.95 quantile *quantile<sub>95</sub>*

In the next sections we will show how this representation will work alone and coupled to others.

**Polynomial Regression** A last representation that we could think about is a polynomial regression of the feature signals. A polynomial regression of order  $n$  represent the signal by approximating it with  $n + 1$  parameters as a function of time. For example, a polynomial regression of order 1 is the straight line represented by two parameters that better approximate the data.

In the thesis we have firstly tried a regression of order 5 (i.e. six parameters) and then a simpler regression of order 2 (i.e. three parameters). Although the order 5 polynomial regression could better estimate the signal we realized that small modifications of the signal generally change the parameters value and that these changes had a very big dynamic. A polynomial regression of order three instead tend to be more stable but bring about the same information as mean value, mean first derivative, and mean second derivative.

---



### 5.3.9 Machine Learning

Machine learning refers to the scientific discipline that is concerned with the design and development of algorithms that allow computers to automatically learn patterns based on data, such as from sensor data or databases (Bishop [2006]). In the attempt to find the best machine learning approach for the recognition of emotions we have tested several different techniques:

- Support Vector Machines (SVM) (Hsu et al. [2003])
- Neural Networks (NN) (Bishop [1995])
- Neural Networks based on Evidence Theory (NNET) (Benmokhtar and Huet [2007a,b])
- Gaussian Mixture Models (GMM) (Bishop [2006])
- Hidden Markov Models (HMM) (Rabiner [1989])

In the followings we would give few generalities about these technologies. Given the available space, we will try to make some simplifications. The interested reader is invited to point to the reference papers cited in the above list and in particular to the more general work of Bishop [2006].

**SVM** Support Vector Machines are a set of related supervised learning methods used for classification and regression. SVM view input data (two classes) as two sets of vectors in an  $n$ -dimensional space and will construct a separating hyperplane in that space which maximizes the margin between the two data sets. To calculate the margin, two parallel hyperplanes are constructed, one on each side of the separating hyperplane, which are “pushed up against” the two data sets. Intuitively, a good separation is achieved by the hyperplane that has the largest distance to the neighboring data points of both classes. In general the larger is the margin the lower is the generalization error of the classifier.

**NN** Neural Networks (also referred to as Artificial NN) are a mathematical / computational tool which simulates the functioning of biological neural networks. NN consists of an interconnected set of *artificial neurons*. Each neuron receives one or more inputs and sums them to produce an output (simulating the synapses). Each input is weighted, and the sum is usually passed through a non-linear function known as *activation function* or *transfer function*. By accurately setting the weights for the sums of all neurons in the networks, NN can successfully classify patterns. The correct setting of these weights is searched with a phase of statistical learning in which the neural network is supervisedly fed with some data and the output error is backpropagated a final condition is satisfied.

**NNET** Probabilistic methods have an inherent limitation: most handle imprecision but ignore uncertainty and ambiguity of the system. Evidence theory allows to deal with these scenarios by explicitly taking into account the uncertainty of the data. NNET are an improved version of Radial Basis Function neural network based on evidence theory created by Benmokhtar and Huet [2007a]. NNET are basically NN with one input layer  $L_{input}$ , two hidden layers  $L_2$  and  $L_3$  and one output layer  $L_{output}$  (see figure 5.9).



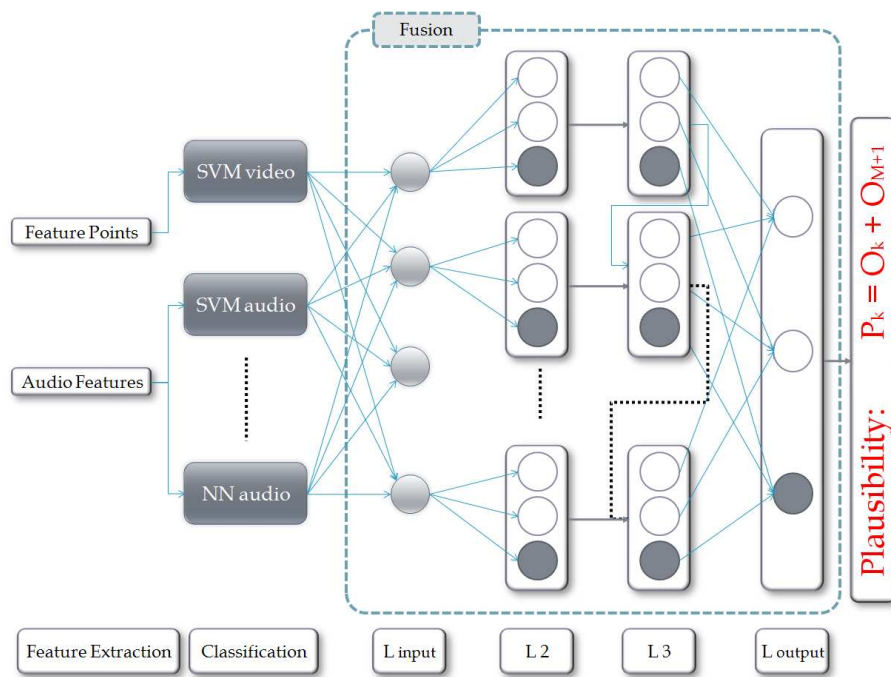


Figure 5.9: NNET classifier fusion structure

**GMM** Gaussian Mixture Models (GMMs) are among the most statistically mature methods for clustering. GMMs construct an estimate, based on observed data, of the data unobservable probability density function.

The simplest GMM consist of one Gaussian per mixture on one-dimensional data. In this sample case the GMM will model the data belonging to the different classes  $C$  with the mean and standard deviation values ( $m(C)$  and  $std(C)$ ). When trying to use this model for the recognition the distances between the value we want to recognize and all the classes  $m(C)$  are computed and weighted by the  $std(C)$ . The smallest “weighted distance” will represent the recognized class.

**HMM** Markov models are mathematical models of stochastic processes that generate random sequences of outcomes according to certain probabilities. Hidden Markov Models represent a finite set of states, each of which is associated with a probability distribution (e.g. GMM). Transitions among the states are governed by a set of probabilities called transition probabilities. In a particular state an outcome or observation can be generated, according to the associated probability distribution. It is only the outcome and not the state to be visible to an external observer and therefore states are “hidden” to the outside.

With HMM we can approach three problems:

1. **The Evaluation Problem:** Given an HMM ( $\lambda$ ) and a sequence of observations ( $O = \sigma_1, \sigma_2, \dots, \sigma_T$ ), what is the probability that the observations are generated by the model? ( $P(O|\lambda)$ )
2. **The Decoding Problem:** Given a model ( $\lambda$ ) and a sequence of observations ( $O = \sigma_1, \sigma_2, \dots, \sigma_T$ ), what is the most likely state sequence in the model that produced the observations?

3. **The Learning Problem:** Given a model ( $\lambda$ ) and a sequence of observations ( $O = \sigma_1, \sigma_2, \dots, \sigma_T$ ), how should we adjust the model parameters  $\Lambda, B, \pi$  in order to maximize ( $P(O|\lambda)$ ) ?

For emotion recognition we want to train the model on some emotional data (task 3) and test it on new observed samples (task 1).

### 5.3.10 Multimodal Fusion

Emotions are a multimodal phenomenon. Multimodal fusion refers to the task of mixing information coming from different modalities (e.g. audio and video) to improve the quality of the estimates. Multimodal information fusion may be performed at different levels and usually the following three are considered (see Figure 5.10):

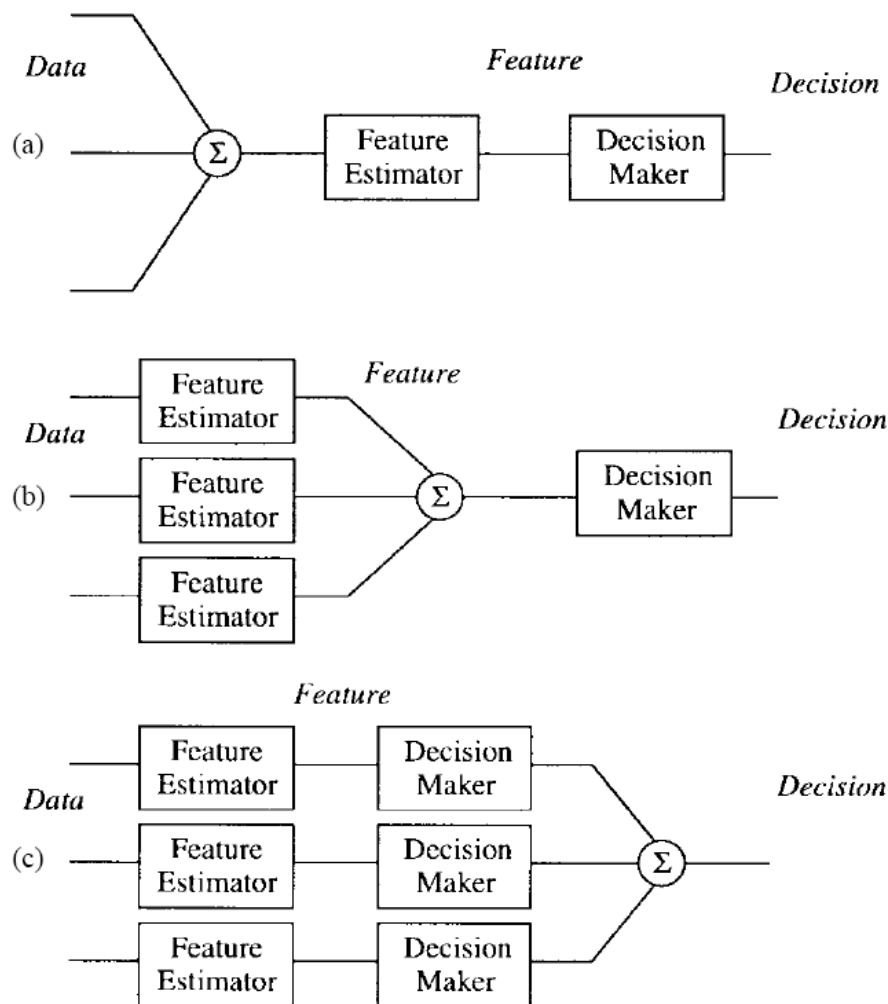


Figure 5.10: Three levels for multimodal fusion (from Sharma et al. [1998]): a) data or signal fusion; b) feature fusion; c) decision fusion

- Signal level

- Feature level
- Decision or Conceptual level

Fusing information at the *signal level* means to actually mix two or more, generally electrical, signals. This can only be done for very coupled and synchronized signals that are of the same nature (e.g. two vocal signals, two webcam signals, etc.). For multimodal fusion this is not often feasible as different modalities always have different captors and different signal characteristics. General frameworks for fusion should include the possibility to fuse at this first level as different, but similar captors may be needed to get better quality on a single modality (e.g. three sensors for the three red, green and blue color components of an image as in high fidelity cams, or microphone arrays).

Fusing information at the *feature level* means mixing together the features outputted by different signal processors. Features must be pseudo-synchronized in order to provide satisfactory results. For example features can be the position of some feature points extracted from a video processor and the prosodic features of a speech signal. This approach guarantees for multimodal fusion a good amount of exploited information but it has some drawbacks. Combining at the feature level needs synchronization and it is more difficult and computationally intense than fusing at the conceptual level (see next) because the number of features is more important and features may have very different natures (e.g. distances and times).

Combining information at the *conceptual or decision level* does not mean mixing together features or signals but rather the extracted semantic information. This implies combining representations obtained from different systems that may also be correlated just at the semantic level (e.g. positions of object, with speech indicating them). Decision level fusion has the advantage to avoid synchronization issues and generally to use simple algorithms to be actually computed.

A complete example showing the three possible levels of fusion may be a multimodal speech recognition system. Let us assume that we want to create a high fidelity gesture and speech recognition system. One of the first things we can do to improve the results is to implement a microphone array to be used in place of a single microphone. The audio signals coming from the different microphones can therefore be fused together to get a better audio signal (*signal level fusion*). This improved signal can be treated to extract audio features (e.g. phonemes). Those features may be coupled with video features obtained from a lip movement recognition system (e.g. visemes) (*feature level fusion*). These coupled features can be used to understand the speech part of a voice-gesture command. At a *decision level fusion*, we can fuse the information about what was said with the information coming from a gesture recognizer and understand to which objects some words refer to (as in the sentence "Put that there").

Several works have discussed multimodal fusion; in particular Sharma et al. [1998] resume the main issues and techniques of multimodal fusion. Sharma et al. [1998] resume the main issues and techniques of multimodal fusion. Several works on multimodal fusion have been developed which follow the well known "*put that there*" paradigm Bolt [1980] in which speech and gesture recognition are fused to interact with a 3D environment; see for example works from Bolt [1980], Corradini et al. [2003], Liao [2002], Kettebekov and Sharma [1999].

More recently some works have described how multimodal fusion mechanisms can be used for emotion/affect recognition, see as example works from Pantic and Rothkrantz [2003], Sebe et al. [2005b], Li and Ji [2005], Busso et al. [2004], Chen et al. [1998] where

---

vocal and facial emotion recognition are performed together to reach better results. Generally fusion is performed at the feature level and the usual clustering techniques are applied to the training data.

Busso et al. [2004] compare the feature level and the decision level fusion techniques, observing that the overall performance of the two approaches is the same, although they present different weakness and strength. Busso et al. conclude that the choice of the approach should depend on the targeted application.

In the case of multimodal emotion recognition, to fuse at the decision level would mean to mix together  $emotion_1$  (extrapolated for example from video signal),  $emotion_2$  (extrapolated from audio signal) etc. If the fusing mechanism is reliable enough then the recognition of the found emotion would be more precise. In other words:

$$P(multimodal) = f(emotion_1, \dots, emotion_n) > P(modality_i);$$

$$\forall i \in [1, n];$$

Where  $P(modality_i)$  represents the precision of  $emotion_i$  which is recognized using the data from the  $modality_i$ . One should note that each  $emotion_i$  component derives from different treatment processes and probably from different input signals.

To fuse at a feature level would mean, for example, to mix together  $features_1$  which will be a set of  $n$  features obtained from video (in the case of the CPT are sets of Ekman's Action Units Ekman [1971], Ekman and Friesen [1978], Ekman et al. [2002]) to  $features_2$  ( $m$  features) obtained from audio (with Scherer's theory *pitch, energy, low frequency energy and duration*) and to search clusters in the multi-dimensional space of dimension  $n + m$ . In this case it is assumed that:

$$P(emotion^*) = f(feature_1, \dots, feature_n) > P(feature_i);$$

$$\forall i \in [1, n];$$

Where  $P(feature_i) = f(feature_i)$  represents the precision of the recognition obtained while using the  $feature_i$ . And the results from Busso et al. [2004] show that:

$$R(emotion^*) \simeq R(emotion);$$

Finally, fusing at the signal level makes the assumption that:

$$P(emotion'') = f(signal_1, \dots, signal_n) > P(signal_i);$$

$$\forall i \in [1, n];$$

Where  $P(signal_i) = f(signal_i)$  represents the precision of the recognition obtained while using the  $signal_i$ .

One of the main limitations of these systems is that they are not upgradeable to new modalities and algorithms. In other words, the described fusion algorithm are designed ad hoc for fusing information coming from two (or more) specific unimodal systems but cannot accept new modalities.

Moreover, usually the existing fusion algorithms are not adaptive to the input quality and therefore do not consider eventual changes on the reliability of the different information channels. A third limitation is that the systems cannot take in account long term affective phenomena; in other words they work on short sequences of video and sound (2 to 5 seconds) and cannot consider affective phenomena such as mood or affect of the user which need to be considered over longer times.

In the following section we propose a system where the different fusion algorithms are automatically able to take into account new uni or multi modal emotion recognition systems and to dynamically adapt to the various channel conditions in order to find an optimal (not always the optimum) solution. The recognized emotion will be computed in real time and considered by our cognitive architecture on different time scales in order to get estimations of the different affective phenomena. In other words the estimated appraisals of the affective processes (i.e. 25 estimations per sec) are considered in time windows and averaged to get different estimations for the different affective phenomena (emotions, moods, affects and personalities). Therefore these affective phenomena estimations are used to describe the user affective model.

### 5.3.11 AMMAF: A Multimodal Multilayer Affect Fusion Paradigm

We hence, propose a framework to perform multimodal fusion at the three possible fusion levels (see figure 5.11) (see Paleari and Lisetti [2006c] for further details).

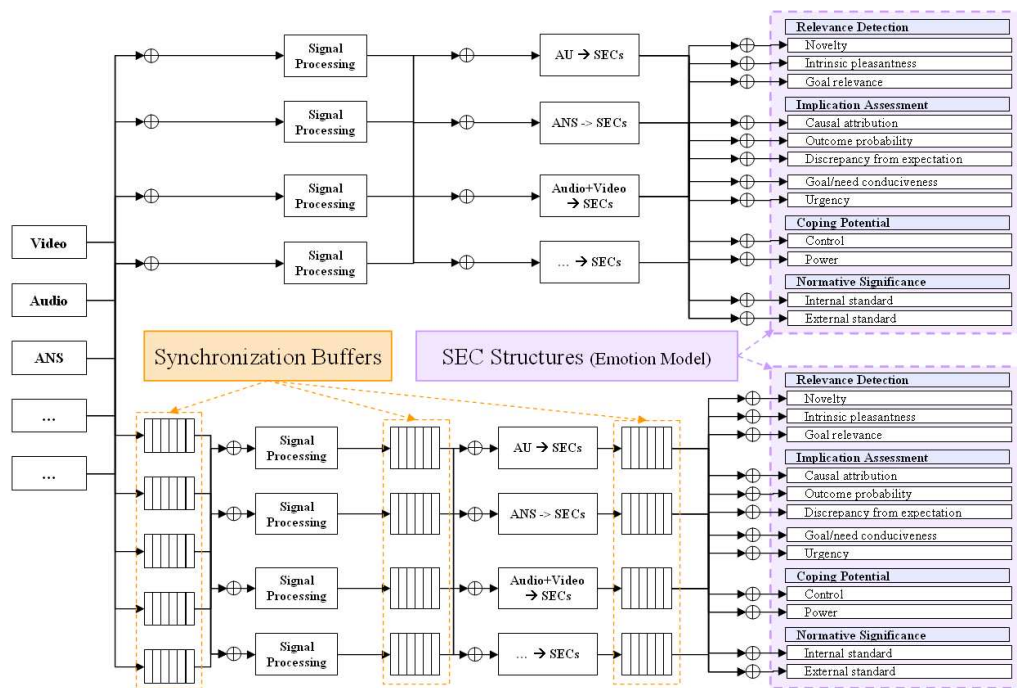


Figure 5.11: Double chain for signal, feature and decision realignment

The objective is to develop a system allowing researchers to add their input signal, if missing, and insert their emotion recognition system into a multimodal context. New multimodal systems will be allowed to use signals, or features from other inputs, and would have constraints on the output format (SECs and confidence values).

In our proposition decision or conceptual level fusion is automatically done by the architecture but is also tunable through the modification of simple text files.

In particular the algorithm analyzes different possible fusion algorithms (e.g. maximum, voting, averaging combining, and product combining) and automatically chooses the most stable one during a training phase.

In the future, the algorithm would also possibly be controlled automatically by the cognitive architecture which will be able to deliberately simulate the user's appraisal of the surrounding events and eventually manipulate it.

The framework, shown in figure 5.11, has been thought for interacting with ALICIA, the affective intelligent architecture we describe in part 6, and works on Scherer parameters. The output of each emotion recognition system must be a vector representing an appraisal in terms of SECs coupled with a confidence value used by the algorithm to fuse data.

Furthermore, two different fusion chains (see figure 5.11) would be active in parallel. The first chain, at top in figure 5.11, will treat close to real time signals and interpretations returning fast interpretations of the recognized emotion. The second chain will work on bufferized and re-aligned data in order to have the possibility to resynchronize data just before fusion.

There are two main reasons for this bufferized approach:

1. there are modalities, like ANS signals that are very interesting and apparently reliable but that have responses time in the order of a dozen of seconds and will not be usable for real time purposes.
2. We are interested in having a better, more accurate appraisal of the user affective state, regardless the computational time it will take.

In other words the objective of this double chain would be to have both a fast but less reliable and a longer but more accurate evaluations of user affective states. Sensory motor (and behavioral) processes would probably use the fast appraisal while conceptual (and sometimes behavioral) processes would use the longer but more accurate version.

Finally the different fusion algorithms may be able to control the dimensions of the resynchronization buffers. In other words, if one algorithm, comparing the different estimations, observes that the outputs coming from the system exploiting ANS signals at time  $t$  correspond to those coming from the other multimodal signals at time  $(t - n)$  it may control the length of the ANS buffer in order to realign the different evaluations.

Buffers at feature and decision fusion levels are used by the algorithms for searching resynchronization patterns; the signal fusion level buffer is then the one used for the actual resynchronization given the commands coming from the algorithms working on the two higher level buffers.

Constraints would be applied to assure the stability of the system by insuring that buffers length cannot diverge. The resynchronization algorithm working on the signal level buffers will not be able to de-align  $signal_a$  and  $signal_b$  more than a certain time  $t_{a-b}$  or less than the time  $t_{b-a}$ .

In both cases of fast and bufferized emotional responses the resulting emotions will be evaluated averaging on different time windows (e.g. 1 sec, 3 sec and 10 minutes) to be able to take into account different affective phenomena, e.g. fast emotional responses like surprise or fear, conceptual emotions like contempt or pride but also moods and affects.

The algorithms make use of the Scherer [2001] component process theory emotion representations. The main reasons for basing this approach on Scherer's theory are three:



1. Scherer's theory models in a very detailed and psychologically grounded way the appraisal process of emotions.
2. CPT allows for a three level model of emotions.
3. CPT links both emotion generation and emotion recognition to the process of appraisal and therefore to the user and agent models.

The use of such a kind of framework impose some constraints, and in particular the use of Scherer emotion representation (see last column of figure 5.11) for the outputs but it also brings three main advantages:

1. The fusion framework automatically takes into account new recognition system.
2. The algorithm computes the channels / recognition system reliabilities, comparing the different recognition system evaluations and adapting the algorithm to find optimal solutions.
3. The complete system treats in parallel two chains, one near to real-time and one bufferized and more reliable one.

In this chapter we have reviewed the phases involved with the extraction of emotional estimates from different modalities and proposed AMMAF, a framework for multimodal fusion of affective cues. In the next chapter we present some of the results we obtained while comparing different possible systems and conclude presenting ARAVER's actual design and obtained results.

## 5.4 Results

In this chapter we will review some of the results we have obtained by combining the different ROI estimation, feature point detection, feature vector construction, and machine learning approach. Most of these works have been presented in international conferences and workshops (Paleari and Lisetti [2006c], Paleari et al. [2007a,c], Paleari and Huet [2008], Paleari et al. [2009a]).

### 5.4.1 Metrics

In this work we make use of 5 main metrics for evaluating the results of the recognition. These are:

1. classification rate of positive samples:  $CR_{emotion}^+ = \frac{\text{correctly\_classified\_samples}_{emotion}}{\text{samples}_{emotion}}$
2. average recognition score:  $m(CR^+) = \frac{\sum_{emotion=1}^6 \text{correctly\_classified\_samples}_{emotion}}{\sum_{emotion=1}^6 \text{samples}_{emotion}}$
3. standard deviation of  $CR_{emotion}^+$ :  $std(CR^+) = \sqrt{\frac{\sum_{emotion=1}^6 (CR_{emotion}^+ - m(CR^+))^2}{6}}$
4. weighted standard deviation:  $wstd(CR^+) = \frac{std(CR^+)}{m(CR^+)}$
5. average recall:  $R = \frac{\sum_{emotion=1}^6 \text{classified\_samples}_{emotion}}{\sum_{emotion=1}^6 \text{samples}_{emotion}}$



$$6. \text{ average precision: } P_R = \frac{\sum_{emotion=1}^6 \text{correctly\_classified\_samples}_{emotion}}{\sum_{emotion=1}^6 \text{classified\_samples}_{emotion}}$$

The objective of a recognition system is twofold:

1. The system should perform the **best possible average recognition score**.
2. The system should return the **lowest possible standard deviation of  $CR^+$** . This translate in finding the best average score while minimizing the difference between the minimum and the maximum  $CR_{emotion}^+$  or in other words to maximize also the minimum  $CR_{emotion}^+$ .

Loosely speaking, generally we would rather prefer a system which recognizes 6 emotions at 50% than one recognizing 3 emotions at 95% and 3 at 10% even though the average recognition score ( $m(CR^+)$ ) would be the bigger in the second case (i.e. 52.5% against 50% of the first system).

The weighted standard deviation  $wstd(CR^+)$  is a measure we have defined to weight the standard deviation  $std(CR^+)$  of the results by the average recognition rate  $m(CR^+)$ . Indeed it is normal that a system with a higher average recognition rate also has a higher standard deviation. Instead, we would like to have a measure of quality which is independent of the average recognition rate.

In practice this is equivalent to normalizing the maximum  $CR_{emotion}^+$  result to 1 (and consequently the others) before computing the standard deviation.

As an example a system performing  $CR^+ = \{50\%, 50\%, 50\%, 100\%, 100\%, 100\%\}$  will have  $m(CR^+) = 75\%$ ,  $std(CR^+) = 0.274$ , and therefore  $wstd(CR^+) = 0.365$ . A similar system performing  $CR^+ = \{25\%, 25\%, 25\%, 50\%, 50\%, 50\%\}$  will have  $m(CR^+) = 37.5\%$ ,  $std(CR^+) = 0.137$ , but the same  $wstd(CR^+) = 0.365$ .

One might notice that the classification rate of positive samples is the same as precision when  $\text{classified\_samples} = \text{samples}$  (i.e. when all samples are classified) and therefore when the recall is equal to 1.

In many cases chance play an important role on the results because of the randomness part intrinsic in most training phases. Albeit, in many of the results we present in the next few sections this is not specified, the presented results depict a particular output but are always motivated by several similar outputs which we have obtained during our studies. Therefore, if not differently specified, when we show that the system A performs better than system B for emotion recognition using a specific machine learning approach, then different machine learning approaches would lead to similar conclusions. Similarly, if not differently specified, when we show that system A performs better than B for video emotion recognition than the conclusions made for video can be generalized to audio emotion recognition too.

## 5.4.2 Dense Motion Flow vs. Feature Point

The first comparison we want to present is the comparison between our feature point system and a dense motion flow technique. In figure 5.12 we show a result obtained training two neural networks (15 neurons in the hidden layer) with the frame-by-frame feature points positions as we have defined them in section 5.3.6 rather than point positions defined by a regular 8x8 grid.

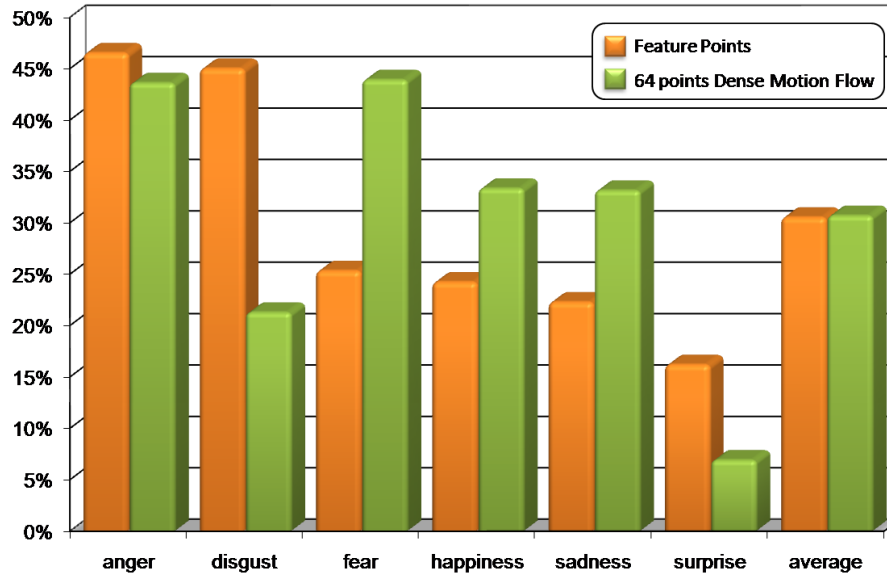


Figure 5.12:  $CR^+$  result comparison between a system based on feature points and one based on 64 points dense motion flow

Without looking at the absolute results, that have here only limited interest<sup>2</sup>, we observe that in average the two systems perform about the same. Nevertheless, the standard deviation of the results changes and in particular is lower for the system based on FP ( $std(CR^+) = 0.127$ ) and higher for the system based on dense motion flow ( $std(CR^+) = 0.142$ ).

This first result validates the use of the extracted feature points and shows a slight improvement of the performances according to the metrics we have defined in section 5.4.1.

### 5.4.3 Audio vs. Video vs. Audio-Video

In this section, we try to demonstrate the advantage of a multimodal approach taking into account both audio and video. Supposing both monomodal system works separately, this kind of approach bring about two main advantages with respect to each one of them:

1. when one of the two modalities is not available then the system could still use the other to estimate the emotion;
2. when both modalities are available, the diversity and complementarity of the information, should bring about a general improvement of the system performances.

While the first point is obvious, the second one might be objectable. In figure 5.13 we report the average and standard deviation recognition rates of the positive samples

<sup>2</sup>In the next sections we will see that the results can be improved with different settings of the machine learning algorithms, by using dynamic information, and by filtering out some noise from the signals.

for three systems exploiting NN<sup>3</sup> for classify video, audio, and multimodal audio–video data respectively.

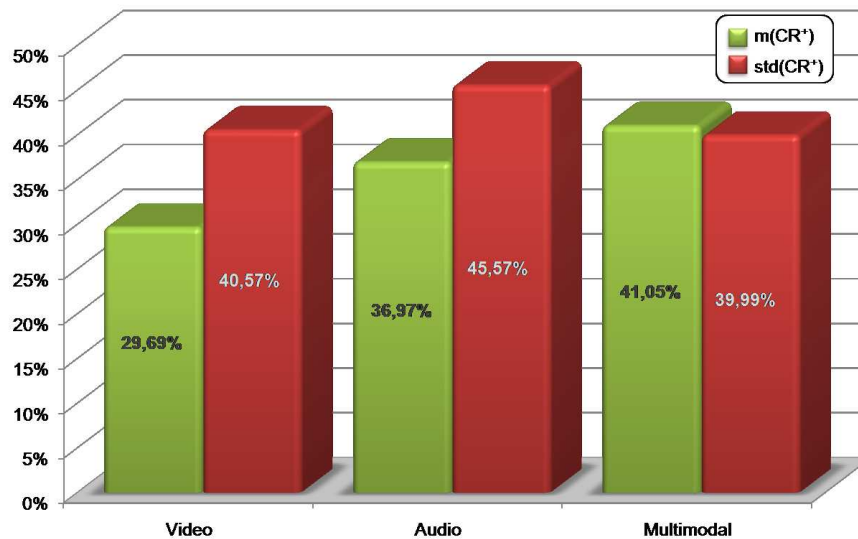


Figure 5.13:  $m(CR^+)$  and  $wstd(CR^+)$  results comparison among video, audio, and multimodal audio–video approaches

As can clearly be seen from this example the system based on multimodal audio–video information outperforms both the system based on video and the system exploiting audio both in terms of average recognition rate and weighted standard deviation.

#### 5.4.4 Feature vector: statistical or polynomial analysis

It can be interesting to know which feature vectors better represent the emotional signal. We have said that, in a first time we have used two different different types of analysis; in this section we briefly compare them. In the next section we will compare, among others things, the raw signal and its representations through the first two derivatives.

In figure 5.14 we show the comparative results of three systems based on NN working respectively on polynomial data only, raw data only, statistical data only, and the set of both statistical and polynomial data.

As it can see from these results statistical analysis outperforms the polynomial analysis performing slightly better than the raw data too. The system exploiting both polynomial and statistical data outperforms all the others in terms of average recognition rate at the cost of a slightly increased complexity.

This result seems to motivate the use of statistical vector analysis for the representation of the feature signals.

As we will show in the next few sections one important factor is data pre and post filtering. Using statistical characteristics of the signals to represent the features is, in part, equivalent to pre–filtering the data. In particular the mean value of the signal can be seen as the result of a phase of low–pass filtering.

<sup>3</sup>20 neurons, 25 frames (i.e. one second) polynomial (5<sup>th</sup> order) and statistical representations (feature fusion by concatenation).

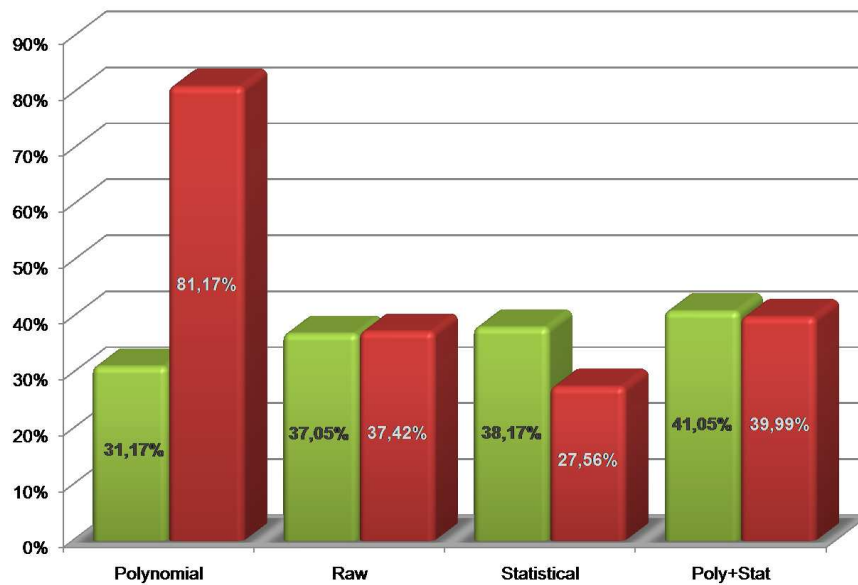


Figure 5.14:  $m(CR^+)$  and  $wstd(CR^+)$  results comparison among different feature vector representations

### 5.4.5 Comparison Among Features

In this section we would like to report a study we have done on feature selection for multimodal emotion recognition. The objective of this section is therefore to understand which features, from the one we have seen before, carry more emotional information. In doing this we are also trying to understand which part of the feature time–signal is responsible of this transfer of information, whether that is the static or dynamic part.

Sections 5.3.6 and 5.3.7 shown how to extract  $24 + 14 + 26 = 64$  different features from video and audio which are related to the emotional expressions. To these 64 individual features we add, for this study, some sets of features by grouping and concatenating them. These sets have been defined as it follows:

- sets of variables from coordinates
  1. mouth region coordinates
  2. eyes region coordinates
  3. nose coordinates
  4. nose and forehead coordinates
- sets of variables from distances
  1. mouth region distances
  2. eyes region distances
  3. head displacements
- sets of audio variables
  1. pitch and energy

2. audio formants
3. LPC coefficients
4. MFCC coefficients

This has been done with the purpose of gathering the information from different features belonging to the same set together. We expect sets of features to perform better for emotion recognition than each one of the single features. Furthermore, we want to compare different groups (e.g. regions of the face) to each other in order to better understand which ones are more interesting for automatic emotion recognition and which one need further development or finer precision.

### 5.4.6 Feature Extraction

As a result of the operations detailed in sections 5.3.6, 5.3.7, and 5.4.5 we have defined 75 sets of individual or multiple features.

For each one of these sets of feature  $f_i$ , we need to extract a feature vector which best represent the affective information. It is expected that affective information is transferred via the dynamics of the facial and prosodic expressions Scherer [2001]. In order to incorporate dynamics to the framework, we have taken partially overlapped sliding windows  $w(f_i)$  of the signals. To be able to understand which window lengths better represent the emotions we have let the size of the window vary from 1 to 50 frames; longer time windows will carry more information about the dynamics of the signal.

Furthermore, in addition to the original signal we investigate its dynamic properties. In particular we consider the following characteristics:

- the feature's values in time  $t$
- the feature's first derivative  $\Delta$
- the feature's second derivative  $\Delta\Delta$

We have anticipated that some statistical characteristics of the signal inside a time window may be interesting as well. For this reason we have considered, beside the signal in time, its mean and standard deviation; therefore, for each one of the three time analysis mode we consider:

- the raw feature values  $raw(w(f))$
- the windows mean values  $mean(w(f))$
- the windows standard deviation values  $stdev(w(f))$

Thanks to the analysis and study of the obtained results, we aim at better identify the most significant features for emotion recognition and possibly at better understanding the process of production of multimodal emotional display.

### 5.4.7 Analysis Procedure

For each possible combination of *feature\_set*, *mode* and *window\_size*, 5 different Neural Networks (NN) have been trained and the different scores averaged. This was done in order to reduce the “randomness” intrinsic in NN training.

We have randomly split the database into two parts for test and training having the shrewdness to select different subjects for the *Test* and *Train* bases. Forty subjects were used for training and four for testing.

The idea of keeping four subjects for testing instead of one originates from the need of having reliable results without performing a complete leave one out test<sup>4</sup>. Four subjects provide a reasonable amount of testing samples and represent meaningful infra-subject differences without impacting too much the size of the training base. To validate the result we have tried part of the NN trainings with two different sets of training and testing subjects and checked that the results were significantly consistent.

Since some emotions are represented by an higher number of frames (and therefore training samples), we have, eventually, normalized the number of samples for each emotion to the one emotion which was represented the least. This was done by randomly skipping samples of the emotions that were represented more often.

All tests were carried out using the MATLAB neural network toolbox. We used 10% of the samples in the *Train* set as validation samples. We have set a variable number of neurons for the input layer (i.e. one per input feature), 20 neurons as hidden layer, and 6 neurons (one per emotion) for the output layer.

The size of the hidden layer has been chosen arbitrarily; our preliminary tests did not show any meaningful improvement given by a further increase in the size of the hidden layer and we are not, in this study, looking for the best possible performance. A question that could arise is whether the number of neurons in the hidden layer should be adapted to the size of the feature vector used as input. Although this may optimize the results, it was impossible with this number of experiments (more than 100.000 NNs) to take the number of the neurons on the hidden layer as an extra free parameter. We think that, in this case, it is more suitable to present results obtained with a common parameter for every training (knowing that sets of variables and longer time windows might be disadvantaged) than choosing arbitrarily a rule setting the number of hidden neurons.

Finally, the maximum number of epochs has been set to 50. In our preliminary tests this value resulted as being a good one avoiding to over-fit the testing data while giving acceptable results. Once more, this number could have been set according to the complexity of the network and/or of the inputs but we preferred to keep the value unchanged and to keep in mind that sets of feature could be disadvantaged for this reason.

For each variable we have then searched the best average score for the recognition rate of the positive samples  $m(CR^+)$ <sup>5</sup>.

Some variables may work very well for the recognition of some specific emotions but below average for the recognition of the others. To evaluate this phenomenon, we have evaluated, for each emotion and each variable, the best obtainable score. In this case, the free parameter, for finding the  $best\_CR^+$  out of all the different  $CR^+$  is the triple composed by the *window\_size* ([1,50]), the *feature\_set* ( $t$ ,  $\Delta$ , or  $\Delta\Delta$ ), and the *mode* (*raw*,

<sup>4</sup>Indeed, while performing more than 100.000 different NN training, it is computationally too demanding to also perform a complete leave one out test.

<sup>5</sup>Please note that, for each triple, the  $m(CR^+)$  is obtained by computing the mean confusion matrix among the five trained neural networks.

*mean*, or *stdev*).

Finally, a new measure is defined as follows:

$$average(CR^+) = \frac{\sum_{i=1}^6 best\_CR_{emo}^+}{6}$$

Please notice that  $average(CR^+) \neq m(CR^+)$  because the number of samples in the test base is not the same for each emotion and because  $average(CR^+)$  is found as an average of results of 6 different classification systems. With this measure we would like to have a measure of the maximum possible result obtainable while fully exploiting one variable. Indeed, one emotion might be better estimated by the mean value of the  $\Delta\Delta$  values over a 50 frames window, while another may be better estimated by the simple value of the features.

Analyzing the obtained results could help understanding which variables are, in the case of speaking subjects, more helpful for the task of emotion recognition. Furthermore, the detailed results about the different settings, emotion by emotion and variable by variable, could help understanding which features are more discriminating each single emotion. In the future a system could be designed to take advantage of all the best settings for each variable to optimize or simplify the classification of emotions.

In the next section we will present the results of this study.

### 5.4.8 Results

In this section we report the outcome of this extensive experimental study.

Before reporting the actual results, we would like to discuss few assumptions that we have made basing on our a-priori knowledge of the problematics involved with the automatic recognition of emotions.

Since the subject depicted on the videos are talking, we expect the visual features of the eye region to perform generally better than the features of the mouth region. The later are, indeed, more influenced by the movement involved in speech production. In effect, speech movements (i.e. visemes) are, for the emotions, comparable to noise. On the other hand, this issue could be at least partially solved by taking longer time windows and/or using the *mean* feature as our system does.

Secondly, we expect some coordinates to be more significant than others; we know, for example, that eyebrows do not slide horizontally on the face. Therefore, we can expect the  $y$  movement of the eyebrows to be more significant than the relative  $x$  movement.

A similar reasoning can be made for the forehead or the nose, but these points are supposed to be fixed on the face and can, therefore, represent the movement of the head.

The eyes  $y$  coordinates are biased by blinking and the definition of the video may not always be sufficient to detect the small vertical gaze movements. We do not expect this coordinate to be very interesting for automatic emotion recognition.

We acknowledge that sometimes the point on the forehead tend to drift horizontally from his original position when the head pans (i.e. rotates horizontally); this is an issue of the Lucas-Kanade algorithm caused by the fact that a point is characterized locally. We expect this coordinate to perform worse than others.

For the distances, the general statement made before about the fact that we expect distances on the eye regions to perform better than the ones involved on the mouth region remains valid.



Concerning audio features, it is known (Scherer [2001]) that pitch and energy should carry a simply exploitable source of emotional information. Pitch is supposed to be generally higher for positively valenced emotions (e.g. happiness) and the other way round lower for negatively valenced emotions (e.g. sadness and anger). Energy shall increase for arousing emotions (e.g. anger or happiness) and vice versa decrease for unexciting emotions (e.g. sadness and boredom). These two variable should therefore help distinguish among emotions families, and the set of the two should work quite well distinguishing almost all emotions. Nevertheless, it is important to notice that not all of the speech signal is voiced and that therefore pitch is not always available possibly resulting to the detriment of the final result.

Harmonicity is linked to the stress on the voice, and it has been demonstrated that it carries emotional information Noble [2003]. We expect this feature to provide useful insight into emotions.

LPCs and MFCCs are often used for this kind of studies Noble [2003], Busso et al. [2004]. Both these sets of coefficient should represent mathematically the characteristics of the vocal tract and therefore are thought to carry an important amount of emotional information. The first variables of each set carry the main information about the vocal tract. Therefore, we expect them to carry the main information of the emotion too. The same hypothesis can be made for the formants.

Generally, we expect the sets of features to perform better than the single features. This is because the information of the one feature can completed from the information carried by the others. Nevertheless, we shall keep in mind that, because of our testing settings, sets of features are partially disadvantaged to single features. This happens because more information also comes with much more complexity. Furthermore, the optimum will be reached only if all the features belonging to a set did perform the best with the same mode and time window (e.g.  $mean(\Delta)$  on 34 frames).

Finally, we expect some emotions, and in particular surprise, to be more difficult to recognize accurately than others. The case of surprise is a particular one since humans rarely express unvalenced surprise. Surprise is either positively valenced (as in the case of a sudden happy event) or negatively valenced (as in the case of an unexpected scary event); in either of these cases surprise will be easily confused with happiness or fear/anger.

Of all the emotions under investigation, sadness is the only one characterized by low arousal values. We can expect that this may cause this emotion to be better recognized than the others through the study of many different variables.

In the following sections we will comment the actual results making frequent references to the expected outcomes that we have just overviewed. We will observe that many of our predictions and hypothesis are corroborated by the actual results and some are not.

**Video features** Here, we present the results relative to the video feature sets. For each one of the two feature sets (i.e. “coordinates” and “distances”) we present and comment the average  $CR^+$  result for each emotion as well as the detailed results obtained for each variable and each emotion and some statistics about the analysis modes and the time window lengths that generated these outcomes.

**Coordinate Features.** In this section, we report the results obtained from the coordinates of the 12 feature points and from the set of features of the same kind.

---

In figure 5.15, we can see the average results of the classification rate for the positive samples  $average(CR^+)$ .

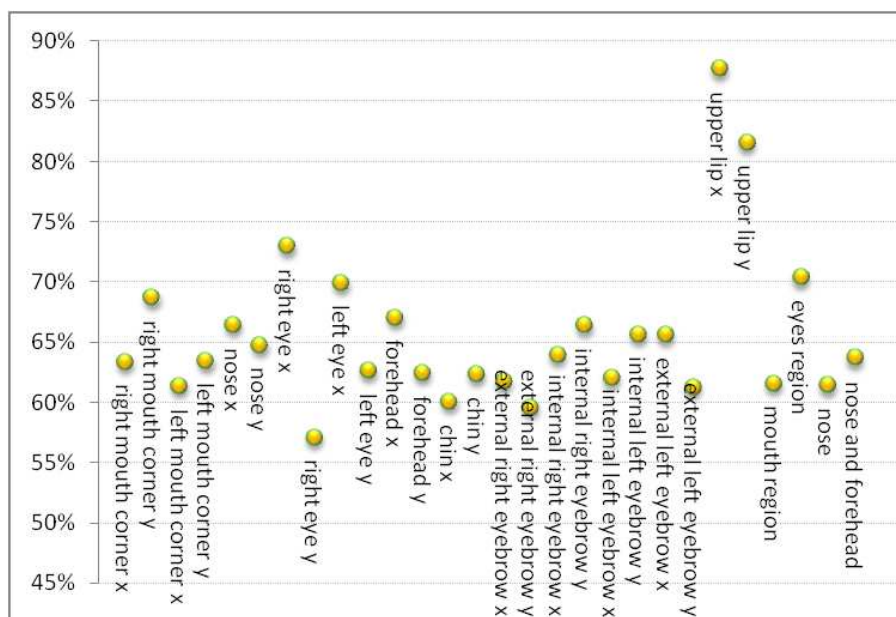


Figure 5.15: average  $CR^+$  for the coordinate features

We can observe that the feature performing the best are the coordinates relative to the upper lip, the worst one is the right eye  $y$  coordinate. The best set of features is the one grouping the coordinates of the eye regions with roughly 70%. In average the coordinate features perform 65.5% (see figure 5.21). We point out that, for the mouth (corners and chin), the  $y$  coordinates works better than the  $x$  for emotion estimation while, for the eyes, the forehead, and the upper lip, the coordinate  $x$  works better. We were expecting these results. The horizontal movement of the chin and mouth corner points is quite limited and/or very much influenced by the voice production; furthermore, we expected blinking to affect the vertical movement of the points belonging to the eyes with few, if any, correlation to the displayed emotion.

	t	$\Delta$	$\Delta\Delta$
raw	14	14	9
mean	21	16	21
stdev	17	21	35

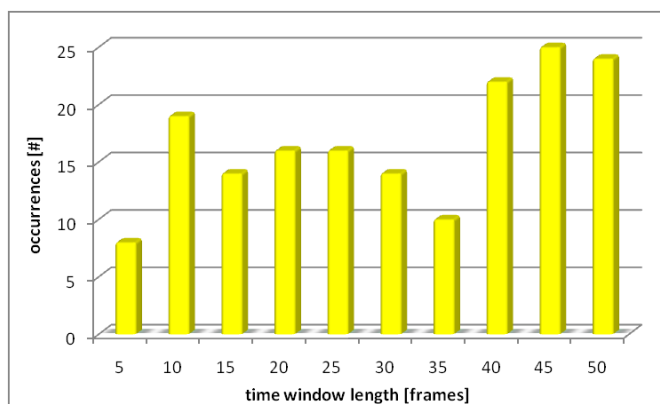


Figure 5.16: Modes and window lengths for the coordinate features

In figure 5.16 (and table), we show the distribution of the modes which have been

selected to get this result. We can notice that the  $\Delta\Delta$  variable is slightly preferred to both  $\Delta$  and *time* analysis and that standard deviation (*stdev*) is being preferred to both *mean* and *raw* signal. We point out that the study of the raw signal is slightly disadvantaged to the other two modes due to the increased size of the feature vector and therefore to the increased complexity of the NN training.

From the same image we can observe the histogram of the window lengths which have been selected, as the one returning the best results. From this graph we observe that, in the case of the coordinate features, windows longer than 35 frames are preferred concentrating around 50% of the best trainings.

In table 5.3 we report the  $m(CR^+)$  score of the single mode which in average return the best result. Obviously here the results are much lower; this is due to the fact that now the same triple of *window\_size* ([1,50]), *feature\_set* (*t*,  $\Delta$ , or  $\Delta\Delta$ ), and *mode* (*raw*, *mean*, or *stdev*) is used for all emotions. In this case, we observe that the the coordinates of the points belonging to the mouth perform slightly better than the ones of the eyes. The best results are obtained by the features sets containing the points belonging to the mouth and the set of nose and forehead coordinates.

We can observe from table 5.3 that most of these results are obtained through the classification of the standard deviation of the signal and for long windows.

Variable	$m(CR^+)$	mode	window
R mouth corner x	0.2785	<i>std</i> ( $\Delta\Delta$ )	41
R mouth corner y	0.3192	<i>std</i> ( <i>w</i> )	46
L mouth corner x	0.2465	<i>mean</i> ( $\Delta\Delta$ )	41
L mouth corner y	0.3200	<i>std</i> ( $\Delta\Delta$ )	36
nose x	0.2428	<i>std</i> ( <i>w</i> )	41
nose y	0.2915	<i>std</i> ( <i>w</i> )	26
right eye x	0.2315	<i>std</i> ( <i>w</i> )	31
right eye y	0.2717	<i>std</i> ( $\Delta$ )	36
left eye x	0.2311	<i>std</i> ( $\Delta$ )	31
left eye y	0.2798	$\Delta\Delta$	46
forehead x	0.2537	<i>std</i> ( <i>w</i> )	26
forehead y	0.2878	<i>std</i> ( <i>w</i> )	36
chin x	0.2608	<i>std</i> ( <i>w</i> )	26
chin y	0.2646	<i>std</i> ( <i>w</i> )	41
ext. R eyebrow x	0.2625	<i>std</i> ( <i>w</i> )	46
ext. R eyebrow y	0.3129	$\Delta\Delta$	41
int. R eyebrow x	0.2480	<i>std</i> ( <i>w</i> )	46
int. R eyebrow y	0.3086	<i>std</i> ( <i>w</i> )	46
int. L eyebrow x	0.2373	<i>std</i> ( $\Delta$ )	36
int. L eyebrow y	0.2874	<i>std</i> ( <i>w</i> )	26
ext. L eyebrow x	0.2684	<i>std</i> ( <i>w</i> )	46
ext. L eyebrow y	0.3189	<i>std</i> ( <i>w</i> )	36
upper lip x	0.2559	<i>std</i> ( $\Delta$ )	26
upper lip y	0.2611	<i>mean</i> ( $\Delta\Delta$ )	31
mouth region	0.3419	<i>std</i> ( <i>w</i> )	46
eyes region	0.2971	<i>std</i> ( <i>w</i> )	41
nose	0.3133	<i>std</i> ( <i>w</i> )	41
nose and forehead	0.3712	<i>std</i> ( <i>w</i> )	46

Table 5.3:  $m(CR^+)$  for the coordinates features

Then, we analyze the results showed in table 5.4. This table reports, for each one of the 6 emotions, the  $CR^+$  score for the best mode<sup>6</sup>. First, we notice that fear (FEA) and sadness (SAD) are more easily recognized than the other emotions. Secondly, we observe

<sup>6</sup>Please note that the score of one emotion for one mode is the result of the averaging of 5 different confusion matrices.

Variable	ANG	DIS	FEA	HAP	SAD	SUR
R mouth corner x	0.45	0.65	<b>0.81</b>	0.57	<b>0.80</b>	0.52
R mouth corner y	0.59	0.54	0.84	0.86	0.89	0.41
L mouth corner x	0.58	<b>0.75</b>	0.64	0.54	<b>0.67</b>	0.50
L mouth corner y	0.56	0.57	<b>0.77</b>	0.60	<b>0.78</b>	0.52
nose x	0.55	<b>0.80</b>	0.85	0.46	0.61	<b>0.70</b>
nose y	0.57	0.45	0.84	0.63	0.86	0.54
right eye x	0.59	<b>0.81</b>	0.90	<b>0.67</b>	0.64	<b>0.78</b>
right eye y	0.55	0.52	<b>0.80</b>	0.52	0.52	0.52
left eye x	<b>0.83</b>	<b>0.77</b>	0.85	0.44	0.61	<b>0.69</b>
left eye y	0.54	0.43	0.87	<b>0.66</b>	<b>0.82</b>	0.43
forehead x	0.60	<b>0.74</b>	0.85	0.61	0.46	<b>0.77</b>
forehead y	0.60	<b>0.68</b>	<b>0.66</b>	0.49	0.84	0.47
chin x	0.43	0.60	<b>0.78</b>	0.54	<b>0.70</b>	0.56
chin y	0.54	0.52	0.63	<b>0.71</b>	0.85	0.50
ext. R eyebrow x	0.64	0.60	0.84	0.58	0.47	0.58
ext. R eyebrow y	0.59	0.46	<b>0.77</b>	0.56	0.84	0.35
int. R eyebrow x	0.60	<b>0.74</b>	0.87	0.51	0.52	0.60
int. R eyebrow y	<b>0.73</b>	0.47	0.84	0.57	0.95	0.42
int. L eyebrow x	0.52	0.55	0.88	0.50	<b>0.70</b>	0.57
int. L eyebrow y	0.64	0.51	<b>0.80</b>	0.62	0.96	0.40
ext. L eyebrow x	0.58	0.40	0.93	0.52	0.93	0.58
ext. L eyebrow y	0.56	0.47	<b>0.69</b>	0.57	0.93	0.45
upper lip x	0.96	0.96	0.94	0.66	0.92	<b>0.82</b>
upper lip y	<b>0.81</b>	<b>0.73</b>	0.99	0.60	0.99	<b>0.78</b>
mouth region	0.54	0.49	0.64	<b>0.74</b>	0.89	0.40
eyes region	<b>0.72</b>	<b>0.79</b>	0.89	0.52	0.85	0.45
nose	0.58	0.49	0.85	0.59	<b>0.77</b>	0.40
nose and forehead	0.57	0.51	0.88	0.59	0.83	0.44

Table 5.4:  $CR^+$  for the coordinate features

that for each emotion there is at least one feature which result in a recognition rate bigger than 80%, the worst being surprise (SUR) with a recognition rate of 82% given by the x coordinate of the upper lip.

Anger (ANG) seems better recognized based on the coordinates of the eyes and eyebrows than for the coordinates of the mouth region; this result is confirmed for the coordinates sets with the point belonging to the eyes 72% and the best other set performing only 54%. The same behavior is found for disgust (DIS) and fear.

Happiness is better recognized using the points belonging to the mouth than for the coordinates of the eye region; this result is confirmed from the sets of features.

Finally, for both sadness and surprise we are not able to say that a particular region performs better than the others. Every region works well allowing to recognize sadness. The few coordinates which works better with surprise are more or less equally spread. This result is confirmed from the scores obtained with the coordinates sets.

**Distances Features.** In the former section we have analyzed the results from the coordinate feature set; in this section we describe how the distances, that we defined in section 5.3.6, perform with the aid of the same graphs and figures.

Figure 5.17 reports the  $average(CR^+)$  scores for the distances features. In average these features score 60% and therefore about 5% less than the coordinates features (see figure 5.21). We notice that the three features best performing are the ones relative to the distance between the chin and the mouth (i.e. the “openness” of the mouth and/or the smiling), the distance between the right eye to the relative eyebrow and the mean x displacement of the head.

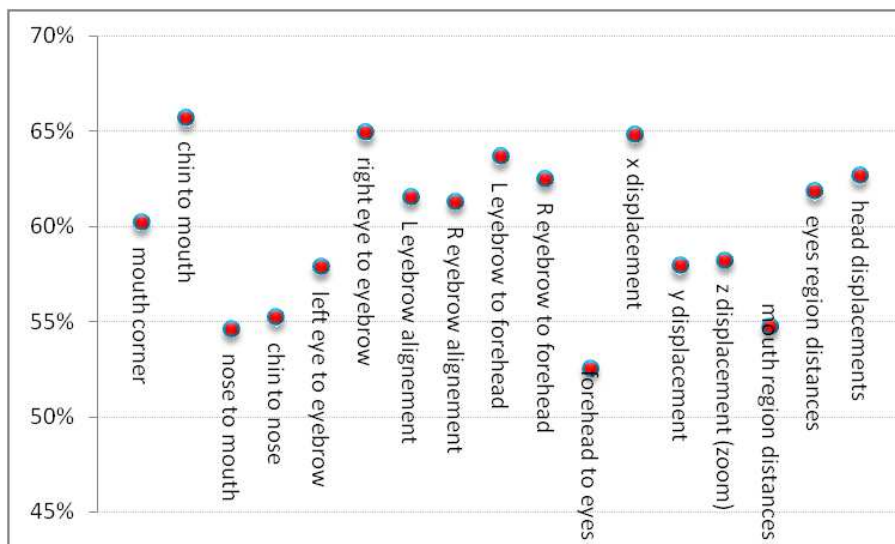


Figure 5.17: average  $CR^+$  for the distance features

The worst feature is represented by the distance between eyes and forehead; this result was, at this point, expected. This distance is, indeed, computed on the  $y$  coordinate, which we know, from the previous section, not to carry significant emotional information for eyes and forehead.

In average the features relative to the eyes and eyebrows seem to perform better than the ones belonging to the mouth region. The same behavior is confirmed from the sets of features: here eye distances and head displacement outperform the mouth region by more than 7%. This result confirms our expectations; speech production influencing negatively the capability of the system to recognize emotions from the video signal.

	t	$\Delta$	$\Delta\Delta$
raw	11	8	16
mean	12	10	12
stdev	7	10	16

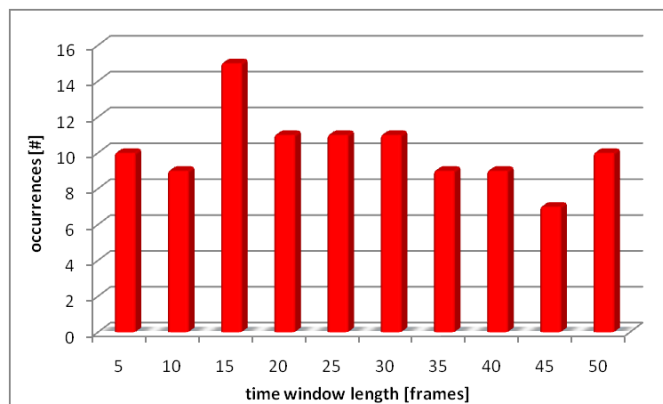


Figure 5.18: Modes and window lengths for the distance features

In figure 5.18 (and table), we can observe the statistics over the modes and window lengths which gave these results. We notice that for the distances features the favorite mode is the second derivative of the signal ( $\Delta\Delta$ ) gathering more than the 43% of the examples. In this case the favorite window lengths are concentrated between 15 and 30 frames.

In table 5.5 we report the  $m(CR^+)$  score of the single mode which in average return

Variable	$m(CR^+)$	mode	window
mouth corner	0.3298	$std(\Delta\Delta)$	46
chin to mouth	0.2807	$std(\Delta)$	46
nose to mouth	0.2665	$std(\Delta)$	46
chin to nose	0.2922	$std(w)$	46
left eye-eyebrow	0.2506	$std(\Delta)$	11
right eye-eyebrow	0.2453	$std(\Delta)$	31
L eyebrow align.	0.2545	$\Delta$	46
R eyebrow align.	0.2405	$\Delta$	41
L eyebrow-forehead	0.2470	$std(\Delta\Delta)$	21
R eyebrow-forehead	0.2783	$std(\Delta\Delta)$	36
forehead to eyes	0.2348	$\Delta$	46
x displacement	0.2513	$std(w)$	46
y displacement	0.3066	$\Delta\Delta$	46
z displacement	0.2973	$\Delta$	31
mouth region	0.3183	$std(\Delta\Delta)$	41
eyes region	0.2773	$mean(\Delta\Delta)$	46
head displacements	0.3230	$std(w)$	36

Table 5.5:  $m(CR^+)$  for the distances features

the best result. In this case it appears that the single distances belonging to the mouth region and the head displacement information carry more information than the distances of the eye region. This result was un-expected. Nevertheless, it is possible that taking into account the derivatives on long windows is enough to filter out the noise-movements due to speech production.

Variable	ANG	DIS	FEA	HAP	SAD	SUR
mouth corner	0.61	0.55	0.64	0.51	0.89	0.41
chin to mouth	0.58	0.52	<b>0.68</b>	<b>0.70</b>	0.89	0.56
nose to mouth	0.39	0.63	0.44	0.57	0.85	0.40
chin to nose	0.43	0.45	<b>0.69</b>	0.59	<b>0.76</b>	0.39
left eye-eyebrow	0.57	0.53	<b>0.74</b>	0.47	<b>0.77</b>	0.40
right eye-eyebrow	0.52	<b>0.77</b>	0.83	0.52	0.57	<b>0.68</b>
left eyebrow align.	0.57	0.55	<b>0.79</b>	0.54	0.52	<b>0.71</b>
right eyebrow align.	<b>0.66</b>	0.59	0.84	0.51	<b>0.66</b>	0.42
L eyebrow-forehead	0.50	0.57	0.83	0.53	0.89	0.50
R eyebrow-forehead	0.56	0.52	<b>0.79</b>	0.50	<b>0.69</b>	<b>0.68</b>
forehead to eyes	0.49	0.53	<b>0.74</b>	0.43	0.44	0.53
x displacement	0.52	0.64	0.88	<b>0.68</b>	0.46	<b>0.71</b>
y displacement	0.49	0.42	0.85	0.37	<b>0.82</b>	0.53
z displacement	0.61	0.52	<b>0.76</b>	0.46	<b>0.79</b>	0.35
mouth region	0.49	0.49	0.61	0.49	0.87	0.34
eyes region	0.51	0.94	<b>0.77</b>	0.50	0.50	0.49
head displacements	0.60	<b>0.66</b>	0.86	0.50	<b>0.75</b>	0.37

Table 5.6:  $CR^+$  for the distance features

In table 5.6, we can observe the best  $CR^+$  obtained by each emotion for the distance features. Once more, we observe that two emotions are better recognized than others: these are, as it was found for the coordinate features, fear and sadness.

The emotion which is recognized with the least accuracy is anger, with a maximum recognition of 66% for the feature relative to the alignment of the right eyebrow. We can also notice that the same emotion is better recognized for the distances relative to the eye region and for the head displacements than for the features relative to the mouth. The same behavior is found for the disgust, fear, and surprise.

The emotion happiness is better recognized thanks to the movements of the mouth region similarly to what it was found for the coordinate features. This could have been



expected: indeed, the main movement related to the happiness facial expression are related to the mouth (smiling); this is especially true in the case of “fake” smile Darwin [1872], Duchenne De Boulogne [1862], Kaiser and Wehrle [2001]. We observe this same phenomenon for the emotion sadness.

Finally, for the sets of features similar behaviors are found: the eye region works the best for disgust and surprise, the head displacements for anger and fear, and the mouth region performs the best for sadness.

We have now discussed the results from the video modality. For most emotions the assumptions presented in section 5.4.8 have been confirmed. Features relative to the eye regions work better than the ones belonging to the mouth region but this is not true for all the cases (i.e. sadness and happiness achieve better scores with the features of the mouth regions).

Unexpectedly, we observe that coordinates perform generally better than distances (see figure 5.21); there are two possible reasons for that: the first is that our definition of the distances is sub-optimal, the second is that concentrating information into distances is, per se, sub-optimal. This was particularly unexpected because using the two feature sets as wholes in our previous tests, the distance feature set always returned better results than the coordinate one. It is possible that single coordinates maintain more emotional information, while distances, representing more “compressed” information, lose this capabilities when split in small sets.

**Audio Features** In this section, we analyze the results of the features relative to the audio modality.

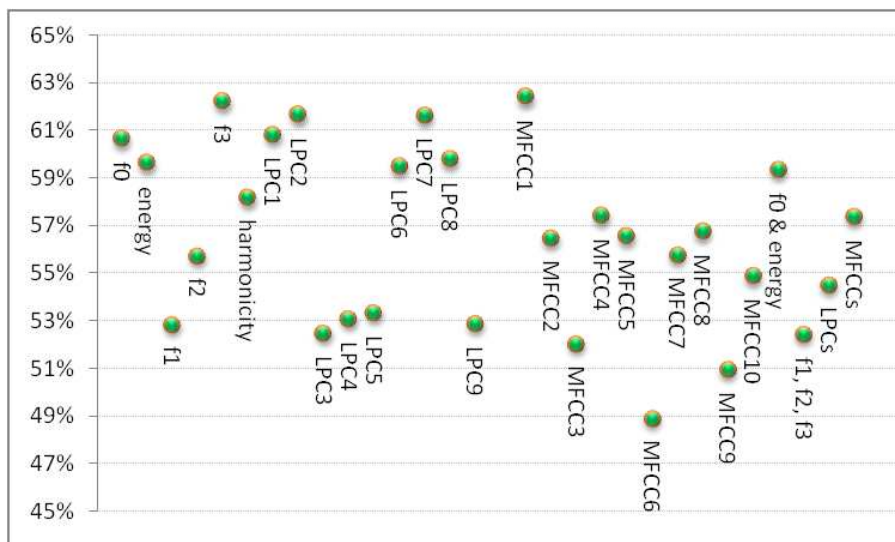


Figure 5.19: average  $CR^+$  for the audio features

In figure 5.19, we report the  $average(CR^+)$  scores obtained by the different audio features that we have listed in section 5.3.7. In average this modality performs with 56.5% accuracy, roughly 9% less than the coordinate features and 4% less than the distance features (see figure 5.21). Nevertheless, it is important to notice that, at the moment, the audio emotion appraisal is returned also for frames which do not present audio (silent pieces of the video shots); simply filtering out these estimation is likely to improve the



scores.

Generally we have seen in the previous section 5.4.3 that audio performs slightly better than video. This was true for the whole set of audio features. We have to conclude that single video features better represent emotions than single audio features. The complementarity of the information carried by the different audio features is enough to overcome this issue while a relevant part of the information of the video features is still common among the features.

In this case, the feature which performs better is represented by the first *MFCC* with 62.4% of  $average(CR^+)$ , the worst one is the 6<sup>th</sup> *MFCC*<sup>7</sup> with 49%. Other variables which are meaningful for emotion recognition are *pitch*, *energy*, 3<sup>rd</sup> *formant*, the first two and the 7<sup>th</sup> *LPCs*.

Regarding the sets of features, the set of *pitch* and *energy* is the one providing the highest  $average(CR^+)$  score followed by the set of the *MFCCs*.

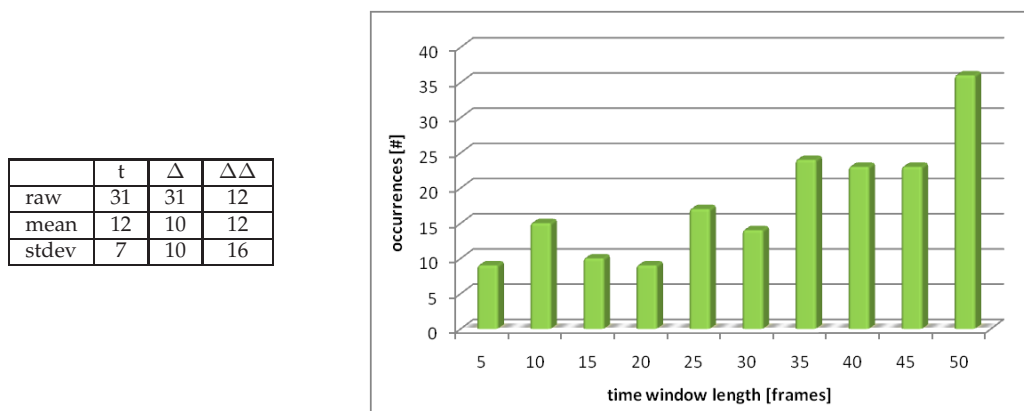


Figure 5.20: Modes and window lengths for the audio features

These results are obtained with the use of raw data (52% of the cases) while the  $\Delta\Delta$  values are disadvantaged as shown in figure 5.20 (and table). On top of that, we also notice that 20% of the best trainings are obtained while using the longest 10% of the available window lengths. This may be due to the system need to somehow filter out samples of the audio signal when the subject were not articulating sounds.

In table 5.7 we report the  $m(CR^+)$  score of the single mode which in average return the best result. We notice that for the audio features sets of features perform better than single features. Longer windows are preferred to shorter ones. The derivative features  $\Delta$  and  $\Delta\Delta$  are preferred to the *raw* time analysis. The best single feature is the second *LPC* coefficient. The best time-based (i.e. non spectral-based) feature is the energy of the signal.

When we observe the results of the individual emotions for the different audio feature in table 5.8 we can observe that sadness is, once more, the best recognized emotion, followed by anger, surprise, and happiness. If we do not consider the 10<sup>th</sup> *LPC* (which is as we said always identical to 0 for all samples) then *LPCs* works better than *MFCC* when these features are taken one by one, but worse when taken all together in a set of features.

Anger is well recognized thanks to the energy (high energy in this case), the first *LPC*,

<sup>7</sup>The 10<sup>th</sup> *LPC* performs worse with 16% (the same as random) but that was expected since the value of this feature is identical to 0 for every sample of the base.

Variable	$m(CR^+)$	mode	window
f0 - pitch	0.2435	$mean(\Delta\Delta)$	41
energy	0.3093	$\Delta$	41
f1	0.2734	$\Delta$	41
f2	0.2777	$\Delta$	46
f3	0.2644	$\Delta$	46
harmonicity	0.2435	$\Delta$	46
$LPC_1$	0.2983	$\Delta$	46
$LPC_2$	0.3397	$std(w)$	46
$LPC_3$	0.2720	$std(\Delta)$	41
$LPC_4$	0.2560	$std(\Delta\Delta)$	41
$LPC_5$	0.2548	$std(\Delta)$	11
$LPC_6$	0.2805	$std(w)$	41
$LPC_7$	0.2857	$std(\Delta\Delta)$	26
$LPC_8$	0.2550	$std(\Delta\Delta)$	26
$LPC_9$	0.2896	$std(\Delta\Delta)$	36
$LPC_{10}$	0.2365	$w$	46
$MFCC_1$	0.2605	$mean(\Delta)$	41
$MFCC_2$	0.2760	$mean(\Delta)$	31
$MFCC_3$	0.2279	$std(\Delta)$	36
$MFCC_4$	0.2591	$\Delta$	31
$MFCC_5$	0.2756	$mean(\Delta\Delta)$	46
$MFCC_6$	0.2571	$std(\Delta)$	46
$MFCC_7$	0.2541	$\Delta$	46
$MFCC_8$	0.2680	$std(w)$	46
$MFCC_9$	0.2595	$std(\Delta\Delta)$	31
$MFCC_{10}$	0.2720	$std(w)$	46
f0 & energy	0.3065	$\Delta$	46
f1, f2, f3	0.3007	$mean(\Delta)$	21
$LPCs$	0.3503	$std(\Delta\Delta)$	46
$MFCCs$	0.3374	$\Delta$	46

Table 5.7:  $m(CR^+)$  for the audio features

and most of the  $MFCCs$ ; the best set of features is, for this emotion, the one formed by *pitch* and *energy*.

Disgust demonstrates to be an emotion which is particularly hard to recognize from audio only; the best features result to be the 1<sup>st</sup> and the 9<sup>th</sup>  $MFCC$  and the *harmonicity* value.

Fear is well recognized using the *pitch* value (high pitch in this case) with 76%; the 2<sup>nd</sup> and the 8<sup>th</sup>  $LPCs$  works quite well too for the single features while, for the sets of features, the one composed of *pitch* and *energy* is the one returning the best result.

Regarding the emotion of happiness we note that the best results are obtained while using the 5<sup>th</sup>  $MFCC$ , the 3<sup>rd</sup> *formant* and the (higher) *harmonicity* value. *Pitch* also gives good results as a single variable. The set of *pitch* and *energy* and the one composed by the different  $MFCCs$  result in the best scores for the sets of features.

Sadness is easily recognized using most features, with the sole exception of the 3<sup>rd</sup>, 5<sup>th</sup>, 6<sup>th</sup>, and 9<sup>th</sup>  $MFCCs$ . Sets of features also performs very well,  $LPCs$  being the best with 100% of  $CR^+$  and  $MFCCs$  being the worst with 80%.

Finally, surprise is well recognized only with the use of the 7<sup>th</sup>, 6<sup>th</sup>, and 4<sup>th</sup>  $LPCs$  and with the 1<sup>st</sup>  $MFCC$ . *Formants* represent the set of features which does best contribute to recognize this emotion.

**Summary of the Results** When summarizing the results we observe that:

- emotions are not always recognized with the same level of accuracy: in particular,

Variable	ANG	DIS	FEA	HAP	SAD	SUR
f0 - pitch	0.50	0.38	<b>0.76</b>	0.64	<b>0.76</b>	0.59
energy	<b>0.86</b>	0.35	0.34	0.47	0.99	0.57
f1	<b>0.66</b>	0.37	0.46	0.37	0.90	0.41
f2	0.57	0.46	0.43	0.50	0.93	0.45
f3	0.46	0.51	0.55	<b>0.67</b>	0.93	0.61
harmonicity	0.54	0.53	0.55	<b>0.66</b>	<b>0.69</b>	0.51
LPC <sub>1</sub>	<b>0.82</b>	0.46	0.37	0.59	0.95	0.47
LPC <sub>2</sub>	0.60	0.33	0.59	0.53	1.00	0.64
LPC <sub>3</sub>	0.40	0.33	0.44	0.44	1.00	0.53
LPC <sub>4</sub>	0.42	0.25	0.40	0.42	1.00	<b>0.69</b>
LPC <sub>5</sub>	0.47	0.34	0.43	0.41	1.00	0.56
LPC <sub>6</sub>	0.49	0.43	0.52	0.39	1.00	<b>0.73</b>
LPC <sub>7</sub>	0.43	0.46	0.52	0.45	1.00	<b>0.84</b>
LPC <sub>8</sub>	0.57	0.48	0.61	0.32	1.00	0.60
LPC <sub>9</sub>	0.48	0.42	0.39	0.37	0.97	0.55
LPC <sub>10</sub>	1.00	-	-	-	-	-
MFCC <sub>1</sub>	0.50	0.62	0.53	0.52	0.91	<b>0.67</b>
MFCC <sub>2</sub>	<b>0.67</b>	0.36	0.36	0.58	0.91	0.50
MFCC <sub>3</sub>	<b>0.67</b>	0.47	0.46	0.51	0.50	0.51
MFCC <sub>4</sub>	<b>0.71</b>	0.47	0.38	0.55	0.88	0.44
MFCC <sub>5</sub>	<b>0.68</b>	0.47	0.55	<b>0.70</b>	0.54	0.46
MFCC <sub>6</sub>	0.63	0.38	0.47	0.46	0.59	0.40
MFCC <sub>7</sub>	<b>0.69</b>	0.51	0.44	0.60	0.64	0.46
MFCC <sub>8</sub>	0.62	0.48	0.35	0.58	<b>0.78</b>	0.59
MFCC <sub>9</sub>	<b>0.69</b>	0.54	0.34	0.47	0.58	0.44
MFCC <sub>10</sub>	0.58	0.48	0.38	0.65	<b>0.69</b>	0.52
f0 & energy	0.61	0.32	0.63	0.61	0.92	0.48
f1. f2. f3	0.48	0.30	0.43	0.44	0.86	0.62
LPC <sub>s</sub>	0.49	0.39	0.41	0.42	1.00	0.56
MFCC <sub>s</sub>	0.60	0.36	0.49	0.64	<b>0.80</b>	0.55

Table 5.8:  $CR^+$  for the audio features

some emotion are generally better recognized than others (see figure 5.22).

- We could not find a specific analysis triple of *window\_size* ([1,50]), *feature\_set* ( $t$ ,  $\Delta$ , or  $\Delta\Delta$ ), and *mode* (*raw*, *mean*, or *stdev*) working well for all the emotions. Depending on the particular emotion and feature different modes should be employed.
- We notice that, generally, longer time windows provide slightly better results. Indeed, the results show that in 40% or the cases the best result is obtained with window longer than 40 frames and than about 15% of these results are obtained with window longer than 47 frames.
- increasing the number of emotionally relevant features does not seem to always improve the result. In the experiment reported here, sets of features often perform worse than the best of the included features<sup>8</sup>.
- with the current settings coordinate features work in average better than either distances and audio features (see figure 5.21).
- anger is best recognized using the  $x$  coordinates of the eyes and of the upper lip, the information about the alignment of the eyebrows; for the audio we will use *energy*

<sup>8</sup>We have also done few test doubling the number of neurons and epochs of the NN to trying overcoming the added complexity of the sets of multiple features trainings but did not result in the significant changes that we were expecting

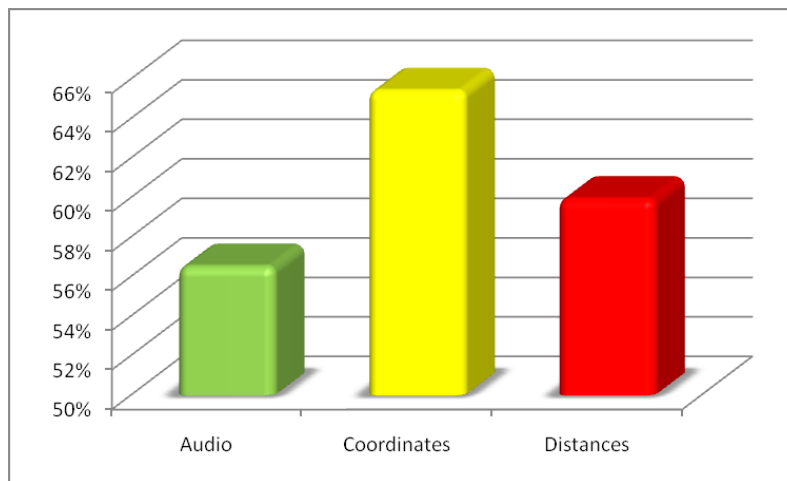


Figure 5.21: Average  $CR^+$  for the different modalities

and the first  $LPC$ .

- ***disgust*** is recognized with the  $x$  coordinates of the eyes, the nose, and the upper lip and the information of the distances of the eye region while using audio features other than the first  $MFCC$  should be avoided.
- ***fear*** is mainly recognized using video features; indeed, only the *pitch* seem to return good results for the audio features.
- ***happiness*** is characterized by the coordinates of the mouth, and in particular the  $y$  coordinates of the mouth corners. The distance chin to mouth may be used too; for the audio features we will mostly rely on the 3<sup>rd</sup> *formant*, the *harmonicity*, and the 5<sup>th</sup>  $MFCC$ .
- ***sadness*** is well recognized using most features and in particular audio seem to better discriminate between sadness and all the other emotions.
- ***surprise*** is best recognized by the use of the  $x$  coordinates of eyes, nose, and upper lip, the mean face  $x$  *displacement*, and the right eyebrow alignment. For the audio features we will use the 7<sup>th</sup>, 6<sup>th</sup>, and 4<sup>th</sup>  $LPCs$  and the 1<sup>st</sup>  $MFCC$

### 5.4.9 Multimodal Fusion

The objective of this section is to present the results relating the performances of some of the different kind of possible fusion systems. As we have seen in section 5.3.10 there are different kind of fusion.

In figure 5.23 we report the average and the standard deviation recognition rates for four different systems. The four system all exploit support vector machines (SVM) and statistical analysis on windows of 25 samples (i.e. one second). We implemented for the “feature fusion” a simple concatenation of the audio and video feature vectors. “Decision fusion” is obtained by averaging the results from the monomodal audio and video systems.

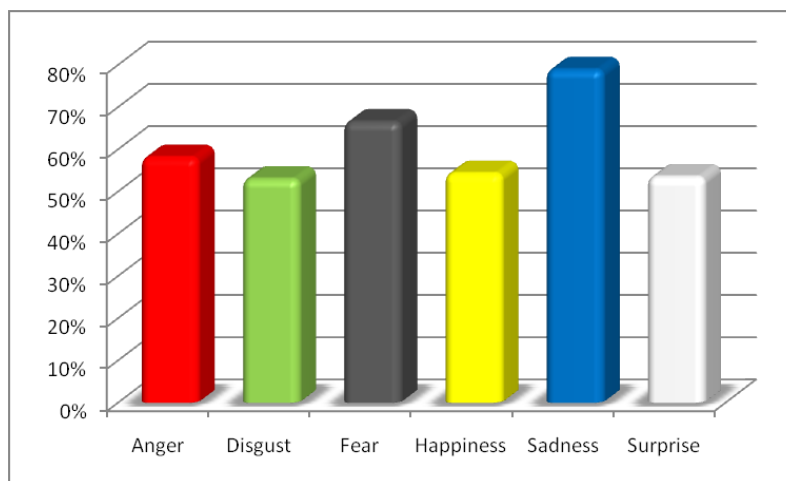


Figure 5.22: Average  $CR^+$  for the different emotions

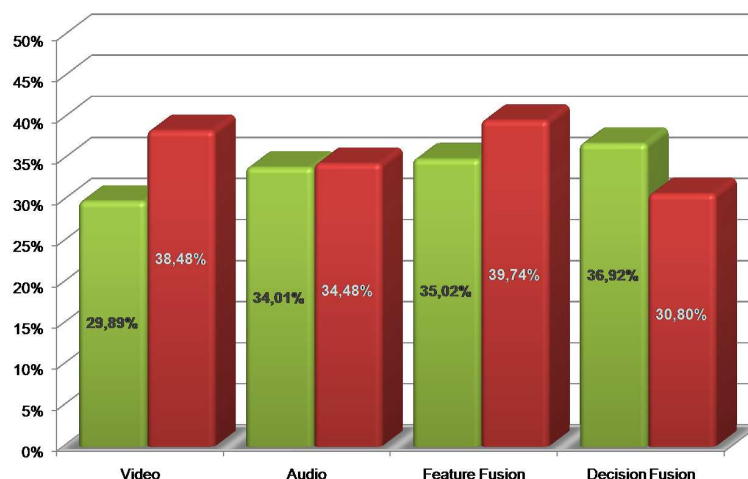


Figure 5.23:  $m(CR^+)$  and  $wstd(CR^+)$  results comparison for basic monomodal and multimodal approaches

As it can be seen from figure 5.23 both fusion techniques works better than unimodal system obtaining higher average recognition rate. Nevertheless, we can observe that the system based feature level fusion obtain this result to the detriment of the weighted standard deviation. On the other hand we note that the system performing decision level fusion outperform, once more, the two monomodal systems and the simple feature fusion system with both average recognition rate and weighted standard deviation.

Decision level fusion is an important step of the classification task. It improves recognition reliability by taking into account the complementarity between classifiers. Several schemes have been proposed in the literature according to the type of information provided by each classifier as well as their training and adaptation abilities. A state of the art is proposed in Benmokhtar and Huet [2006].

ARAVAR performs fusion between estimations resulting from different classifiers or

modalities. The output of such a module boosts the performances of the system. Since with NN and SVM the classification step is computationally cheap, we are allowed to use multiple classifiers at the same time without impacting too much on the performances.

Thanks to the simple use of classifier fusion strategies result can be boosted of around 19% with respect of the baseline that in this example was the simple audio SVM classification<sup>9</sup>.

In table 5.9 and in figure 5.24 We compare multiple decision level fusion strategies that have been tested and evaluated including Max, Vote, Mean, Bayesian combination and the NNET approach<sup>10</sup> (Paleari et al. [2009a]).

	Anger	Disgust	Fear	Happiness	Sadness	Surprise	$m(CR^+)$	$wstd(CR^+)$
Video NN	0.420	0.366	0.131	0.549	0.482	0.204	0.321	0.504
Audio NN	0.547	0.320	0.151	0.496	0.576	0.169	0.354	0.536
Video SVM	0.342	0.342	0.193	0.592	0.426	0.244	0.320	0.443
Audio SVM	0.627	0.220	0.131	0.576	0.522	0.162	0.361	0.624
Max	0.612	0.378	0.120	0.619	0.586	0.185	0.384	0.584
Vote	<b>0.666</b>	0.422	0.142	0.622	0.495	0.161	0.391	0.573
Mean	0.635	0.406	0.150	0.721	0.600	0.206	0.415	0.572
Bayesian	0.655	<b>0.440</b>	0.159	<b>0.743</b>	0.576	0.235	<b>0.430</b>	0.542
NNET	0.542	0.388	<b>0.224</b>	0.633	<b>0.619</b>	<b>0.340</b>	0.428	<b>0.386</b>

Table 5.9: Decision level multimodal classifier with different fusing criteria

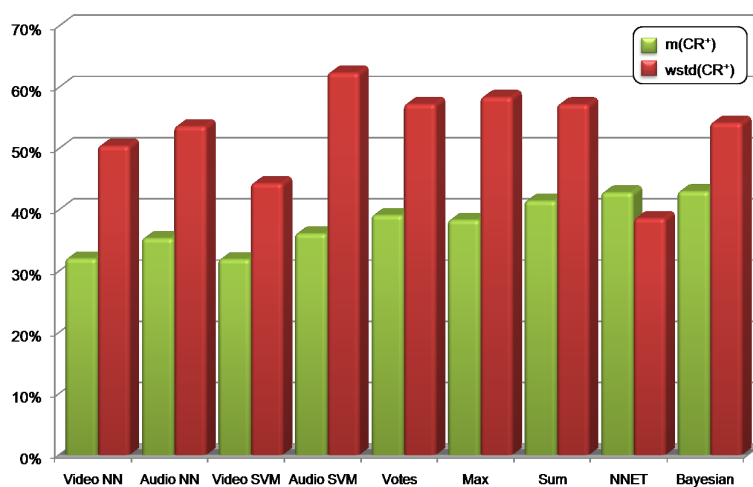


Figure 5.24:  $m(CR^+)$  and  $wstd(CR^+)$  results comparison for different monomodal and multimodal approaches

As the results report we notice that Bayesian combination and NNET approaches outperform other systems in term of  $m(CR^+)$ . Both systems improve the average recognition score of a relative 19%.

When comparing those last two we can observe that the Bayesian approach reach a slightly higher average score. On the other hand, the NNET approach allows to minimize

<sup>9</sup>We have chosen as baseline the best classifier between audio SVM, audio NN, video SVM, video NN without using any thresholding or averaging technique

<sup>10</sup>The system employ the statistical representation of data coupled with a second order polynomial estimation on a sliding window of 25 frames, one second.

the weighted standard deviation by boosting the  $CR^+$  of the two emotions which are usually less recognized (i.e. fear and surprise).

Nevertheless, NNET bring about two main drawbacks. Generally when applying multimodal fusion we are trading improved performances with computational power. When dealing with NNET we also need more data to support the training of this new classifier.

#### 5.4.10 Detectors or Classifier

In the former sections we have always used machine learning algorithms to classify the six different emotions in one step. An alternative technique consists of training one different classifier for each emotion obtaining, therefore, six different emotional-concept detectors. Doing this has two main advantages:

1. training a different classifier for each emotion reduces the complexity of the single classifier;
2. training a different classifier for each emotion allow to use different machine learning technique, different features, or different processing. Indeed, different emotions may be better recognized by different machine learning approaches.

In this section we report a simple result showing how much improvement is brought by the first advantage. In figure 5.25 we report the average recognition rate and the weighted standard deviation result for NN trained with video data.

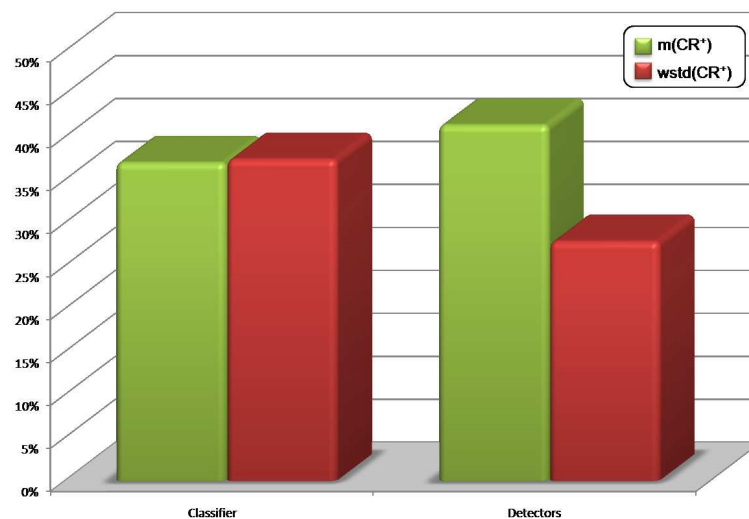


Figure 5.25:  $m(CR^+)$  and  $wstd(CR^+)$  results comparison for an approach based on one classifier and one based on 6 detectors

As it can be seen from figure 5.25 six detectors outperform one single classifier both increasing the average recognition score and reducing the weighted standard deviation. Using this approach simply means trading the improved performances of the system with increased computational load and slightly improved complexity.



### 5.4.11 Machine Learning Approaches

In the former sections we have detailed results mainly obtained using NN and SVM. One might wonder how the different classifiers performs relatively to each other for the specific task of multimodal emotion recognition.

It is very hard to compare different machine learning techniques because of the high number of different parameters that play a role in the training. For this reason, rather than give some numeric results, we will try to resume the conclusion we have made during our studies on this matter.

Since the emotional expressions are a dynamic phenomena and HMM automatically take into account the dynamic component of the signals we expect HMM to outperform the other classifiers in this particular scenario. In particular, the comparison between GMM and HMM should be very simple as we modeled the HMM as a simple left-to-right, as opposed to ergodic, (i.e. the states are only connected to themselves, to the next, and to the second-next but not on previous states as opposed to a fully connected network of states) sequence of GMM states (as validated by Wagner et al. [2007]). SVM are known for their capability to handle high number of input features; we expect SVM to perform rather well in multimodal feature-fusion scenarios and whenever the number of features increases.

Despite our expectations, the results of our tests show that in most scenarios different classifiers influence only partially on the quality of results. All classifiers practically give similar results both in term of  $m(CR^+)$  and  $wstd(CR^+)$ . In most cases the results of our analysis showed to depend more on random seed of the particular training than on the employed classifier. Different training on the same data might make vary the results of as much as 6%  $m(CR^+)$  while the absolute difference between two different classifiers is, with the employed database, lower.

In particular, contrary to our expectations GMM and HMM return basically the same results in all cases. This may due to the relatively small size of the eINTERFACE database and in particular to the limited length of the single video files. Indeed training HMMs involves using a bigger quantity of data.

To conclude the debate about the quality of different classifiers is left it to other works more specifically centered on this particular task. Given the lower computational complexity of neural network training this specific classifier has been chosen for most of our tests.

### 5.4.12 Data Post-Treatment

The objective of this section is to present some basic post-processing techniques for boosting the results. The idea is that we have applied a certain number of classification engines to our feature vectors and we have the output values (one for each time frame and emotion). This section shows how we can process these output values to extract more precise emotional estimations.

**Low-Pass Filtering** Until now we have been presenting systems doing frame-by-frame classification: each system might as well work on a sliding time window, but it is nevertheless classified independently from the others. One first technique which could be used to improve the results takes advantage of the basic temporal correlation that there should be among output classification samples. Indeed it is not possible that the analyzed face

---

displays anger in one frame (and surrounding window) and happiness on the following frame just 1 25<sup>th</sup> of second after.

In figures 5.26 and 5.27 we show the results obtained by applying a simple low-pass filtering technique (mean-like filter) to the output probabilities of the video SVM system of section 5.3.10.

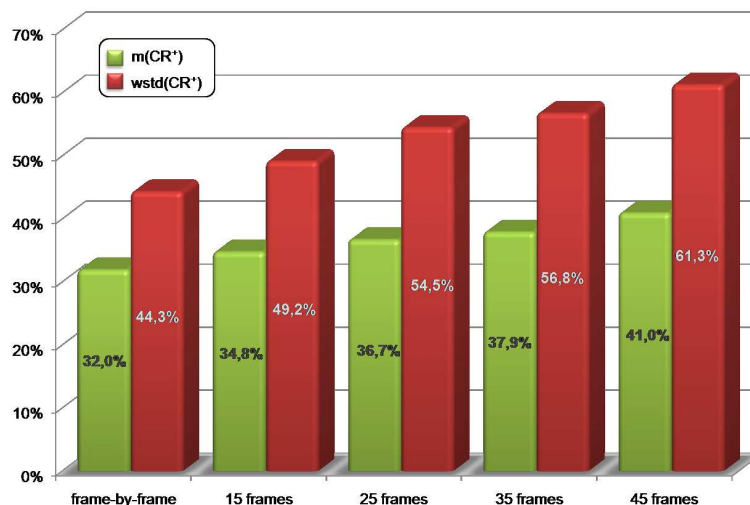


Figure 5.26:  $m(CR^+)$  and  $wstd(CR^+)$  results comparison for different low-pass filtering sizes

As it can be clearly see from figure 5.26 the average recognition rate augments almost linearly with the filter size. Unfortunately also the weighted standard deviation has the same behavior.

Indeed, as it can be seen in figure 5.27, the score of emotions that were well recognized augment and the score of fear, which was already controversial, worsen.

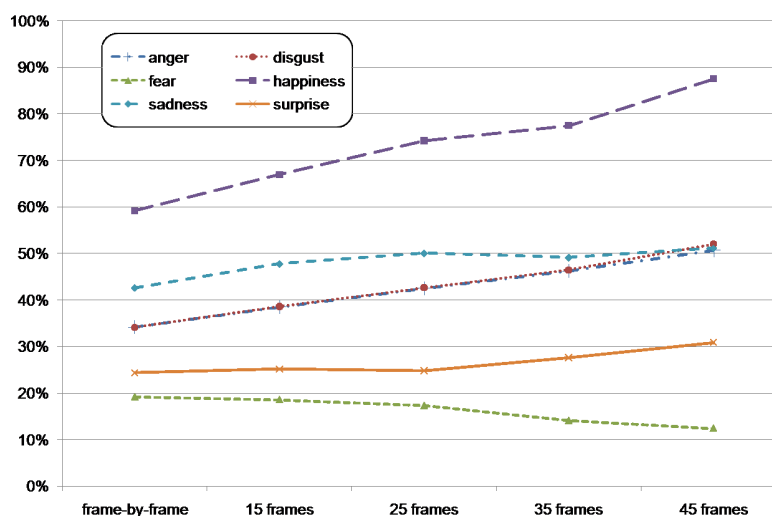


Figure 5.27:  $CR^+$  results comparison for different low-pass filtering sizes

If we apply this technique to more reliable results, such as the one outputted by the

multimodal approach based on NNET than the improvement is even more marked. Indeed no  $CR_{emotion}^+$  does worsen in this case and the weighted standard deviation remains more constant. Standard deviation remains more constant.

Experimentally, we have noticed that the borderline score for an emotion to get improved by this approach is of about 22%. Please note that this technique does not reduce the number of the estimations. This is, that with this technique we can still extract a different estimation per second.

One could wonder how far we could go on filtering the signals. In figure 5.28 we show the average recognition rate results obtained by training several NN with the distances data defined in section 5.3.6 while varying the length of the low-pass filter

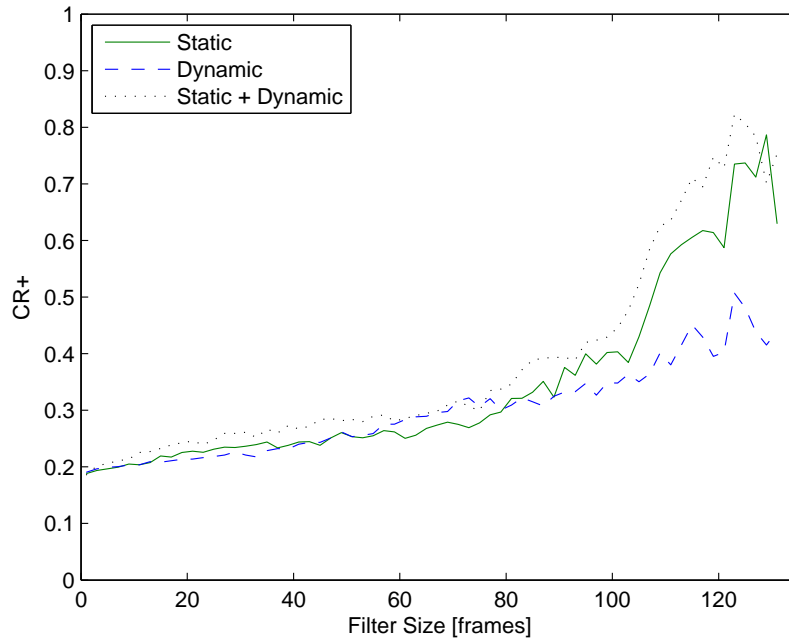


Figure 5.28:  $CR^+$  results comparison for different low-pass filtering sizes

We let the filter vary from 1 frame (i.e. no filtering) to 127 frames (i.e. more than 5 seconds of video). Values of filters bigger than this one have no sense since this is the maximum length of the files included into the eNTERFACE'05 database. We used first and second derivatives to represent the dynamics of the system.

It is interesting to notice that also dynamic features shall be partly filtered to improve the results. It is also interesting to notice that longer filters almost always improve the obtained average classification results. This may indicate that the signal are very noisy. From the other hand, in this experiment, we have been reducing the number of the trained and tested samples to only include samples from which all data belonged to the same video (i.e. when filtering with a 3 frames filter then 2 samples are lost at the borders); the reduced number of samples may have simplified the training and testing procedures while also reducing the influence of noise and the one of the first frames of the video in which the subject may show a neutral or misleading emotion (see section 5.3.2).

**Thresholding** Another simple solution to improve the results exploits the thresholding of the results. This is, if the best estimation value  $e(t, emo)$  for a specific emotion  $emo$

and time  $t$  is lower than a certain threshold  $tr$  than that is not considered reliable enough and no estimation is given for that certain time  $t$ . This technique reduces the number of output estimations.

In figure 5.29 we plot the average recognition rate, the weighted standard deviation, and the percentage of the total sampled which are classified when thresholding is applied to the video SVM system of section 5.3.10. As one can see from this graph, the average recognition rate improves when the threshold decreases.

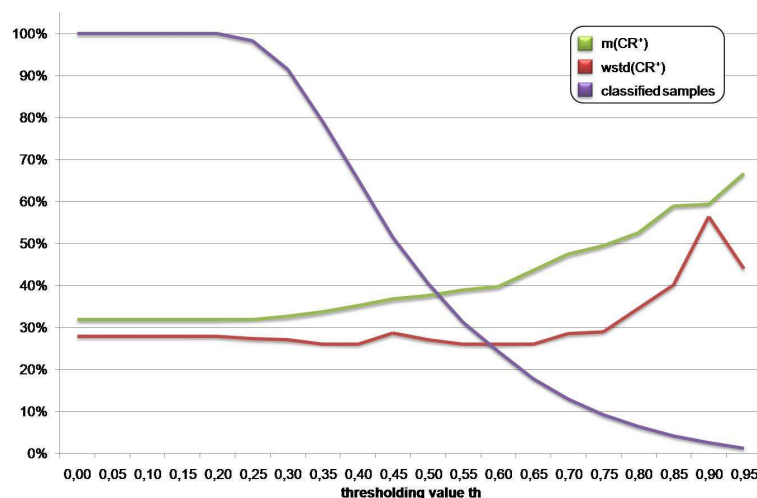


Figure 5.29:  $CR^+$ ,  $wstd(CR^+)$ , and `classified_samples` comparison for different thresholding values

Furthermore, one could notice that the improvement of the average recognition rate is coupled with a almost constant weighted standard deviation for thresholding values  $th < 0.75$ . After that threshold the first emotion disappears (i.e. no samples are classified for the emotion fear) because of the size of the database and the little reliability of that particular emotion and the standard deviation starts increasing faster than the average recognition score.

The drawback of the thresholding operation is linked to the decreased number of samples which are actually classified.

We can directly see the relation between the number of correctly classified samples and the total number of classified samples thanks to the precision and recall graph in figure 5.30.

In figure 5.30 we compare the performances of the multimodal systems based on the Bayesian and NNET approach described in section 5.3.10. In this case only the best *recall* percentage of the samples are kept which have the highest likelihood to be correct.

**K-best Approach** Yet another way to improve the results is to accept more than the first result as the classified one. It is obvious that selecting more than one class to classify one sample improve the chances to extract the right class. With a number of classes extracted variable and equal to  $k$  this problem is often represented as a graph of  $k - best$  results as in figure 5.31.

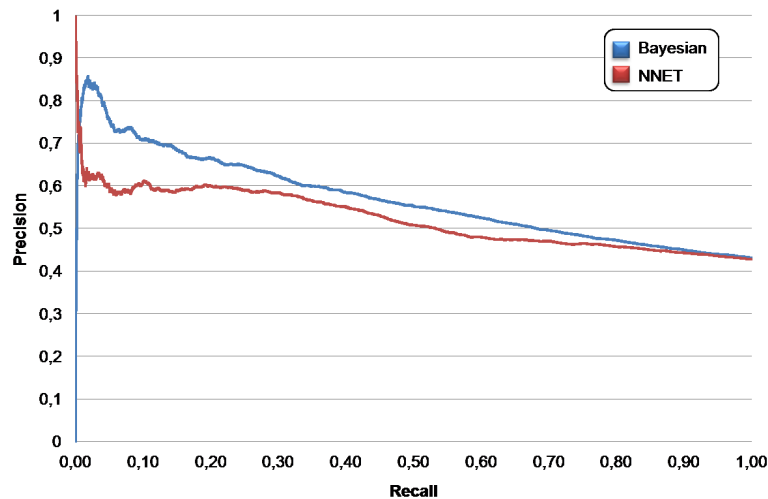


Figure 5.30: NNET and Bayesian: comparison of precision and recall

The more the graph second derivative is negative (i.e. the more the curve are convex toward the low part of the graph) the more we know that the system work. Indeed, it would be normal for all system better than random to improve linearly the average recognition score by linearly augmenting the number of accepted classes.

**Inverse Thresholding.** A soft way of accepting more than one classification would be the one which we call “**inverse thresholding**”. Using inverse thresholding means to classify a sample with all emotions whose classification score is above a certain threshold  $th^{-1}$ . In this case more than one estimation per sample can be found but this number is variable and dynamically adapts to needs of the moment.

In figure 5.32 we report the average recognition rate, the weighted standard deviation, and the percentage of the total sampled which are classified when inverse thresholding is applied to the video SVM system of section 5.3.10.

If one compares the result of this system to the ones of the k-best one might notice that the average recognition results are about the same for both the systems. The interest of this second technique is that one could select any desired inverse threshold  $th^{-1}$  while the k-best has to be chosen as an integer number.

It is practically possible to chose an inverse threshold value such that less than 100% are selected. Nevertheless, as it can be seen from the results in figure 5.32, this operation is to be discouraged as it reduces the average recognition rate while increasing the weighted standard deviation.

We have detailed, in this section, few techniques that can be performed a-posteriori to improve the result of the system. We have seen that with the right post processing techniques the average recognition score can as much as been doubled and more.

We would like to point out that all of these post processing techniques also bring some drawbacks: if low-pass filtering then the system might not be able to follow fast emotional changes, if thresholding than we cannot classify all samples, and when inverse thresholding we may have to handle multiple estimates for one single sample of video.

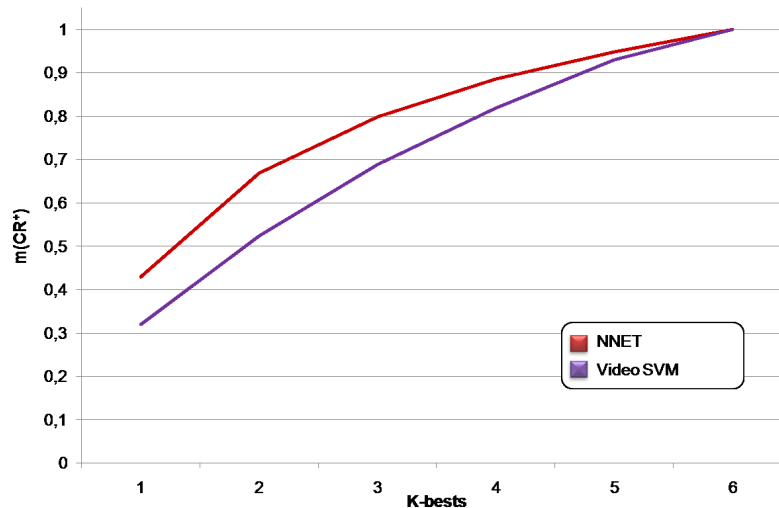


Figure 5.31: NNET and video SVM: k–best average recognition rates

### 5.4.13 Resulting System

Given the results we have shown in the few previous sections we have designed a system to perform multimodal emotion recognition. For each emotion we have computed three different neural networks using data respectively from the audio, the coordinate, and the distances feature sets. We have used the results of the previous study in section 5.4.5 to determine which data shall be used to recognized each different emotion.

In table 5.10 we detail the features we have selected for this study.

Emotion	Audio features	Coordinate features	Distances features
Anger	Energy, Pitch, & HNR	Eye Region	Head Displacements
Disgust	LPC Coefficients	Eye Region	Eye Region
Fear	MFCC Coefficients	Eye Region	Head Displacements
Happiness	Energy, Pitch, & HNR	Mouth Region	Mouth Region & x Displacement
Sadness	LPC Coefficients	Mouth Region	Mouth Region
Surprise	Formants	Mouth Region	Mouth Region

Table 5.10: Selected features for the different emotions

Video features are pre–filtered with a five frames long low–pass filter to reduce the complexity of the feature point movement. We have evaluated that we did not need such a phase for the audio features since the extraction phase is much more precise and reliable. We have included harmonicity into the set of pitch and energy and taken off the 10<sup>th</sup> LPC coefficient which was equal to 0 for all samples as seen in section 5.4.5.

For each feature we have, then, computed the mean value and the standard deviation, as well as the mean and standard deviation values for the first two derivatives. We have decided to adopt a single window length for each features and selected a sliding, overlapped (30/31) sliding window of 31 frames. According to our study in section 5.4.5 longer windows generally returns better results than shorter ones. We have selected this value as a trade–off between the stability of the input data and the capability of the system to follow fast changes of the user’s emotions. At the same time, the decision of having one single window length allow us to adopt much simpler fusion techniques.

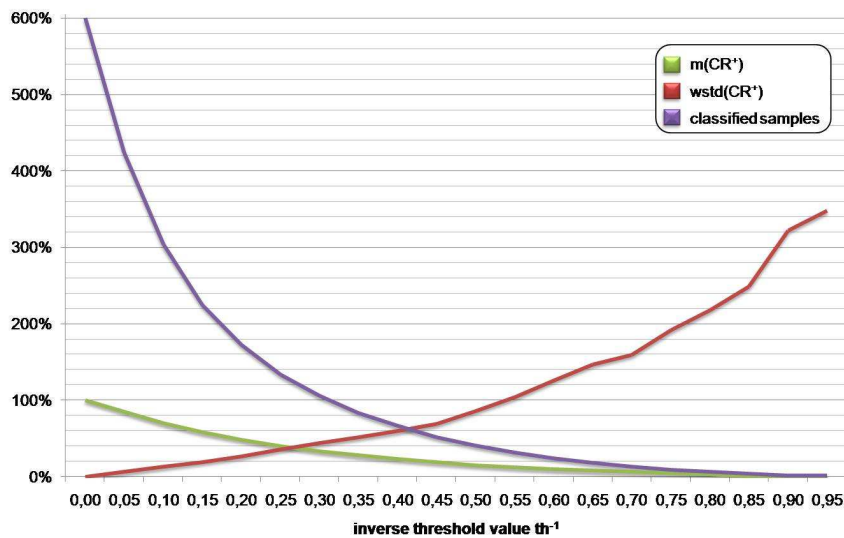


Figure 5.32:  $CR^+$ ,  $wstd(CR^+)$ , and `classified_samples` comparison for different inverse thresholding values

We have employed neural-networks with one hidden layer composed of 50 neurons which has been trained on a train set composed of 40 randomly selected subjects from the eINTERFACE'05 database (Martin et al. [2006]). The data was extracted with software written in C++ using Intel's OpenCV (IntelCorporation [2006]) library and saved the various features on three different files (one for each modality) per video-shot. I's OpenCV (IntelCorporation [2006]) library and saved the various features on three different files (one for each modality) per video-shot. These data were fed to the networks for a maximum of 50 times (epochs) using the MATLAB 2009 neural network toolbox. We have used 10% of this dataset for validation purposes. The remaining 4 subjects from the eINTERFACE database were used for test. It is interesting to notice that the whole training process last about 1 hour and 10 minutes on a Pentium(R) 4 2.10 GHz with 1,00 GB of RAM memory; the extraction of one estimate with the constructed neural networks last less than  $5 * 10^{-5} seconds$ . We have repeated these operations 5 times using different subjects for test and training and averaged the results.

The output of the 18 resulting neural-networks have been filtered with a low-pass filter of 25 frames to improve the results as seen in section 5.4.12.

For each emotion we have employed a Bayesian approach to extract a single multimodal emotion estimate per frame  $o_{emo}$ . The Bayesian approach as been preferred to other simple decision level fusion approaches and to the NNET approach (section 5.4.9) as one returning very good results without the need for training. The resulting system, simply detecting the most likely emotion by searching from the maximum estimation between the 6 different detectors perform an average recognition rate equal to 45.3% ( $wstd(CR^+) = 0.73$ ).

We have computed the minimum, maximum, average  $m(o_{emo})$ , and standard deviation  $std(o_{emo})$  values for each one of the detector outputs. We have then proceeded to normalize the output of the six different detectors to have a minimum estimate equal to 0



and by normalizing the average output for each one of the 6 different emotions. This operation raised the mean recognition rate to 50.3% and decreased the weighted standard deviation to 0.19.

Finally we have defined two thresholding profiles returning respectively around 12.5% and 50% of the samples by setting both a lower thresholding value and an upper inverse thresholding values as seen in sections 5.4.12 and 5.4.12. Other thresholding profiles are possible which also give similar or different percentages of recognized samples. Indeed, by increasing the threshold or decreasing the inverse threshold we also decrease the number of recognized samples and vice versa. Almost infinite profiles can, therefore, be defined which return about the same number of estimations.

We have selected these two specific ones having in mind different possible application scenarios. In the case in which one would want real-time frame-to-frame estimation than no thresholding will be applied. After a couple of seconds all buffers in the system will be filled, and the system will start returning one emotion estimation per frame.

A second scenario could be the one in which the user needs precise estimations without being too much concerned about how often these estimations are coming. In this case a scenario tagging around 12% of samples will be suitable to increase the recognition while having an average of roughly 3 estimation per second. We want to observe than an average of 12% of emotionally tagged samples may turn out not to estimate any samples for few seconds and than estimate one second of video frame-by-frame. Lower recall values, while still improving the precision, have to be generally discouraged because they might bring the system not to tag same long consecutive part of the video (or, in our case, whole video-shots).

A third application scenario is the one which stays in the middle to these two (50% of frames are tagged with an emotional estimation). In this case quite often (in average every other frame) one or more emotional estimations are returned to the user.

The two new obtained systems are capable of correctly evaluating respectively 61% and 75% of the recognized samples. In table 5.11 we report the specific thresholding settings and the originated results.

Recall	Thresholding Values	Inverse Thresholding Values	$m(CR^+)$	$wstd(CR^+)$
49.7%	$m(o_{emo}) + 1.2 * std(o_{emo})$	$m(o_{emo}) + 2.0 * std(o_{emo})$	61.1%	0.29
12.9%	$m(o_{emo}) + 3.0 * std(o_{emo})$	$m(o_{emo}) + 5.0 * std(o_{emo})$	74.9%	0.29

Table 5.11: Selected features for the different emotions

The two systems maintain low weighted standard deviation values while improving the mean recognition rate of the positive samples. Thresholds were defined as a function of the output mean and standard deviation values making the assumption that the distributions of the outputs of the detectors are Gaussians. In reality, the analysis of the output data did not confirm this hypothesis and suggested that better chosen threshold may further improve the results.

In table 5.12 we report the confusion matrix for the second system.

With the sole exception of the emotion surprise which is often confused with anger, fear, and sadness all emotions are recognized in more than 50% of the cases and that happiness is recognized in 99% of the samples in our test bases. Anger and disgust are sometimes reciprocally confused as well as fear and sadness.

Surprise is the emotion which our system recognizes with most difficulties. This result was to be expected from our previous studies and from the theory. As we pointed out

in \ out	Anger	Disgust	Fear	Happiness	Sadness	Surprise
Anger	87%	11%	0%	2%	0%	0%
Disgust	14%	60%	0%	6%	21%	0%
Fear	0%	10%	75%	0%	15%	0%
Happiness	1%	0%	0%	99%	0%	0%
Sadness	15%	20%	1%	0%	64%	0%
Surprise	21%	2%	14%	7%	15%	41%

Table 5.12: Confusion matrix of the resulting multimodal system

previously, surprise is theoretically hard to distinguish to other emotions as it is most often at least slightly valenced: therefore, we have positive surprise as in sudden happiness or negative surprise as in fear.

Applying machine learning such as a second NN could further improve the results by taking into account intra-emotional information at the cost of the increased complexity and the need for some extra training data. For example a NN could train to recognize surprise when all other estimations are in between certain values.

## 5.5 Perspectives: SAMMI

Multimedia information indexing and retrieval research is about developing algorithms, interfaces, and tools allowing people to search and therefore to find content in all possible forms Lew et al. [2006].

Although research in these fields has achieved some major steps forward in enabling computers to search texts in fast and accurate ways, difficulties still exist when dealing with different media such as images, sounds, or videos.

Current commercial search methods mostly rely on metadata such as captions or keywords. Such metadata must be generated by a human and stored alongside each medium in the database. On the web this metadata is often extracted and extrapolated through the text which is close to the media, assuming a semantic connection between the two. Although this is often true, in many cases this information is not sufficient, complete, or exact. Furthermore, in some cases this information is not even present.

Content-based methods have been designed to search these kinds of media through the semantic information intrinsically carried by the medium itself. The term “content” in this context might refer to colors, shapes, textures, tempos, frequencies or any other information that can be derived from the medium itself. Content-based techniques are designed to automatically extract information about the content of the media through computer based algorithms. Other efforts are spent in finding smart ways to use this information.

The trends of this kind of research are twofolds: on the one hand academia is looking for new feature sets which represent media in better ways; on the other hand an effort is done in designing algorithms, and interfaces, which allows for queries which better exploits current feature set dimensionality.

Multimedia content analysis addresses both low-level features extraction and semantic content recognition. Low level features are characterized by the estimation of visual, auditory (and others) features such as colors or frequencies; the semantic content recognition tries to appraise the semantic meaning of the media extracting information such as

---

objects, events, genre, and others.

One of the main challenges in content-based multimedia retrieval remains the bridging of the semantic gap still. The semantic gap characterizes the difference between two descriptions of an object by different representations. In the domain of indexing and retrieval we refer to the difference of abstraction which subsists between the extracted low level features and the high level features requested by the human's natural queries.

Emotions have been demonstrated to influence many different human cerebral functions (Picard [1997], Frijda [1986], Ortony et al. [1988], Lazarus [1991], Damasio [2005], Scherer [2001], Lisetti and Gmytrasiewicz [2002]). Among the other influences one is more interesting than the others when dealing with indexing and retrieval systems: this is the influence emotions hold on memory.

It is demonstrated that events and object appraised as emotionally relevant are memorized in more permanent ways Hamann et al. [1999], McPherson [2004] but also, that the organization of memory is such that similar memories (i.e. which elicit similar emotional reactions) are linked among them. Therefore, there are suggestions that emotions are an important characteristic of human memory which allow us to easily retrieve the memories we are looking for (Damasio [2005], Lisetti and Gmytrasiewicz [2002]).

Since Emotions are a fundamental characteristic of all sorts of art forms, it seems, in many cases, very reasonable to use emotions for indexing and retrieval tasks. For example in music recommendation systems one could argue it would be much simpler to ask for "romantic" or "melancholic" music than to define its genre, author, album or title (One example could be yahoo launchcast (music.yahoo.com) which allows defining radio preferences called "mood").

The same observation can be made for films or books. Film genres (and book too) are strongly linked to emotions as can clearly be seen in the cases of comedies, romances, or horrors. In other cases such as in adventure or musical movies the link between emotions and film genres is less clear. Salway and Graham [2003], Chan and Jones [2005] suggest that there may be, in these cases, links between the evolution of emotions in films and their classification. Action movies could be, for example, characterized by having an ongoing rotation of surprise, fear, and relief.

Albeit studies from the indexing and retrieval community (Lew et al. [2006], Ornager [1995], Dimitrova [2003], Sebe et al. [2003]) acknowledge that emotions are an important characteristic of media and that they might be used in many interesting ways as semantic tags, only few efforts have been done to link emotions to content-based indexing and retrieval of multimedia.

Key works in this domain started with Salway and Graham [2003] with the extraction of emotional feature from the transcriptions of audio-descriptors of films for visually-impaired people. 679 different words were considered as emotion tokens belonging to one of the 22 different emotions described in the Ortony et al. [1988] model.

A second study from Miyamori et al. [2005] shows how the extraction of such characteristics can arise from the written commentary of people watching a TV show. The researchers took emotional text from blogs and used it to summarize American football games.

Chan and Jones [2005] decided to focus on film audio; through the simple analysis of pitch and energy of the actors' speech signal the two researcher extrapolated information about the expressed emotion. This emotion is used to index films. A very simple evaluation of the retrieval system was conducted with a positive outcome.

Kuo et al. [2005] used films music and algorithms exploiting features such as tempo,

---

melody, mode, and rhythm to classify music into one out of 22 different emotions (Ortony et al. [1988]). Experiments on the retrieval system showed that the proposed approach could achieve 85% accuracy in average.

Finally, Kim et al. [2005] used information about texture and colors of an image to extrapolate the emotion elicited by that picture in humans. Experiments on a base of 160 pictures have shown that their approach can achieve nearly 85% accuracy.

These studies show the interest for such a kind of emotional content-based retrieval systems, but often present few main limitations. Firstly, most of these systems lack of an appropriate evaluation study. Secondly, the emotion recognition algorithms, which are actually very simple and not really up-to-date, are in some cases not evaluated at all; the evaluation study is often limited to a subjective estimate.

Furthermore, we argue that emotions should not represent the only media characterization; many other tags about the content of the media shall be used together with emotions to have complete systems. In the case of music recommendation systems, it is true that people often wants to listen to melancholic or happy music but it is also true that sometimes we may be interested in a specific music genre, band, or song; in these latter cases emotion-based approaches will not be very useful.

Another example showing the importance of a multi-disciplinary approach could be where one is trying to retrieve an action movie: one possibility is to look for explosions but the very same explosions will be also present in a documentary about controlled building demolitions. Another possibility will be to recognize an action movie only through its emotion evolution but this recognition may be very complex. Both monomodal systems have good chances to fail the task, retrieving un-relevant movies. Instead, combining the two systems could facilitate good results with relatively low complexity: videos are selected which contain explosions and documentaries will be cut off since their general mood and their emotion dynamics are usually very different from the one contained in action movies.

A second interesting application where emotions could play a role is media summarization. The principle is simple: given a media which we want to automatically summarize we summarize it by selecting the more emotionally relevant parts. Miyamori et al. [2005] apply this principle to football matches with satisfactory preliminary results. Another possibility is that given a video and its general emotion categorization (e.g. fear-horror), we only select scenes which exhibit that specific mood (i.e. that are fearful) or that at least we use this information to modify the summarization process.

Current systems analyze the media and look for low level features such as color histograms, motion detection, and others. As for retrieval and recommendation, coupling emotions to this kind of features will probably give better results.

For instance, while summarizing an action movie, one may look for scenes regarding gunfights and therefore looking for shootings. Supposing there are, in the film, scenes in a shooting range, we may not want to select them. Looking at the content alone would return these scenes together with the real gunfights while only looking for emotionally relevant scenes instead would result in finding scenes which do not contains shootings at all. The combination of the two, however, will be able to return scenes which are emotionally relevant and do contain shootings and that are, therefore, likely to belong to gunfights.

We have said that the main limitation of current content-based retrieval system is linked to the semantic gap. We have now shown examples where emotions, thanks to their highly abstracted nature, could help to bridge this gap.

---

Another point in favor with the emotion-based systems regards the human centered approach in general. As such systems are generally designed for humans we ought to develop systems which work on similar basic principles to generate more natural, and therefore both pleasant and effective, interactions.

In this section, we have described how emotions can join other media content descriptors on order to improve upon the performance of content-based retrieval and semantic indexing systems. In the next section we describe SAMMI, a framework we propose which allows creating such a kind of systems.

**SAMMI Architecture** The objective of this section is to describe SAMMI, a framework explicitly designed for extracting reliable real-time emotional information through multi-modal fusion of affective cues and to use it for emotion-enhanced indexing and retrieval of videos.

There are three main limitations for the existing works on emotion-based indexing and retrieval that we have overviewed in section 5.2:

1. emotion estimation algorithms are very simple and not very reliable;
2. emotions are generally used without being coupled with any other content information;
3. the evaluation of the experiments is preliminary and quite incomplete;

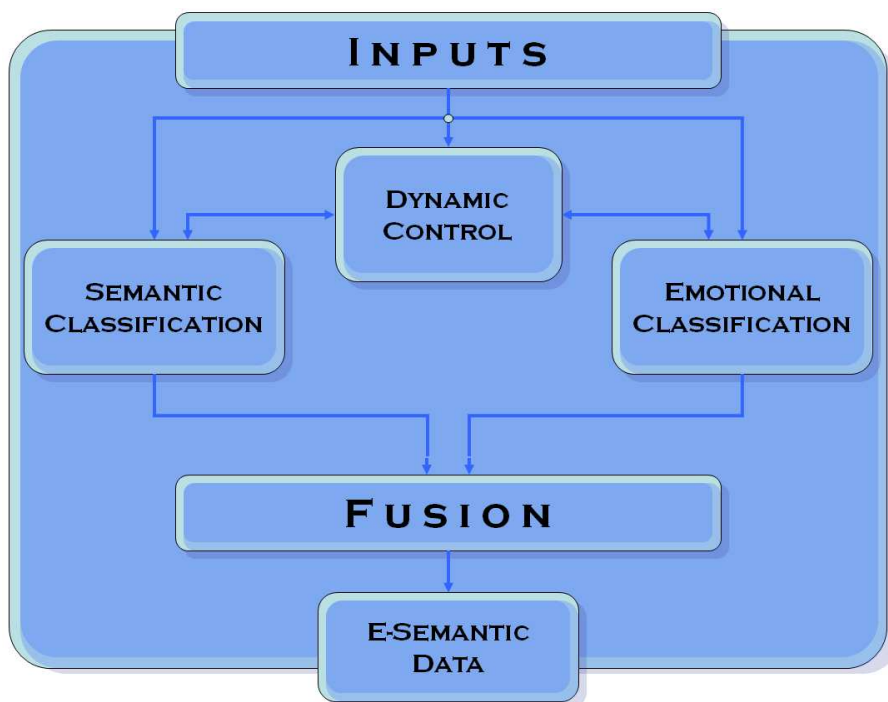


Figure 5.33: SAMMI's architecture

**SAMMI estimates emotions through a multimodal fusion paradigm.** Speech is analyzed and different feature sets are extrapolated: pitch, pitch contours, speech formants,

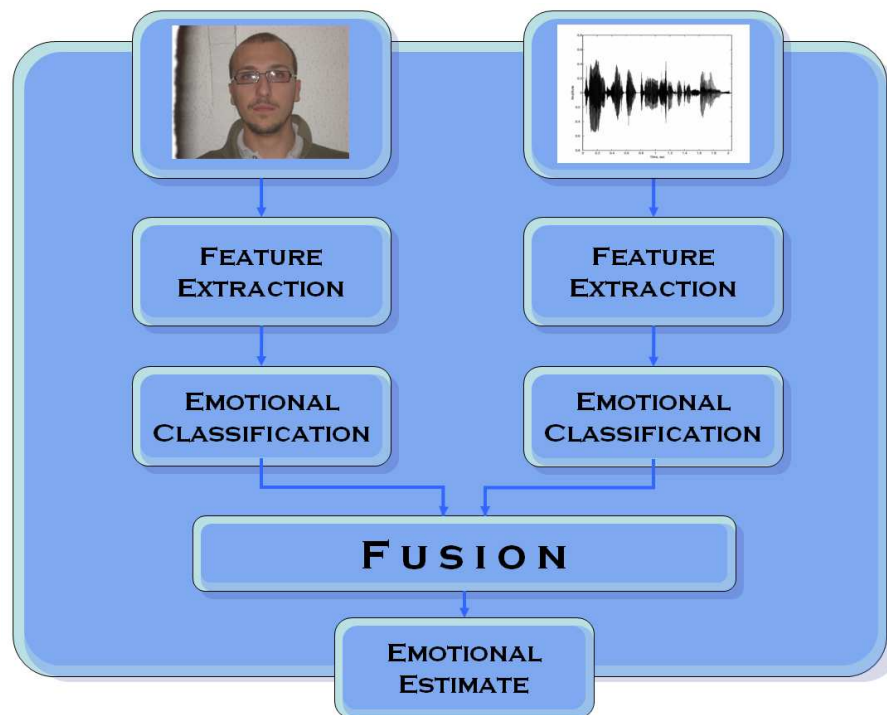


Figure 5.34: Multimodal Emotion Recognition

energy, MFCC, and Rasta-PLP. Those feature sets are fed to different classification systems (e.g. HMM, GMM, and SVM) in order to have different emotion estimates to compare.

At the same time a face is found in the video and the expression is analyzed through motion flow and feature point positions and movements; these features are also fed to different classification systems. Multimodal feature fusion will be experimented which will lead to additional emotion estimates.

The result is an array of different emotion appraisals which are therefore fused on extrapolating a single emotion estimate (see Fig. 5.34) characterized by arousal and valence as found in Chan and Jones [2005].

Dynamic control (Fig. 5.33) will be used to adapt the multimodal fusion according to the qualities of the various modalities at hand. Indeed if lighting is inadequate the use of color information should be limited and the emotion estimate should privilege the auditory modality.

**SAMMI couples emotions and other semantic information.** The emotional information will be then coupled with other semantic information extrapolated in parallel from other algorithms Benmokhtar and Huet [2007a], Galmar and Huet [2007] as can be seen in figure 5.33. Albeit emotional information is obviously a subset of the semantic information, in this figure we decided to display the dedicated recognition module as separated from the one devoted to other semantic information to point this former module out and because we believe more importance, than the one currently given, should be given to such a form of semantic information.

The extraction of different feature sets from the same media, as well as the application



of different classification techniques and the use of different modalities are all characteristics which assure good reliability; the use of dynamic control assure stability in presence of noise.

One important improvement involves the fusion of emotion-based features with other content-based semantic features.

## 5.6 Concluding Remarks

In this part of this document we have presented the complex topic of emotion recognition.

We have presented ARAVER a system to perform automatic emotion recognition from the facial expression and the vocal prosody of a speaking subject and SAMMI, a framework designed for semantically tag multimedia excerpts with emotional information.

We have also proposed AMMAF a multimodal multilayer affect fusion paradigm which is designed to optimize the results of unimodal emotion recognition system by fusing affective information in an intelligent way by performing different steps of filtering, multimodal fusion, and synchronization.

In the following sections we have presented results obtained with different setup of our recognition system on the eNTERFACE'05 multimodal emotional database. In particular we have seen how different feature vectors, machine learning approaches, multimodal fusion paradigms, and post-processing techniques can be combined to recognize emotions.

Finally, we have presented the results of the ARAVER system exploiting the knowledge acquired during the tests presented in the former sections. The system employ six multimodal detectors (one per each emotion) which are each obtained by fusing the output of three neural networks (one per each modality: audio, coordinates, distances) with a Bayesian approach. Each detector uses different data which was evaluated as the best for recognizing that particular emotion in a previous study.

Pre and post-processing low-pass filtering are applied to the input video data and to the output of the different classifiers. A simple normalization is employed to normalize the six different detectors which are then compared using the “*max()*” operator.

Both thresholding and inverse thresholding are applied respectively to reduce the number of samples which are classified with a too low probability ( $P_{emo} = 0$  if  $P_{emo} < th$ ) and to increase the number of samples which are classified when their likelihood is not the maximum one but still bigger than the inverse threshold  $th^{-1}$  ( $P_{emo} = 1$  if  $P_{emo} > th^{-1}$ ). Thanks to this processing we obtain a final system with an average precision equal to 75% for roughly 13% of recall. Without changing the system better average precisions can be obtained by increasing the thresholding value with the drawback of further reducing the recall of the system.

Future work on this subject shall concentrate on four main axis:

1. design and implementation of a system allowing to extract “cleaner”, more precise, and more reliable video features (e.g. using active appearance model to gain robustness to all head movements);
2. design and implementation of multimodal feature fusion techniques (e.g. fusing audio and video information about the mouth movement to reduce the influence of speech production to facial expressions);



3. implementation of a multimodal dynamic decision fusion systems which also take into account the reliability of the different modalities to compute the multimodal estimation;
  4. use of other non-intrusive modalities such as pose, gestures, word semantic, etc.
-

---

## Chapter 6

# Emotion and Artificial Intelligence

### 6.1 Introduction

Since the fifties, researchers started to investigate the possibility of replicating human intelligence through software. The idea that a machine could “think” was at that time only fiction but, in the years that followed, it has become more and more real.

Despite the success of AI in replicating human basic reasoning mechanism, starting from the eighties, more and more people have argued, or demonstrated, that rationality is not enough for modeling human thinking nor to replicate or simulate human behaviors (Damasio [2005], Picard [1997], Lisetti and Gmytrasiewicz [2002]). It seemed that something was missing in the model of the agent, something that did not allow the agent to make a decision fast enough nor necessarily to take the decision a human would have really taken. As it is said by Dean Spooner (a.k.a. Will Smith) in the movie “I Robot” (on the homonym Isaac Asimov’s novel) “*A human would have understood*” what a robot could not.

Studies show that a key missing component was emotion. Marvin Minsky [1988], one of the founders on Artificial Intelligence in his “*Society of mind*” said: “*The question is not whether intelligent machines can have emotions, but whether machines can be intelligent without any emotions*”.

As already pointed out in section 3.1.1, Damasio [2005], in his book “*Descartes error*”, shows how people with difficulties in showing or feeling emotions present also difficulties in making decisions. Subjects discussed in his book (e.g. Phineas Gage and Elliot) demonstrate problems in understanding emotional behaviors and they often show problems with decision making. On the other hand their intelligence, or rather their capabilities of logical reasoning seem to be untouched.

Consequently one can say that rationality, as it is defined by Descartes and Newton, is not sufficient to model human behavior. To solve this problem, Lisetti and Gmytrasiewicz introduced the definition of *rationality*<sub>1</sub> (Lisetti and Gmytrasiewicz [2002]). They write: “*rationality*<sub>1</sub> includes the role that emotions play in rationality in their plethora of ways during the human decision making process [...] Contrary to the Cartesian rationalist tradition, *rationality*<sub>1</sub> and emotion can be considered not as necessarily opposed, but clearly different faculties, and their differences can be considered as allowing each to serve a division of labor in which their distinct capacities contribute to a unified outcome...” In their opinion, *rationality*<sub>1</sub> cannot be seen separate from the affective state of a human being. According to the new definition of *rationality*<sub>1</sub> the reasoning part of the mind cannot be considered separate

---

from the emotive one.

On the other hand studies have demonstrated people have a tendency to treat computer as social being (Reeves and Nass [1996], Cassell [2000]) or yet how people prefer to interact with agents capable of demonstrating some form of emotional capacity (also while the same subjects were assessing that the agent should not “have” emotions)(Koda and Maes [1996]).

Affect influence *rationality*<sub>1</sub> in different ways at several levels :

- Firstly, affect act as a filter for what a person sense from the environment or in other words When an agent passes from an emotional state of say, sadness, to one of anxiety, the agent reduces the set of the states of the world to  $S_0 \ll S$ , where  $S_0$  mostly contains states associated with negative valence. That happens both from the physical (i.e. the person only senses some states) and the mental sides (i.e. the person interpretation about those states changes).
- Secondly, affect can influence the desires a person has and in particular the importance of a desire with respect to the others changes. That can be translated in two main phenomena: firstly people do not usually desire with strong intensity what it is impossible to have as that will finally elicit a bad mood (a.k.a. behavioral disengagement); secondly, people usually subconsciously appraise desire “importance” differently regarding their mood.
- Thirdly, affect can create, modify or delete beliefs about the environment. For example, take when one meets someone who makes one feel vaguely uncomfortable. One cannot necessarily formulate a belief about the person that justifies that emotion, but one can infer from that emotion that one has such a belief.
- Finally, emotions and mood changes for sure the way a decision is made, as we have said by changing the initial parameters, (belief and desires), but also by helping the process selecting the most appropriate, or rather preferred, possible actions. This also passes through a process that Damasio [2005] defines somatic markers.

From those considerations it seems both useful and interesting to develop agents not only able to behave emotionally, like humans do, but also able to use emotions in a wide range of *reasoning*<sub>1</sub> functions like decision making, memory and the appraisal of events.

In the next chapter we will overview some of the existing techniques and technologies for building intelligent and affective agents.

Chapter 6.3 presents, then, the result of our work on AI and the affective computing topic of *emotion synthesis*. In particular we will present ALICIA, the architecture we developed to build affective intelligent agents. ALICIA makes use of the techniques overviewed in the next chapter and improves them by simulating part of the Scherer [2001] appraisal process and of the influences that affect hold on *rationality*<sub>1</sub> as they are described by Lisetti and Gmytrasiewicz [2002].

## 6.2 Relevant Work

This chapter presents the most relevant theories about modeling agent’s internal states, behaviors and interactions with the environment. The chapter will then evolve in a brief description of existing intelligent and affective agents. We will present three agents in

---

particular: GRETA (de Rosis et al. [2003], Poggi et al. [2005]), EMA (Gratch and Marsella [2004], Marsella and Gratch [2009]) and VALERIE (Paleari et al. [2005]). In the next chapter we will present ALICIA, our generic intelligent and affective architecture.

## 6.2.1 Basics on Artificial Intelligence

Artificial Intelligence (AI) officially started in 1956 when the name was coined but the basic idea of computers acting intelligently is much older.

As early as in 1950 Alan Turing proposed a test called the Turing test which was designed to provide a definition of intelligence. Turing said that a computer can be defined “intelligent” if it has the ability to achieve human-level performance in all cognitive tasks. In his idea a computer needs to possess abilities for natural language processing, knowledge representation, automated reasoning, machine learning, computer vision and robotics. The basic idea for the Turing Test is: *“if a person speaking to the computer through a teletype cannot tell if there is a computer or another human at the other hand then the computer has passed the test”*.

In its first years, the fifties and the sixties, AI (Artificial Intelligence) reached great success. While intellectuals were saying a machine would never do something programmers demonstrated it can. Neural Networks were used and there were people believing that before year 2000 a human mind would have the possibility to be copied and stored on a computer.

In 1961, Stanley Kubrik, in his “2001, space odyssey”, anticipate the idea that machines can be not only intelligent but also emotive.

After this “golden era”, artificial intelligence started to show some limitations. Computers could not easily translate text or understand speech; machines were not really able of learn and find solutions outside the fields they are programmed for. Such kinds of limitations were found and neural networks seemed not to work well for all problems.

In the eighties, AI becomes an industry, commercial software using theories and techniques originally developed for artificial intelligence started to be created and sold. In the same years the first theories about emotions were developed.

In the recent times AI has become more and more available: videogames make intense use of AI, some of them uses emotions to make their characters behave in a more believable way, software are sold and are claimed to be intelligent, etc.

### 6.2.1.1 BDI

BDI or Belief Desire Intentions is the most common technologies for modeling human goals and knowledge (Rao and Georgeff [1991, 1995]). The model is based on three basic entities, the *beliefs*, the *desires* and the *intentions*.

**Beliefs** are used to represent the knowledge of the agent. Everything the agent knows about the world, about objects in the world and their mutual relations would be beliefs. A belief can also be related to an absolutely imaginary world. For example one can have beliefs about creatures that do not exist and that would never exist outside one’s imagination.

**Desires** represent the goals of the agent, what the agent desires, or the final states the agent would like to reach. Intermediate states the agent has to reach to attain one of its

goals would become sub-goals and therefore would be considered as desires themselves. Examples of desires are “eat a banana” or “be in London”. Usually, in the literature states the agent would like to avoid (i.e. aversions) are considered as desires. For example “not being killed” would still be a desire.

**Intentions** represent the possible plans the agent has to modify the environment or his internal state. Another name which is usually given to intentions is coping strategies. Therefore when saying “to cope” it is meant to select an intention. Possible intentions are “take the car”, “go to the supermarket”, or “buy a banana”.

At all times reality can be represented as a list of states. For example if Alex and Bob are at home looking at the television then an agent can represent this knowledge as a set of states like:

- State 1 = Alex at Home
- State 2 = Bob at Home
- State 3 = Alex watching the television
- State 4 = Bob watching the television
- ...

In this case desires can also be defined like a set of state the agent would like to reach. Therefore, if the agent Charles is in Milan in Italy, but he would like to go to Turin the state “Charles in Milan” is a belief while the state “Charles is Turin” is both a belief (never happened and therefore with likelihood equal to zero) and a desire.

Intentions are the actions the agent can do to modify current states. Referring to Charles example then the action “take a car and go to Turin” is a plan the agent can take. If the agent selects it then this action would become an intention of the agent. Also the action “eat a sandwich” would be a plan and therefore a possible intention of the agent but it will not help the agent to reach its desires.

Agents within this BDI architecture would need to create new plans.

**Plans** are set of actions that change the current state approaching new states and avoiding other ones.

For example, Alex is at home and would like a banana. Alex knows there are not bananas in her fridge, that her car is in the garage, and that one can find bananas at supermarkets. Current state could be:

- State 1 = Alex at Home
- State 2 = Car in the garage
- State 3 = No bananas in the fridge
- ...

Alex would have a goal:

- Goal 1 = Eat a banana

In this case Alex could for example plan to:

- 
- Step 1 = Go to the Garage
  - Step 2 = Take the car
  - Step 3 = Go to the supermarket
  - Step 4 = Eat the banana

Probably a real person would like to take more than one banana, to pay for that and to go back home to eat the banana. Alex would not do any of these actions because her only goal was to eat the banana so she does not mind anything else; neither to breach the law or to go to prison. If her goals would have been:

- Goal 1 = Eat the banana
- Goal 2 = Not go to prison
- Goal 3 = Eat at home
- ...

In such a situation Alex's behavior would have been much different. We have, therefore, already seen that to have a believable behavior from Alex than a large number of desires to be set, some of them would be related to Alex's needs and objectives, some others to social rules, internal standards or simple preferences. At the same time, we have also seen that states need some basic characteristics such as likelihood. In section 6.3.2 we will overview a possible representation for the main components in BDI architectures. In the next sections we describe some techniques which can be used to represent interactions among states, desires, and intentions.

### 6.2.1.2 Belief Networks

One of the simplest ways to organize belief is the one commonly known as belief network (BN). A BN is a graph in which:

1. Beliefs make up the nodes of the Network
2. Arrows connect pairs of nodes. For example an arrow could connect node A to node B. That would mean that node B depends on A and that node A has direct influence on node B
3. Nodes are parents of the nodes they have arrows pointing on
4. The graph has no direct cycles
5. Each node memorizes the probabilities the sons occurs given that it is occurred

Let's consider this example from "*Artificial Intelligence: a modern approach*" (Russell and Norvig [1995]). One has a burglar alarm fairly reliable at detecting the burglary that also, sometimes, starts in occasion of an earthquake. One also has two neighbors, Alex and Bob, who promised to call if they heard the alarm. Alex always heard the alarm but sometimes confuses the telephone ringing with the alarm and calls then too. Bob, on the other hand, likes loud music and sometimes misses the alarm. One would like to compute the probability of a burglary given the evidence of who has called.

---

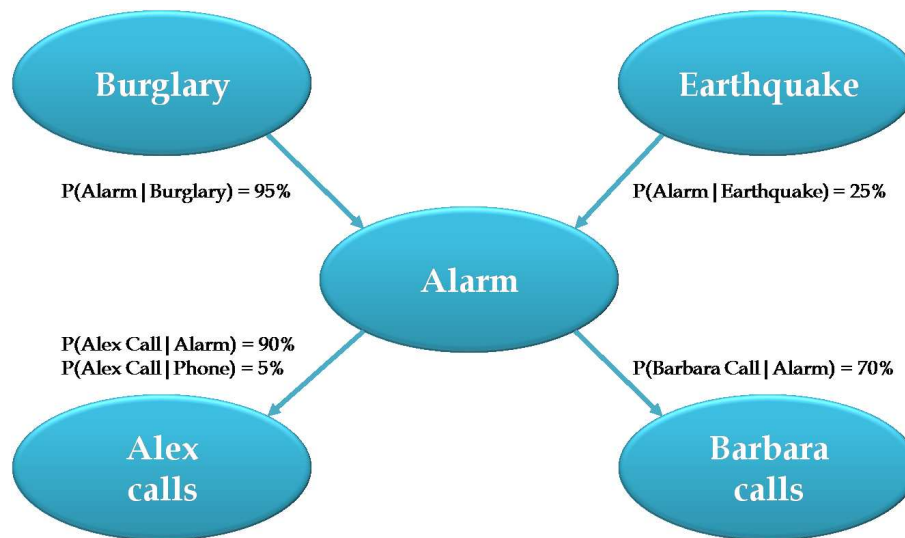


Figure 6.1: A simple Belief Network

The resulting network looks like the one in figure 6.1.

In the example all probabilities related to the events that can occur are given and stored in some way as information relative to the beliefs themselves. If one receives a call from Alex, he would know that with the 90% his alarm rang but it also know that there is a 5% that Alex called without reason and so on and so forth.

### 6.2.1.3 Decision Networks

Belief Network doesn't allow to see where actions have their influences on current states and therefore how the agent plans influence future states. To solve this problem Decision Network (DN) were defined. DN are basically BN but instead of having only belief as node of the network DN can also have action or *decision nodes*.

The probability of the states will be related to the action taken. For example, Dr. Tom has a problem with his six years old patient Jimmy (Marsella and Gratch [2003a]). Jimmy has cancer and is in pain; giving morphine to Jimmy Dr. Tom will end his suffering but might as well hasten Jimmy's death. In this case, Dr. Tom's decision network while trying to appraise the possibility of administering morphine to Jimmy would be as in figure 6.2.

In this example the probabilities of hastening Jimmy's death and of end Jimmy's suffering depends on the chosen medical treatment. For example, Dr. Tom believes that giving morphine would have with a likelihood of 15% the effect of hastening Jimmy's death and 95% chances of ending Jimmy's suffering. Other medical treatment could have different probability so DN can be seen as a way to make BN dynamic, as the values of the probabilities change as a function of the decisions.

Those probabilities would depend on the agent history and beliefs. For example if a treatment did not work in the past the agent may think it will not work this time either.

### 6.2.1.4 Dynamic Belief and Decision Networks

Dynamic Belief and Decision Networks (DBN and DDN) are a tool developed to take into account the dynamic aspect of the environment as well to its uncertainty. When the



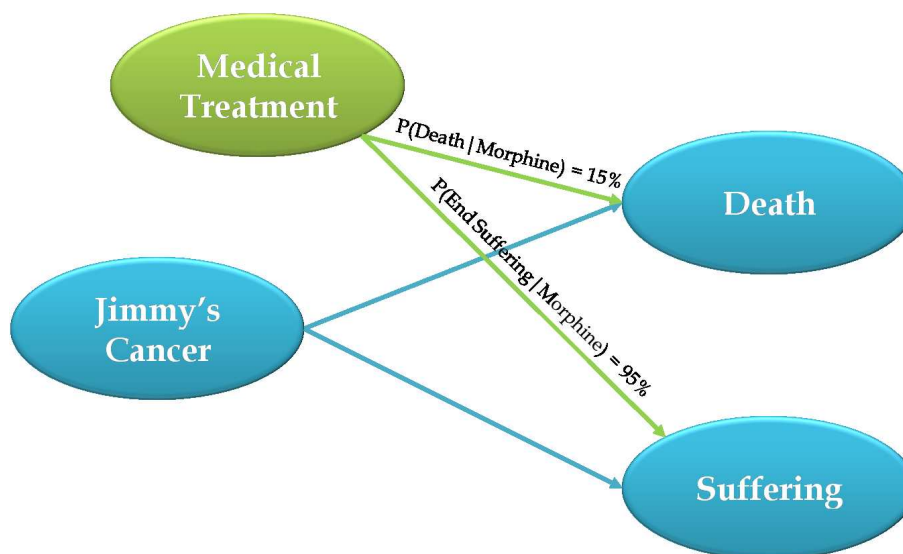


Figure 6.2: A simple Decision Network for Dr. Tom problem

environment is inaccessible (only partial, imperfect knowledge of the world is available) Believe Network and Decision Networks are not enough for decision making.

Dynamic Belief Networks have, for each time step, nodes for states and nodes for sensors variables. An agent will have access to the sensors but it would not have access to states. At the same time, the agent links sensors with states and estimates these latter. A simple DBN will look like the one shown in the figure 6.3.

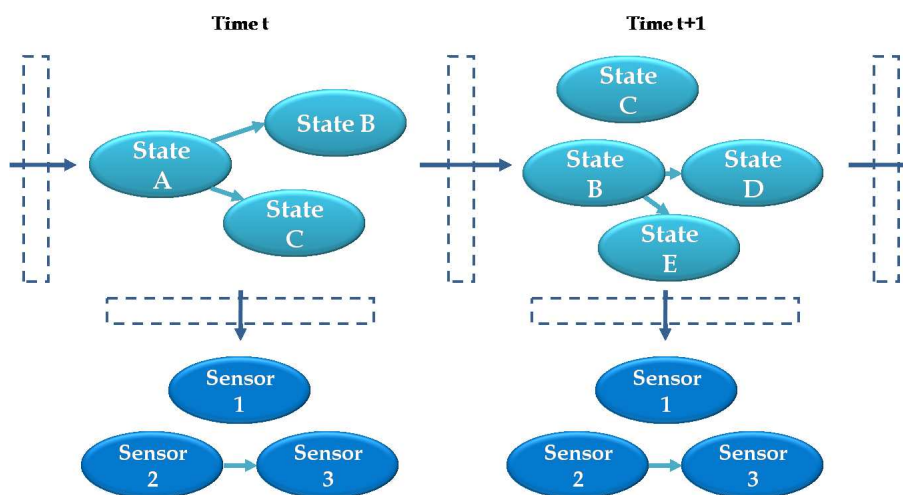


Figure 6.3: A simple Dynamic Belief Network

The simple addition of decision nodes to this model would generate Dynamic Decision Networks (DDN). The idea is fairly the same as for DN but as states are not accessible and current state are computed on a probability model, the states at the time  $t+1$  could be only computed with a likelihood of their occurrence. Usually the model used is based on Markov chain therefore literature also refers to this kind of problems as “*Partially Observable Markov Decision Problems*”.

### 6.2.1.5 Utility Theory

Once beliefs of the agent are modeled together with the evolution of the environmental states one needs to implement some mechanism to actually make decisions.

The simplest way to solve this kind of problem is to refer to a certain utility function and maximizing it. The utility function is a function linking a set of states to a number representing its *utility*. The main characteristic of a utility function is that if a given set of states is preferred to another, than it has a bigger utility.

Utility shall generally have two main characteristics:

1. when there are conflicting goals and only some of those can be achieved at the same time then the utility function shall be able to evaluate which one shall be pursued first.
2. when there are several goals that can be achieved, none of them with certainty, utility function shall be able to weight the likelihood of success with the importance of the desires.

When an utility function is correctly set then the agent should be able to follow one strategy in order to achieve its desires by comparing utilities of the sets of states which the different actions/decisions/intentions leads to (Mao and Gratch [2004]).

## 6.2.2 CPT and OCC Model of Emotions for AI

In section 3.3.2.3 overviews of the Ortony et al. [1988] OCC and the Scherer [1984] CPT componential models of emotions are given. Albeit all of existing systems we know about use the OCC model of the emotion because of its simplicity, the advantages of using the latter theory by Scherer would be manifolds. Three of these advantages are particularly relevant:

1. Scherer's CPT of emotions can represent an almost infinite number of emotions through a simple process of appraisal but the OCC model can eventually only represent 22 different emotions. When we observe closely the OCC model we are under the impression that the model was developed only for computer simulations; this is not the case of Scherer CPT of emotions in which the process complexity seem to better model the human appraisal.
  2. Scherer gives, within the CPT of emotions, hint on how emotions influence the production of multimodal emotional expressions in terms of facial expressions, pose, gaze, vocal prosody, and autonomous nervous system signals. Such information facilitates the development of computer software performing recognition and display of emotions within Scherer's CPT paradigm.
  3. Scherer models three level of cognition, namely the reactive, the schematic, and the deliberative, facilitating the implementation of architectures based on the same paradigm. Building an architecture on this paradigm better allows simulating non cognitive/deliberative agent's behavior and in particular fast responses to stimuli.
-

### 6.2.3 Existing AI with Emotions

In this section, we will briefly describe how the theories about AI (BDI, utility functions, BN, DN, DBN, and DDN) and the ones about emotions (the OCC model by Ortony et al. [1988]) can be used to develop emotive agents (i.e. agents that cope and interact with the environment taking emotions in account).

The intent of this chapter it is not to give a fully comprehensive overview of affective intelligent agents, but only to describe the main features of the most relevant ones. For further information about the agents please refer to the works of Prendinger and Ishizuka [2004], Cassell [2000], Pelachaud et al. [2007], Prendinger et al. [2008], Ruttkay et al. [2009].

#### 6.2.3.1 GRETA

GRETA's body has been already discussed in section 4.2.2.5. In this section we overview the main features of GRETA's mind (de Rosis et al. [2003], Poggi et al. [2005]).

The objective of the MagiCster project (i.e. the European project under which GRETA has been developed) was not to mimic underlying mental processes or to replicate surface behavior but simply to produce a "believable" behavior for their agent. GRETA was designed to be a health assistant for eating diseases like bulimia or anorexia and she has a specific personality. Albeit different agent could have been developed with the same architecture, GRETA is a specific individual agent and not a generic platform for building agents.

GRETA's mind was designed within the BDI paradigm and the authors did include in GRETA's mental state the information about beliefs and goals that drive the emotions as well as the possible intentions of displaying or hiding them. Changing Beliefs Desires and Intentions should be enough to change GRETA personality and application scenario.

GRETA's personality is defined in terms of the Five Factor Model (a.k.a. OCEAN or CANOE) by Norman [1963]; beliefs are modeled in a Dynamic Decision Network.

Emotions are computed as consequence of the information about the mental state at time  $t$  and  $t - 1$  and are computed in function of only two variables: 1) uncertainty and 2) utility.

With those two values and according to the utility theory GRETA computes the variation in intensity of on a given emotion. It is interesting to know that GRETA's Dynamic Decision Network includes nodes for *Beliefs*, *Goals*, *Goal achievement* (describing the belief that one goal would be achieved), *Events*, and nodes that the authors call *Observable* nodes that they used to represent observable effect of events.

GRETA can have different layers of superimposed emotions, to represent emotional mixtures.

The communicative meanings are associated with dialog moves by APLM. For more details about this markup language please refer to section 4.2.2.5.

#### 6.2.3.2 EMA

EMA is Marsella and Gratch [2009] implementation of "A domain independent framework for modeling emotion" (Gratch and Marsella [2004]). The agent is based on a BDI architecture.

EMA takes deep inspiration from *Steve*, *Jack and Steve*, and *Carmen Bright Ideas*, three previous works of the same authors.

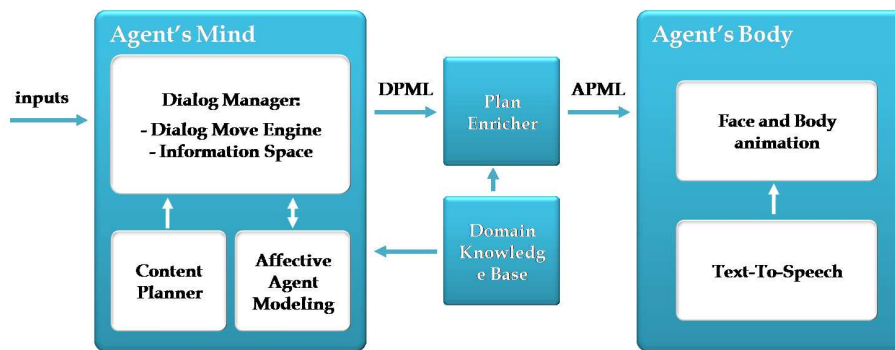


Figure 6.4: MagiCster system architecture

**Steve** is a 3D virtual agent able to interact with his environment explaining the user how to use instruments. For doing so Steve is able to perceive the state of the environment and to execute plans to reach his goals. The world could change in several ways as users interact with Steve and the environment itself can change unexpectedly for example due to simulated equipment failure. The model does not comprehend emotions and Steve only has limited set of possible reactions which are linked to objectives and goals.

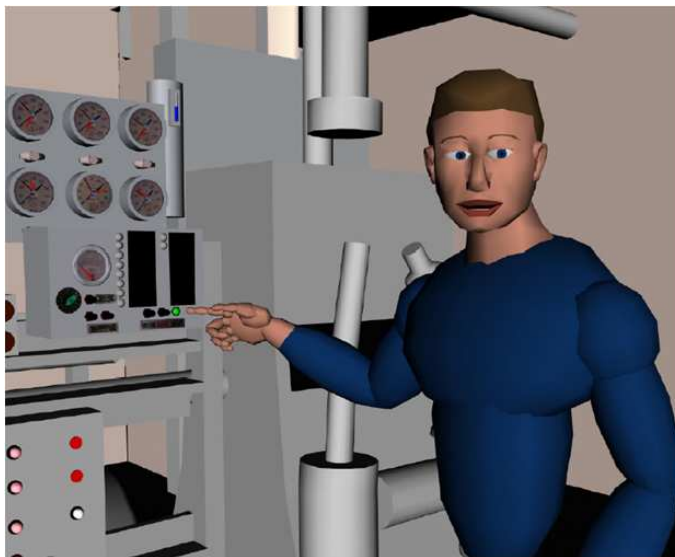


Figure 6.5: Steve describing equipment

**Jack and Steve** updates Steve model to include emotions and, above all, a way for an agent to take in account social interactions and other's plans and objectives. The new designed agents are now able to store each other goals and plans and, if personality traits allow that, to change their personal plans whenever the two agents contend a shared resource. Although Jack and Steve can show emotions they don't actually cope with regards to those emotions and emotions do not affect the agents' cognitive abilities.

**Carmen's Bright IDEAS** is an update of Steve AI showing how an agent can interact with the environment coupling logic and emotions. The system was designed to realize

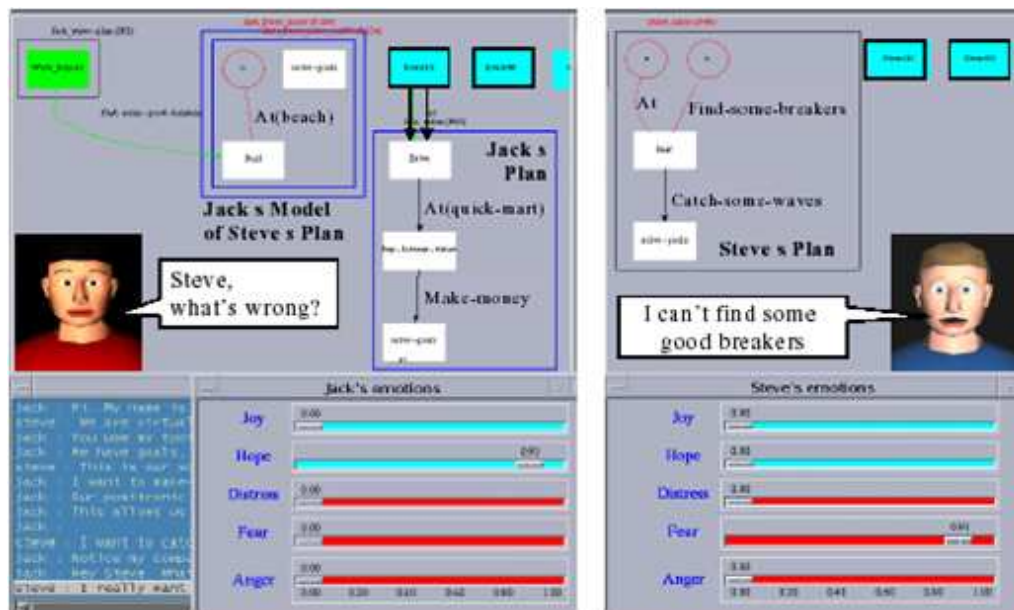


Figure 6.6: Jack and Steve appraisal representation

an *Interactive Pedagogical Drama*, an approach to learning that immerses the learner in an engaging, evocative story where Carmen interacts with other realistic character. The idea is to help a mother dealing with the stress she would face if she had a son or daughter affected by cancer. A mother learns by taking decision on behalf of a character in the story and by seeing the consequences.

The character in Carmen's Bright IDEAS shows believable behavior and emotions. The system is perfectly able to compute emotions felt as reactions to those emotions.

EMA supports interactive task-oriented dialog, real-time control over verbal and non-verbal behavior and responsiveness to external events (Gratch and Marsella [2004]).

As it was for Carmen's bright IDEAS cognitive processes and decision making depend on the appraisals of the events occurring in the environment and on the internal state of the agent. Gratch and Marsella described two different ways of coping to the events occurring in the environment:

1. by motivating actions that change the environment (problem-focused coping)
2. by motivating changes to the interpretation of this relationship (emotion-focused coping).

Gratch and Marsella [2004] consider twelve different appraisal variables (see table 6.1) but only five of them are actually used. These are: 1) *relevance*, 2) *desirability*, 3) *causal attribution*, 4) *ego involvement* and 5) *coping potentials*. It is not well specified if the sub-variables for causal attribution and coping potential are used or not.

The reason for not using the other variables is almost always linked to the lack of time information in the EMA software. The appraisal is based on Ortony Clore and Collins model but it is not better specified if and to which extent the model has been modified or adapted.



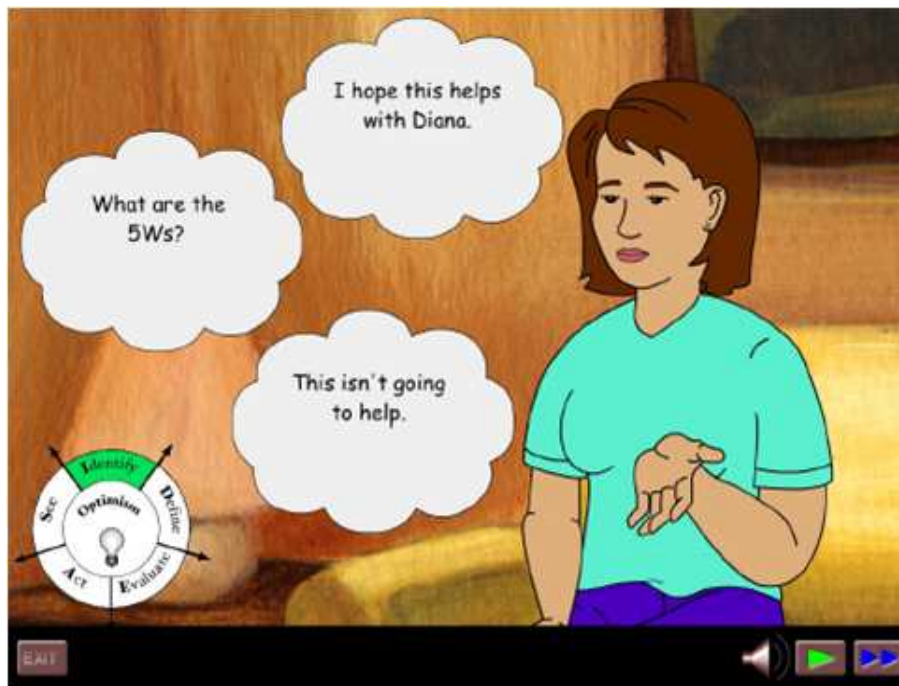


Figure 6.7: Learner influences Carmen by selecting Thought Balloons

Each time an event, either environmental or internal, occurs EMA tries to identify the so called coping opportunity. For doing this the coping module computes a “*coping elicitation frame*” that consists of some coping related fields. In their papers Gratch and Marsella describe three of them:

1. focus - agency: is the agent that provoked or initiated the event;
2. interpretation objects: tasks, states or individuals referenced by the causal interpretation of the event;
3. max interpretation: is the interpretation object with the strongest appraisal.

The successive task consists in elaborating a coping situation or rather to elaborate the “*elicitation frame*” taking into account other factors like relations with other agent that are interpretation objects of the eliciting frame. The idea, then, is to elaborate and propose alternative coping strategies. Some common coping strategies can be seen in table 6.2.

Once EMA has computed some alternative coping strategies there is a phase called “*assess coping potential*” that correspond to computing a utility for all the proposed actions. Finally, thanks to the utilities of the possible coping strategies one intention is selected and applied.

This approach basically uses all theories explained before in a simplified version. The work differentiates itself from others in this field because the system uses emotion interpretation and social relationships to bias the search of appropriate coping strategies. The described system is actually a framework that can be easily used for many different tasks. For example, Gratch [2000a] makes use of this technology in an e-learning scenario while Hill et al. [2003] simulates the behavior of military forces in Iraq.

Relevance		Does the event require attention or adaptive reaction
Desirability		Does the event facilitate or thwart what the person wants
Causal Attribution	Agency	What causal agent was responsible for an event
	Blame & Credit	Does the causal agent deserve blame or credit
Likelihood		How likely was the event; how likely is an outcome
Unexpectedness		Was the event predicted from past knowledge
Urgency		Will delaying a response makes matters worse
Ego Involvement		To what extent does the event impact a person's sense of self (social esteem, moral values, cherished beliefs, etc.)
Coping Potentials	Controllability	The extent to which an event can be influenced
	Changeability	The extent to which an event will change of its own accord
	Power	The power of a particular agent to directly or indirectly control an event
	Adaptability	Can the person live with the consequences of the event

Table 6.1: Appraisal variables according to Gratch and Marsella [2004] EMA

### 6.2.3.3 VALERIE

VALERIE (Virtual Agent for Learning Environment Reacting and Interacting Emotionally) (Paleari et al. [2005]) is a project built within the MAUI framework shown in figure 6.8 (Lisetti and Nasoz [2002]).

MAUI, Multimodal Affective User Interface, by Lisetti and Nasoz [2002] is a framework for building agents that can take into account affective input as Autonomic Nervous System signals, facial expression and vocal signal, to determine user affective state. The agent can talk and show emotions through an avatar designed with Haptek technologies. The intent of the project was to work on both agent body and mind modeling.

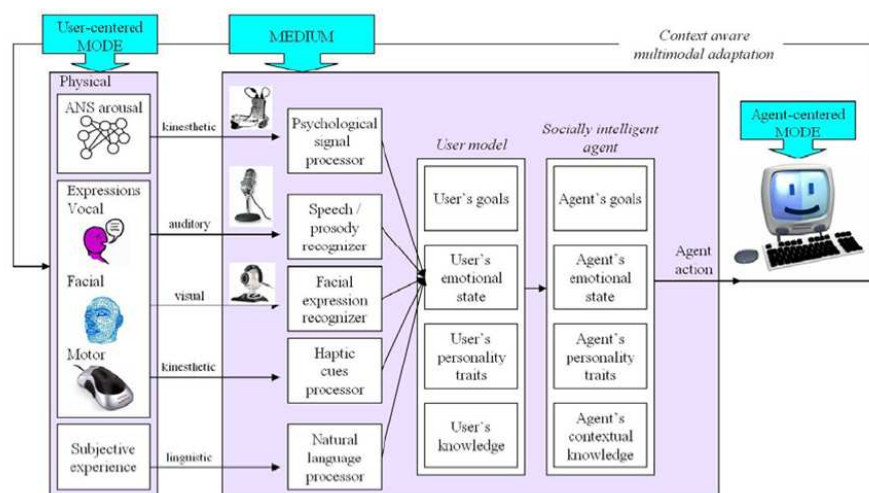


Figure 6.8: Multimodal Affective User Interface framework from Lisetti and Nasoz [2002]

For this project we designed a new avatar within the Haptek technology (see section 4.2.2.8 for details about Haptek avatars) We firstly animated it by developing new expressions within the software People Putty by moving nine sliders which are mapped to character's *happiness, sadness, anger, mellowness, suspicion, curiosity, ego, aggressiveness* and *energy*. Secondly we have animated it with the expressions generated within Scherer's



Problem-focused Coping	<b>Active coping:</b> taking active steps to try to remove or circumvent the stressor.
	<b>Planning:</b> thinking about how to cope. Coming up with action strategies.
	<b>Seeking social support for instrumental reasons:</b> seeking advice, assistance, or information.
	<b>Suppression of competing activities:</b> put other projects aside or let them slide.
	<b>Restraint coping:</b> waiting till the appropriate opportunity. Holding back.
	<b>Seeking social support for emotional reasons:</b> getting moral support, sympathy, or understanding.
	<b>Positive reinterpretation &amp; growth:</b> look for silver lining; try to grow as a person as a result.
Emotion-focused Coping	<b>Acceptance:</b> accept stressor as real. Learn to live with it.
	<b>Turning to religion:</b> pray, put trust in god (assume God has a plan).
	<b>Focus on and vent:</b> can be function to accommodate loss and move forward.
	<b>Denial:</b> denying the reality of event.
	<b>Behavioral disengagement:</b> Admit one cannot deal. Reduce effort.
	<b>Mental disengagement:</b> Use other activities to take problem out of mind: daydreaming, sleeping.
	<b>Alcohol/drug disengagement:</b> Make use of alcohol or drugs to take problems out of mind.

Table 6.2: Some common coping strategies from Gratch and Marsella [2004] EMA

component process theory (Paleari and Lisetti [2006d], Paleari et al. [2007b]).

We tried to model VALERIE's mind in the easiest possible way allowing believable behavior. The base for our work was the BDI architecture.

VALERIE has in the resulting model only two **desires** as a teacher:

1. giving feedback;
2. giving support.

We also implemented in VALERIE logic the implicit but relevant desire of having a good student.

These desires are mapped in term of three possible **intentions**:

1. give positive feedback if the user is performing well;
2. give positive/empathic support if the user is performing badly but it is trying to do well;
3. give annoyed or strict comment while the assumption that the student is concentrated is not satisfied.

The information about user effort is simply extrapolated by the information about the times used to answer questions and correct errors. In the future this information would be integrated by the affective information MAUI can extrapolate from Autonomous Nervous System signals.

The expressed emotion of the agent is computed from the information about the accomplishment of VALERIE's desires. Some parameters influence the "felt" and expressed emotions which are influenced by the agent's **mood** and **personality**.

Agent mood represents an unfocused, slowly modifiable, affective phenomenon. The agent automatically adjust its parameters trying to maintain its mood as described and

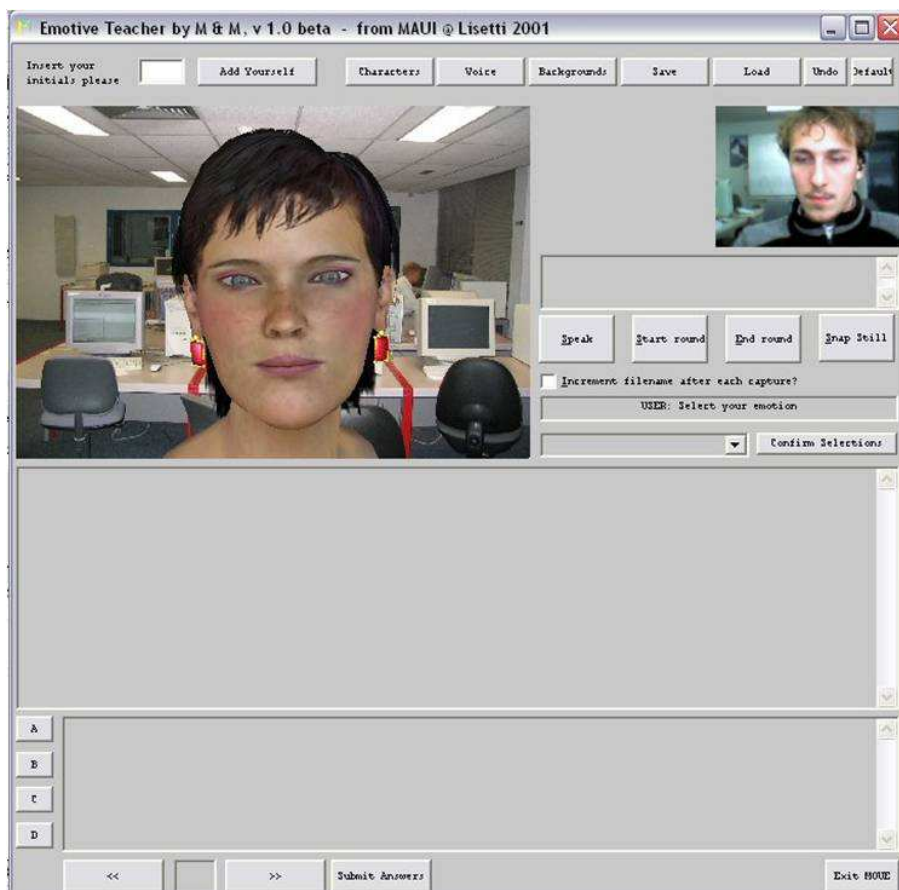


Figure 6.9: VALERIE Interface

justified by many psychologists (Frijda [1986], Picard [1997], Davidson et al. [2002] and many others).

The resulting architecture is really simple but it is able to simulate, either in an explicit way either with embedded code, a BDI system with some Beliefs relative to user average, and user effort, three Desires and three Intentions. Affective phenomena are represented in the form of emotions, mood, and personality. In Paleari et al. [2005] we present the results obtained by VALERIE when different personalities are applied.

In this chapter the main techniques and technologies for modeling and build intelligent affective agents have been reported. Next chapter presents ALICIA, the affective intelligent architecture we have developed.

### 6.3 Generic AI: ALICIA

ALICIA is a generic framework for modeling agents within the BDI architecture. We added Emotions creating the so called **BDI+E** architecture

The architecture of the project takes deep inspirations from the works from Gratch [2000a,b], Gratch and Marsella [2001, 2004], Marsella et al. [2003], Marsella and Gratch [2003a,b] for the artificial intelligence part, and from works from Lisetti and Nasoz [2002],

Lisetti and Gmytrasiewicz [2002] Scherer [1984, 2001], Leventhal [1979, 1984], Leventhal and Scherer [1987] for what it regards the general architecture and the basic psychological theories.

In the next sections we will describe the architecture in more details but essentially ALICIA is based on three layers: *reacting layer*, *behavioral layer* and *deliberative layer*.

**The reactive layer** is responsible for all those kind of reactions that are subconscious and limbic, as close eyes and turn head if faced to a strong light, or the desire of running if faced to a lion.

**The behavioral layer** is responsible for reactions that require no reasoning but that are nevertheless learned, as to respect rules. Examples can be to stop at the traffic light or to answer when asked. One person can usually choose not to follow reactions from the behaviors if he wants, but cannot usually choose not to execute one automatic reaction. In other words behavioral responses are the favorite and more common responses to specific stimuli, but not mandatory ones.

**The deliberative layer** is, on the contrary, responsible for the actual reasoning of the agent. Things like planning and long term decision making belong to that layer.

In the attempt to deploy a generic framework, Beliefs, Desires and Intentions are not hard coded but loaded as databases from external files. Beliefs are organized in term of a Dynamic Decision Network. A database of sample emotions is loaded too. This would be useful when trying to recognize an emotion given its appraisal.

### 6.3.1 BDI+E

ALICIA is structured in three layers: reactive, behavioral or schematic, and deliberative. There are two main reasons for this approach:

1. Firstly, building such a kind of architecture allows to better simulate the three layers model of emotion and cognition described by Scherer [2001] and Leventhal [1979].
2. Secondly, a three layer architecture allows to distribute different problems to one or the other layer of the architecture and, therefore, to subdivide the real world complexity into simpler problems according to the well known software engineering strategy of the "*divide and conquer*".

For this latter reason to be exploited the three layers of the architecture need to be mostly independent to each other and being assigned each a specific set of tasks and capabilities.

The general approach is based on the principle that in climbing the levels of the architecture from the reactive to the deliberative layer, the decision making process becomes less automatic and more cognitive, less limbic and more deliberative. The reasoning process therefore becomes more complex and takes in account more complex affective phenomena. Consequently, the architecture is developed to have a strong integration of emotions into each level of the reasoning process.

In the figure 6.10 ALICIA's architecture is shown; please note that emotions represent a layer superimposed on the third dimension. The emotional process has its effect at all layers as explained by Lisetti and Gmytrasiewicz [2002], and it cannot be separated easily in the three layers described before.

---

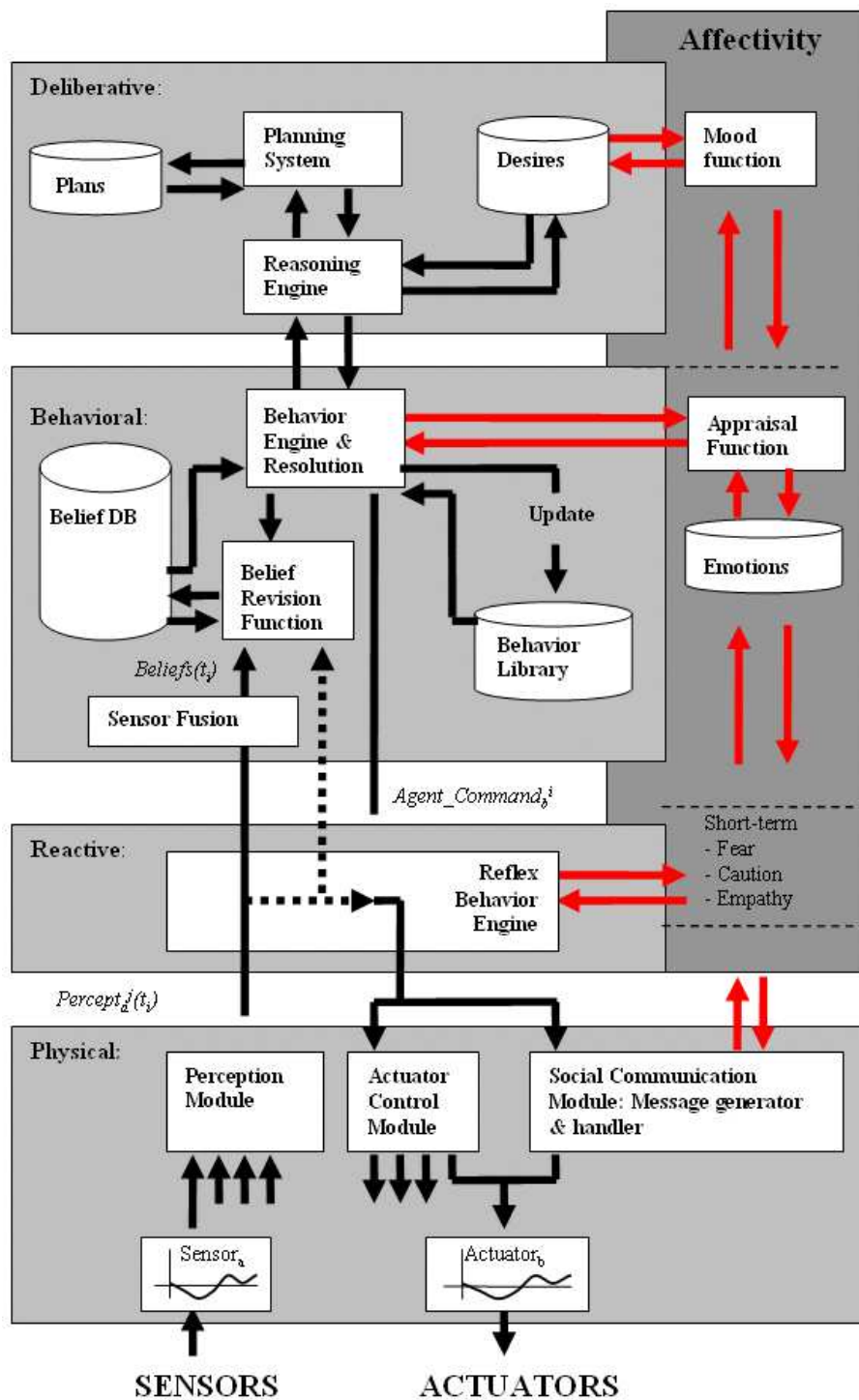


Figure 6.10: ALICIA architecture and its 3 layers

### 6.3.1.1 The reactive layer

The reactive level provides real-world robustness to unforeseen events where fast reflexive actions protect the system. It directly reacts to those incoming percepts that are considered "life threatening". At this layer only simple automatic and involuntary reactions to external stimuli are taken in account, like moving back if something big approaches (response to danger eliciting fear). Physical responses are also represented here; for example the shaking generated by a trill of pleasure generated by a caress could be implemented at that level rather than at one physical (fourth) layer. Basically, the reactive layer contains a list of rules that enable the agent to act rapidly without preliminary detailed consideration. To make the system the fastest possible, reactions are implemented through a hard coded list of beliefs to which the agent responds automatically with rules like:

```
if(recognize(belief, current_state))
    reaction(i);
```

This is similar to say that every human being has some specific kind of reactions when facing some specific stimuli which are in some way coded in our very DNA. For example, when faced a sudden and intense noise all of us will turn his head to face the source the noise, and when touching something hot the natural reaction would be to remove the hand and let the object go.

Usually the function `reaction(i)` shall contain calls to the higher levels to treat the event at behavioral and deliberative layers too. At the same time the function `reaction(i)` may invoke some kind of `eliciting(emotion_j)` function to elicit the right emotion to the system. For example an automatic reaction to the percept of a loud sound could elicit fear and so on.

**Affective Influence.** At the reactive layer the system has only basic affective mechanisms, with information of particular relevance and valence. The system is basically able to discern if an event is appraised as pleasant or not, if one is afraid of something or not, or if one is at its ease or not.

In addition, this layer would be influenced by the affective state of the agent. If the agent is in a good mood, or rather in a comfortable positive situation, a stimulus would be appraised in a very different way in comparison to the way it would be appraised if the agent is in a bad mood or rather in an uncomfortable situation.

*Example:* Even if one acts as a reflex - there's a different reaction to an approaching, flying object when it appears in a familiar environment among friends and therefore one is in good mood (that is, one tries to catch the object), then the reaction in an unknown environment with the feeling that something is wrong - then one would try, for example, to avoid the un-identified object.

### 6.3.1.2 Behavioral Level

The behavioral layer is committed to those learned reactions implying very limited or no reasoning that are still volitional to some extent. In other words it is responsible to induce those automatic reactions that are not merely simple signals, but that are encoded and recognized has objects, situations, events or people using sensing fusion strategies.

---

In this layer, the system is able to give responses to simple sets of beliefs, representing situations of the environment (i.e. learned schemata).

Examples of a behaviors could be the rules: if one sees a red traffic-light, then he or she has to stop or if somebody sneeze around us then we wish him/her "bless you". These kinds of behaviors necessitate recognizing a red traffic-light, but would not require the agent to think of the danger of not stopping at the traffic-light. In other words, to stop at a red traffic-light is a schema one learns while learning to drive and only imply a limited degree of reasoning and abstraction capabilities.

This layer is also responsible for storing belief about the physical and social environment representation. Beliefs are added, deleted, or modified by the Belief Revision Function (BRF). The BRF takes its inputs from the reflexive layer, as beliefs, and updates the beliefs database if the beliefs were recognized or adds new beliefs if the matter of the beliefs were unknown.

This level incorporates a reasoning engine where inbuilt behavior resolution mechanisms draw on information from the Belief database, the Behavior Library, and the medium-term emotional motivations.

**Affective Influence.** Emotional influence at this level would be relatively basic, such as mood and affect, but the layer itself would have the capacity to influence and generate emotions (as Lisetti and Gmytrasiewicz [2002] defines them) with specific contexts relating to single events, objects or people in the systems environment.

### 6.3.1.3 The deliberative layer

The third and last layer is the deliberative one. At the deliberative layer the agent is able to "think" about long term plans and actions with the objective of approaching its desires and avoiding its aversions. The idea is that, while the agent is not acting responding to schema from the behavioral layer or to rules of the reflexive layer, then it is deliberately reasoning at how to approach desires and avoid aversions (see figure 6.11).

In our architecture, Desires and Aversions are stored here in two different parts of the same database. The layer would also have the main reasoning engine that will be used by the system to make decisions according to its preferences and plans. At this purpose the layer does also include a module dedicate to plans generations and searches into a Plan database containing the history of recently learned/used/built plans. Possible actions the agent can perform are stored here in the form of the simplest possible plans.

The reasoning engine contains the information about the intention of the agent on the form of a pseudo plan of the next few actions of the agent. The reasoning engine would be committed to create, update and modify this "intended plan" according to the agent's desires, aversions and to the current state of the environment.

**Affective Influence.** At the deliberative layer emotions are complex and completely appraised. Although this layer is the more rational, in Descartes sense, it is here that emotions have the biggest and most subtle influence. At this level of consciousness emotions would be appraised looking at all pieces of information. For example if at the schematic, or behavioral layer, it can be enough to say an agent likes chocolate, at this layer that information should be mixed with the information, for example, that the agent don't want to get fat, or that there is eventually other people who wants the last chocolate cookie. The information about the appraised emotion will lead the decisions and the creation and



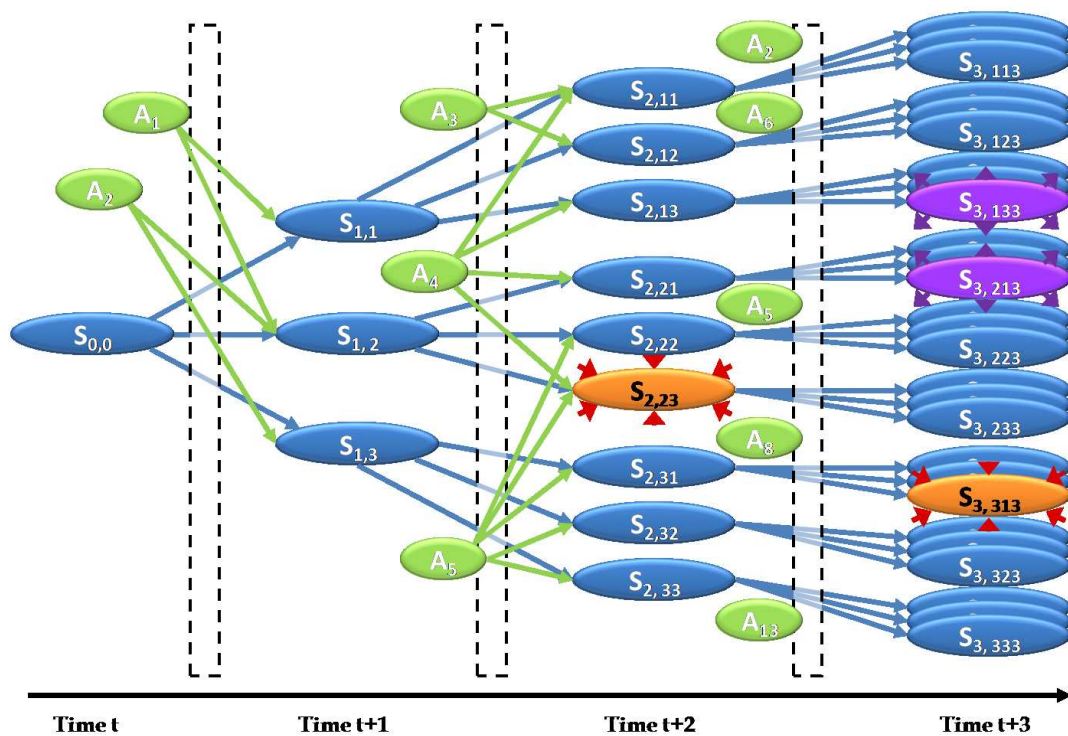


Figure 6.11: Exploring the DDN for desires. Desires are represented in orange, aversion in violet.

updating of the current intention. The current emotion influences the way the reasoning engine weights the different possibilities through the utility function.

#### 6.3.1.4 The physical layer

Below the reactive layer we defined a fourth level of the architecture. This level is the physical one. The physical layer depends on the specific platform. Indeed, the architecture (reactive, schematic, and deliberative layers) is independent from the platform, the sensors and the actuators.

The physical layer has the goal of outputting sensors information in a format compliant to ALICIA's representations and to convert ALICIA's intentions into actuators commands.

While the physical layer translates the electric signals coming from the sensors to percept for the reactive layer, it is a task of the behavioral layer, with its sensor fusion, to take the different percept and create beliefs about the environment.

This layer needs to contain a database with the capabilities of the agent, robot or computer, to be connected directly with the upper layers. In this way, the agent does only take decisions to make actions that are allowed by the platform. The action `move`, for example, would not be available for a desktop agent; at the contrary the action `read_CD` would probably not be available for an agent running on a robot.



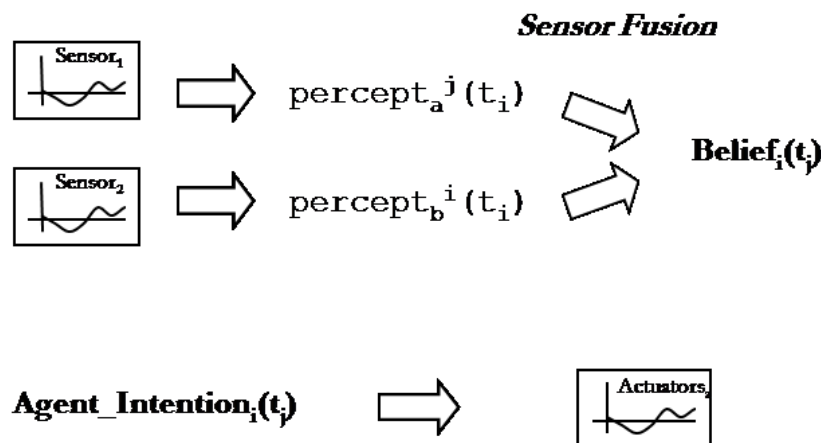


Figure 6.12: Sensor fusion and command translation

### 6.3.2 Modules descriptions

In this section we give some basic definitions about the structures included in ALICIA's architecture. In particular, we will define concepts such as Belief or Desire while giving few examples of the processing involving these entities in ALICIA.

**Belief** A belief is a structure used to represent all the information that the agent has referred to a particular object, person or social rule. Beliefs, therefore, constitute the systems knowledge of its physical and social environment, i.e. what the agent believes about his environment (e.g. `car_color = red`) or what the agent believes about a person (e.g. `user_name = Claudia`).

In ALICIA there are three main families of beliefs (see figure 6.13):

- beliefs about objects;
- beliefs about other agents;
- beliefs about social rules and conventions.

Inside the database, beliefs are organized in three different ways to make different behavior possible. Firstly, beliefs are organized as a cyclic double linked list ordered regarding their *goal relevance*, from the most relevant belief to the least one. In practice, beliefs belonging to desires or aversions, are more relevant than beliefs that are not associated and do not lead to desires. The search of beliefs would then start from the most relevant one and go to the least relevant one. Therefore, when a belief is found the agent can be sure that it is the most relevant one responding to the requisites of the search, and it can possibly stop the search because the following beliefs will, anyway, be less relevant.

Secondly, beliefs are organized amongst them by connection from parent to son, and vice versa or rather by connection as "*part of*" and "*contains*" designing a graph. For example `car0001` is "*parent of*" `wheels`, `doors`, `the engine0001` etc. while each of this part is "*son of*" the car. `Son_Of` and `Father_Of` are therefore used to situate a belief into some sort of objects' hierarchy.

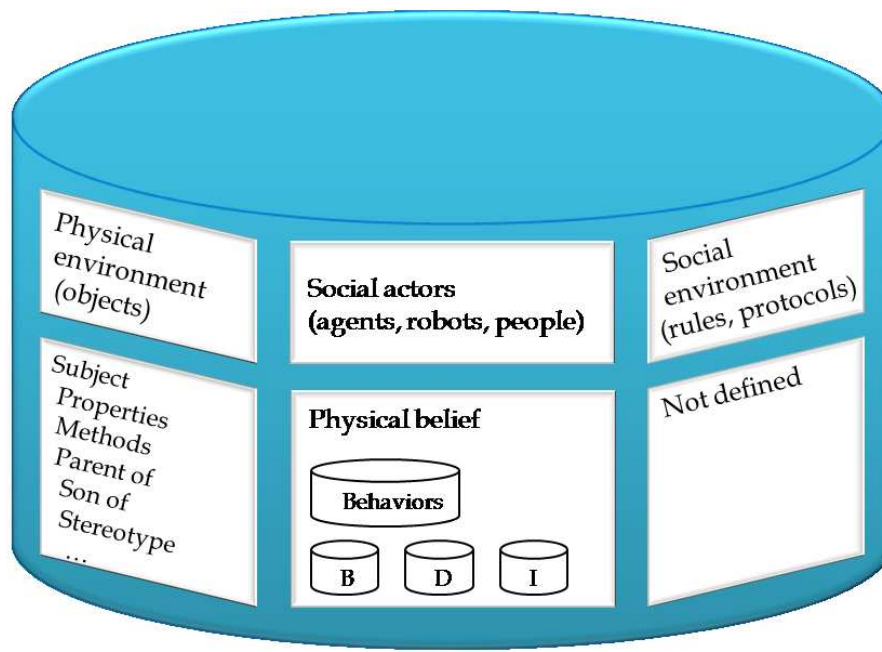


Figure 6.13: Belief Database structure

Thirdly, beliefs are connected in a semantic tree of relationships from the instances to their *stereotypes* or model in a hierarchy that goes from the instances up to their most abstracted version of it. For example *car0001* is a *car*, which is a *vehicle*, etc.

An object is represented as a set of attributes and capabilities. To this information it is added information necessary to place the belief into a hierarchy plus a link to a special stereotype-belief. For example Marco's car would be represented like in example 1. In this case the car has a color, an owner, a health, ... as well as some capabilities like "Go0001", an action that allow to move from a point  $x$  to a point  $y$ . The car also includes an engine, doors, and other parts and it is linked to the stereotype of *car*<sup>1</sup>.

**State** The environment is seen by the agents as a *State*. A State is a structure including one or more beliefs. In general the environment–state could include a very large number of beliefs.

**Behavior** The Behavior database stores information about rules defining in a formal way how the agent reacts to certain belief sets. Herein, the internal standards of the agent, and behavior rules that are either given a-priori or self-generated can be found.

The behavior database is updated to enable the agent to act more quickly in situations that happen regularly. This is one of the commitments of the behavioral engine. The other one is to find in the behavior database, the most relevant and urgent behavior available (if any) for the current situation. Practically, the agents search in the beliefs database for the presence of the initial eliciting state of each behavior.

<sup>1</sup>Please note that since Marco's car is linked to the stereotype of car we may not need to actually store all the car parts. In this example we have reported them for the sake of the example pretending that each one of these part have something peculiar the agent knows about.

### Example 1: A possible representation of the belief about Marco's car

```

Subject: car0001          -> an alphanumeric label (CString)

Attributes:

  Name: color             -> a label (CString)
  Value: 000.255.000     -> in this case RGB (long int)
  Precision: 010.010.010 -> the confidence in the measure (long int)
  Strvalue: RGB          -> in this case explains the format
  Certainty: 99%         -> likelihood of the information (char)
  Changeability: V       -> Automatic, Voluntary, Unchangeable (char)
  Time: 16/06/2007 12:34:23 -> timestamps (time / long int)

  Name: owner
  Value: 1               -> the ID of the model (long int)
  Precision: 0
  Strvalue: Marco        -> the name of the owner
  Certainty: 99%
  Changeability: V
  Time: 16/06/2007 12:34:24

  Name: health
  Value: 90              -> how much the car works (long int)
  Precision: 5
  Strvalue: 0-100        -> the range for the value
  Certainty: 70%
  Changeability: A
  Time: 16/06/2007 12:34:25
  ...

Capabilities:

  Name: Go0001           -> an alphanumeric label (CString)
  Certainty: 70%         -> likelihood of the information (char)
  Action: ..move(x, y).. -> are the lines of code to be executed (CString)
  Initial state: 1       -> ID of the initial set of beliefs (prerequisites)
  Final state: 2         -> ID of the final set of beliefs (final condition)

  Name: Take_gas
  ...

Certainty: 99%          -> likelihood of the existence of the car (char)

Son Of: none            -> the car is not part of something

Father Of:              -> links to the beliefs representing the car
  Engine00001
  Door00001
  Door00002
  ...

Stereotype: car         -> link to the belief of the stereotyped car
Goal relevance: 80      -> updated each time a desire is added or changed
Active: 0               -> whether the object is active (1) or passive (0)

```

Behaviors are implemented at the Behavioral or Schematic layer to define, for a given situation, standard or default reactions that the agent has learned (learned reactions). There is no cognitive reasoning in doing that; simply, while taking the driving license one has learned the rule of stopping when driving to a red traffic light.

To store that kind of information the model includes a link to a state and to the relative behavioral action plus some other information about the affective state elicited from that situation.

For example the behavior for the red traffic light would be like represented in a way similar to the one showed in the Example 2. As one can see a behavior is implemented as a state (i.e. a set of beliefs), a favorite action to perform if the case the state was recognized as belonging to the current environment state, and a list of possible consequences (i.e. a list of possible future states) listing what could happen if the behavior was not followed.

Behaviors are organized as a doubled linked cyclic list sorted according to relevance and urgency.

## Example 2: Possible representation of the “red traffic light” behavior

```

Relevance: 80                                -> the relevance of the behavior (0-100 char)

State:                                         -> a set of beliefs
  Belief 1:                                    -> the first belief of the set
    Subject: Traffic_light001
    Attributes:
      Name: light_color
      Value: 255.000.000
      Precision: 010.000.000
      ...
      Name: position
      Value: 010.050.005
      ...
  Belief 2:                                    -> the first belief of the set
    Subject: Me
    Attributes:
      Name: speed
      Value: 110
      Precision: 10
      ...
      Name: direction
      Value: 010.050.005
      Precision: 001.001.001
      ...
  ...

Action: stop_at(traffic_light001) -> pointer to the action stop_at(X)

Urgency: 1                                    -> the value of available seconds or cycles (int)

Consequences:                                  -> a list of consequences, not yet used
  BeliefXX
    Subject: Driver0001
    Attributes:
      Name: alive
      Value: 0
      Precision: 0
      Certainty: 20%
      ...
  ...
  ...

```

**Desires and Aversions** Desires represent the belief sets or states that one wants to reach. There can be several desires at the same time, even some with little correlation between them. An example mentioned in Scherer [2001] would be, one can have the desire to eat a piece of chocolate cake and at the same time have the desire to lose weight, two desires that in fact counteract each other.

Aversions are the states one wants to avoid. Desires and Aversions are stored in the deliberative layer of the architecture and used by the deliberative engine to plan future actions. In the phase of generating new behaviors Desires and Aversions are used when adding the information of relevance and urgency. For example, the fact that the “red traffic light” rule is more relevant than a “answer to friends” one depends on the fact that not stopping at the traffic light would imply death, or serious injury or possibly the destruction of the car. Such very strong undesirable states override the consequences of not answering a friend which are, with respect to these, quite acceptable.

The representation for both aversion and desires are the same but they are stored in two separate parts of the same database. The reason for this is to have fast and simple mechanisms to search desires or aversions separately.

While one is in a good mood, he/she usually easily perceive good news and find it hard to see bad ones (Lisetti and Gmytrasiewicz [2002], Picard [1997], Frijda [1986]). By separating the two databases, this mechanism is easily implemented: when in a good mood the agent can, for example, search for desired states 80% of time and for the other 20% of time search for undesired states. This mechanism does allow the agent to be more sensitive to good news and only prioritize very relevant bad ones or vice versa

Furthermore looking at the previewed possible future states it would be easier to mark states as desirable or undesirable and therefore facilitate a fast appraisal of the future possible states without needing to look into the desires/aversions structure to read the values. In fact, for example, future states presenting only undesired states are less desirable than future states presenting only desired states (see figure 6.11).

All the mechanisms described above base their working on the idea that desires and aversions are ordered inside by relevance and are differentiated from each other through a valence value (for desires is positive, for aversions is negative). Also in this case the organization for desires and aversion would be the one of a double linked cyclic list.

In the example 3 we report an example for the desire “eat chocolate cake”. The desire is simply a state (in this case a chocolate cake) with the beliefs attributes (in this case the cake has been eaten) and a partial description of the emotional appraisal related to that state.

Example 3: A possible representation of the “chocolate cake” desire

```
%State:
  Belief1
    Subject: chocolate cake  -> in this case is the stereotype

  Attributes:
    Subject: eaten
    Value: 1
    ...
    Subject: eater
    Value: 0001             -> the ID of the Agent
    ...

  Relevance: 20            -> 0 - 100 (char)
  Urgency: 5              -> 0 - 100 (char)
  Valence: +              -> + (char)
  Intrinsic/Extrinsic: I  -> I - E (char)
  Duration: 90            -> (int)
  Modifiability: 40       -> (int)
  Likelihood: 40         -> if computed is the likelihood of
                          reaching this specific desire
```

**Plans** A plan is a sequence of Actions. An action is the application of a belief method which lead from an initial state I to a final state F.

For each belief in the final state F a likelihood is set so that the action `give_morphine` can in 90% of the case reduce the patient pain and in 15% of the cases hasten the patient death. Please note that the likelihood does not have to sum up to one. Every single output is potentially independent from the others

The planning database contains the more common plans the agent can use. The main function of the planning system is to search the Desires DB for the most important desires, pass them to the plan database to see, if and which plans are available to reach those desires. Once this information is returned, it is the job of the planning system to find the best plan to follow and update the current Intention.

For example if the agents wants to go to Paris then it can evaluate the choices of taking the train, the plane, the car, the motorbike or other means of transport. For each transportation available the agent can create a plan<sup>2</sup>. All of these plans would start from

<sup>2</sup>Please note that to follow the plan each action contained in it needs to be executed; plans will, therefore, be only composed of the capabilities of the physical agent.

## Example 4: A possible representation for the plan “go to London”

```

Plan Initial state:
  Belief1
    Subject: Actor X
    Attributes:
      Subject: position
      Value: Y
      ...
    Subject: ID
    Value: N
    ...
  Belief2
    Subject: Money00001
    Attributes:
      Subject: quantity
      Value: 1.000.000
      ...
    Subject: belongs to
    Value: N
    ...
    -> the Id of ActorX

Actions
  Action1:
    Initial State:
      Belief1:
        ...
    Final State:
      ...

  Action: reserve plane ticket(Z)
    Initial State:
      ...
    Final State
      Beliefn:
        ...
    -> the link to the action reserve ticket

  Action: prepare suitcase
    Initial State:
      ...
    Final State
      Beliefn:
        ...
    ...

Plan Final State
  Belief1
    Subject: Actor X
    Attributes:
      Subject: position
      Value: Z
      ...
    Subject: ID
    Value: N
    ...
    -> note: position is changed
  Belief2
    Subject: Money00001
    Attributes:
      Subject: quantity
      Value: 900.000
      ...
    Subject: belongs to
    Value: N
    ...
    -> note: quantity is changed
    -> the ID of ActorX

  Time: 100
  Fatigue: 10
  Money: 100.000
  Somatic marker: 0
    -> the various costs the agent has to afford
    -> to pursue this plane

```

the initial state with the agent located in “location A” (for example Sophia Antipolis) and end with a final state where the agent is in “location B” (for example Paris). Each plan is coupled with some additional information which tells the agent how much does the plan cost in term of money, time and fatigue. The parameters of cost in term of money, time and fatigue would help to assess rapidly whether the plan is acceptable or not. For example the agent may not consider the action of going by feet 700 Km far spending a couple of month walking nor to take a private jet spending 50.000 euro for the flight.

As it can be seen from example 4, plans are basically represented by an Initial State, a list of Actions and a Final State.



Simply looking at the initial and final state and making the difference between the two one should be able to understand what the plan do. To know exactly how that is done the agent needs to evaluate each action initial and final states.

Evaluate a plan would be a question of recognizing desires and aversions into the initial and final states of each action belonging to the plan itself.

Therefore, to make a complete evaluation of a plan, the agent need evaluate all the history of states reached at each step.

**Intentions** The reasoning engine at the deliberative layer would also contain the `Intention` of the agent. An intention is represented in this system with the same model of plans and therefore is a list of actions the agent would like to do the following time steps.

The intention is updated each time the engine does some computation. In particular at every time the engine asks the planning engine for plans to the agent most relevant and urgent desires. If something is found, then the plan is compared with the current intention. If the two are not equals or do not leads to the same final state the intention can eventually be updated, modified or created as new.

When the deliberative function is ended and the intention is set, then the agent can eventually send the first action of the list through the architecture to be finally translated into actuators by the physical layer.

**Emotions and moods** The model used for the emotions comes from the Scherer [2001] and the Lisetti and Gmytrasiewicz [2002] models. The architecture implements emotions, moods and personalities. At the lower level of the architecture, the emotion influence would rather be the one from the mood, at the higher the one from emotions. Anyway at all layers the affective phenomena elicited are emotions. Emotions are appraised by dedicated functions from a state (current, future or past, true or simulated etc.). Mood are then computed by "averaging" all the appraised emotions and by the mood history.

The representation of mood is composed of two appraisal variables, namely intensity and valence (see example 5)

Example 5: An example of mood

Intensity: 80	-> from 0 to 100 (char)
Valence: 1	-> 1 ' positive or -1 ' negative (char)

A representation of one emotions, which is derived from the works of Scherer [2001] and Lisetti and Gmytrasiewicz [2002], can be seen in example 6. In this case we are detailing an emotion of "happiness" which is due to the result of a certain action we do not know by agent 10; the emotion is legitimate and comply with all of the agent internal and social norms: for example the other agent could have offered a gift to us.

The process of appraisal of a state looks the state for the most relevant active beliefs (non active or non goal-relevant beliefs are not considered). Once it has selected those beliefs it simulates all the possible combination of actions the agents in the current state can perform (in the future it will simulate only the goal-relevant actions) and computes that way the possible future states (see also Rao and Georgeff [1991, 1995]). Once all possible, and considered relevant, states are computed, desires and aversions are searched inside these possible future states. This search would lead to a list of desires and aversions for the current and for the future states.

### Example 6: An example of Emotion

```

Intensity: 80          -> from 0 to 100 (char)
Valence: 1            -> 1 ' positive or -1 ' negative (char)
Name: happiness
Facial Expression: happy
Duration: 32          -> from 0 to 100 (char)
Focality_controll: 'A' -> (char) used to identify one from the next
                        'P' ' physical, 'A' ' actor, 'R' ' rule or 'N' ' none
- Social Focality: 10 -> on who the emotion is focused on
- Social Environment Focality: NULL
- Physical Focality: NULL
- Event Focality: NULL

Agency: 10           -> the ID of the actor who is responsible for the emotion
Novelty: 70           -> if the emotion was expected

Intentionality: 1     -> if the triggering event was intentional
                        (1 ' yes, 0 ' no)
Controllability: 10   -> from 0 to 100 how much the agent can cope with the event
                        causing the emotion
Modifiability: 10     -> from 0 to 100 referring to the time necessary to cope
Certainty: 60         -> referred to the anticipation effect to come
Legitimacy: 1         -> if the emotion is felt as legitimate (1 ' yes, 0 ' no)
External Norm: 100    -> if the triggering event is acceptable from the others
Internal Standard: 100 -> if the triggering event is acceptable for oneself
Action Tendency: 1    -> the link to the most appropriate strategy
Causal Chain:         -> the history of states that caused the triggering event

```

From the differences between the two lists the system computes the basic parameters of the appraised emotion. Then, all the appraised emotions are “averaged” to characterize the average emotion one state and the future ones elicit.

It is interesting to note that all of these models apply to both the agents and the users. Indeed the representation of the self is, in ALICIA, the same representation used for the others (agents or users).

#### 6.3.2.1 Jimmy and Dr. Tom Example

This example is taken from Gratch and Marsella [2004] paper. The scenario consists of three actors, the little Jimmy, a 10 years old boy suffering from cancer in its final stadium, Dr. Tom, who has to decide about giving or not some morphine to his suffering patient, and Jimmy’s mother.

Let the system simulate Dr. Tom thoughts. Dr. Tom beliefs would be:

1. Jimmy, the 10 years old boy suffering from cancer
2. Jimmy’s mother
3. Morphine, with the two possible actions (if given to someone)
  - (a) hasten Jimmy’s death that leads to the final state where Jimmy is dead
  - (b) end Jimmy’s suffering that leads to the final state with Jimmy not suffering
4. Dr. Tom itself with the actions:

- (a) administrate morphine
- (b) do nothing

#### 5. other non relevant beliefs

In a similar way its desires, as defined by Gratch and Marsella would be:

1. End people suffering (a state containing a person not suffering)  
relevance = 20 valence = 1
2. Hasten death (a state containing a dead person)  
relevance = 100 valence = -1
3. other non relevant desires

Morphine can, in the idea of the agent, lead with the 20% to death hasten and with 90% to end Jimmy's suffering.

If Dr. Tom will simulate the action of giving morphine the resulting state will contain the belief morphine which will be appraised looking at the possible future states. There are two possible future states; the first one contains the belief of Jimmy dead and the second containing the belief of Jimmy not suffering.

Looking at the first state the agent will find the occurrence of the aversion "Hasten death" with relevance 100 and valence -1, which will lead to sadness and in general to a negative emotion. Looking at the second state the agent will find the occurrence of the desire "End suffering" with relevance 20 and valence 1, which will lead to happiness and in general to a positive emotion. At the same time the probability of hastening Jimmy's death by giving of the morphine is very low (10%) while the likelihood of ending Jimmy's suffering is very high (95%).

The emotion will be then computed as a difference of current emotion (neither positive nor negative as no desires or aversions are present) and therefore leads to respectively fear and hope. The emotion felt appraising the action "administrate morphine" would then be a mixture of fear and hope. The aversion "hasten death" is, in the current setup, much more relevant than the desire "end suffering" but when weighting the resulting emotions by the likelihoods of the output then the emotion linked to the drawback of administrate morphine would be made less intense. The overall emotion would be the one of hope.

If the agent will have to choose if to give morphine to Jimmy or not it would then choose to administrate the patient with that drug. If the agent will be obliged not to give morphine as Jimmy's mother decision then the emotion elicited would be of anger as it will be obliged to do something that Dr. Tom doesn't want to do.

If morphine would work out Jimmy's pain then Dr. Tom would feel relieved from the fear of hastening his death and happy for the attainment of one of his desires. At the contrary, if morphine would turn out to hasten Jimmy's death then Dr. Tom would feel sad and ashamed for the result of his action.

### 6.3.3 Demo Implementation

We have developed a proof-of-concept demo version of ALICIA. The deployed system only implements part of the architecture. In particular these parts have been developed:

- a simple physical layer (i.e. the interface)

- a simple reactive layer, to basically show how the reactive layer would work,
- the Belief Revision Function,
- the Belief database (physical beliefs only),
- the Behavior engine and Behavior database,
- the Desires and Aversions databases,
- the Emotion database,
- the Emotion engine,
- the Plan database (but no plan has been designed)
- some simple mechanisms of the deliberative reasoning engine.

We have created an interface (shown in figure 6.14) which allows to load all the various databases<sup>3</sup> and to send new beliefs to the architecture.

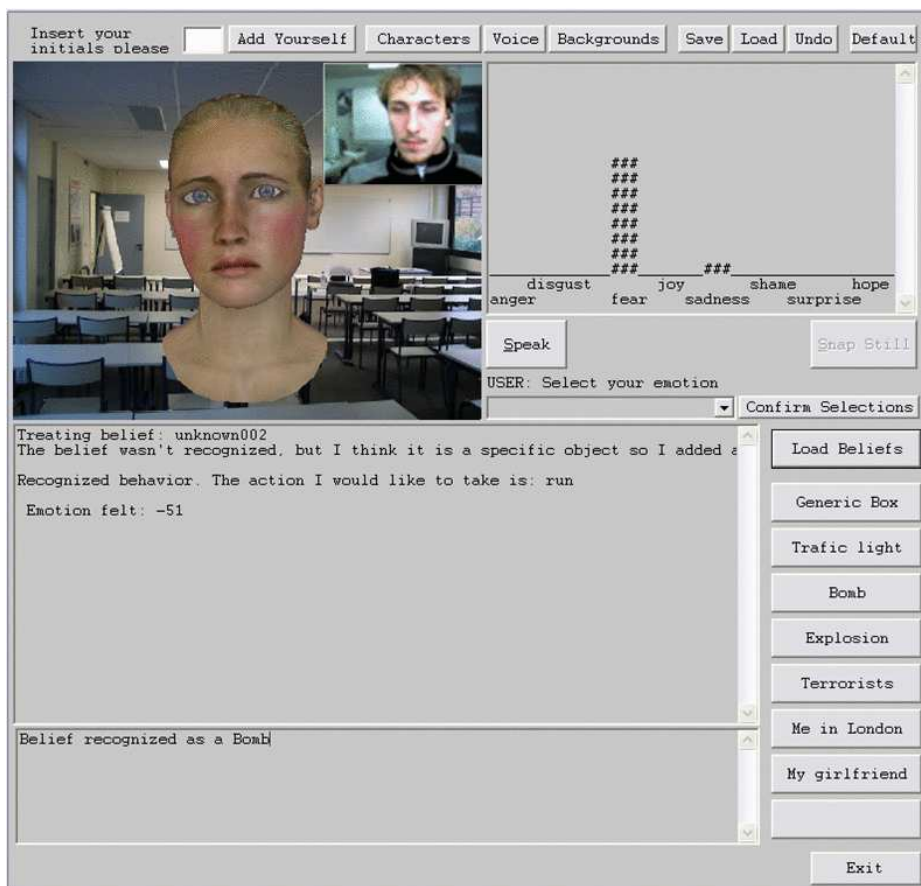


Figure 6.14: Interface for the demo

<sup>3</sup>databases contain beliefs, desires, intentions, behaviors, plans and a database containing examples of stereotype emotions of the agent and are stored in text files.

The panel at the top right corner represent the emotion felt by the agent as a mixture of 8 basic emotions (see figure 6.15): anger, disgust, fear, joy, sadness, shame, surprise, hope. The graph has, therefore, 8 bars representing the emotion as a mixture of those 8 basic emotions. As it has been said before, appraisal is computed looking at all possible future states. For each future state an emotion is computed then, each emotion is weighted by its relevance or intensity and mixed in the graph.

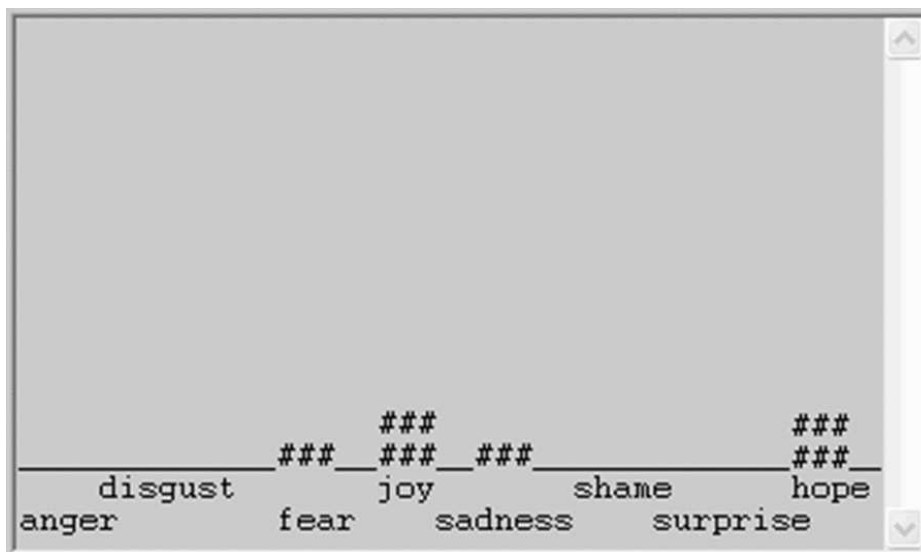


Figure 6.15: The top right panel representing emotions

We included in the interface an emotional avatar which is able to display facial expressions as counterpart of the appraised emotional responses and to communicate with the user.

In figure 6.14 we show the resulting interface.

The top buttons are used to load and save favorite avatars, to change background or avatar's face or voice. The "Add Yourself" button is used to add the user to the list of current users. A webcam film the scene in front of the computer and let the agent, through third part software called Visionics, perform face recognition on the user. If the user is recognized then the system can load his preferences in term of agent, voice and background, if it is not then the system asks the user to add his/her name. Face recognition is actually triggered by the "Snap Still" button on the top right quarter of the interface.

The middle and bigger panel is now used by the agent to prompt text and output reactions to the user but it is thought for generic purpose text.

The lower panel is currently used for prompting debug information and to send commands to ALICIA. When writing text to the lower panel and pushing the text speak the agent simulates having recognized the user speech. Some sentences (small talk like) are recognized thanks to a simple text parser and ALICIA respond triggering behavior.

Thanks to a simple text parser it is possible to send with the same input new beliefs to the reactive layer from the physical one (i.e. the interface itself). The text "`--command=`" will trigger this behavior; the following text will be treated as a command to the interface. Typical commands would be like "load agent agent\_name", "load scenario scenario\_ID" or "add belief belief\_ID". Some buttons allow, for simplicity

and speed, to send particular beliefs that we used frequently during the testing phase. Beliefs are automatically treated, revised, stored and evaluated at the three different layers of the architecture.

### 6.3.3.1 Proof of concept scenarios

We have developed a demonstrator to test some relatively simple scenarios with ALLCIA's architecture. The architecture was tested with several scenarios and worked as expected.

Some simple scenarios are described which show the various competences of the different levels of the architecture.

**A Simple Unknown Box** . Let us consider, as a first scenario, the case where the agent sensors recognize the presence of a simple box near the agent. The belief which will be formed will represent the box according to its dimensions, position and color. Other data can only be computed afterward. At the reactive layer, the belief does not activate an automatic reaction.

The belief is therefore sent to the schematic layer where it is treated by the belief revision function. Here the new belief will be compared with the existing ones; if the information contained in the belief is considered too generic then generic beliefs will be recognized with maximum likelihood and the belief will be temporarily stored, awaiting further information. If not, then the belief can be recognized as a specific existing object or as a new occurrence of a stereotype.

It is not expected that specific behavior will be initiated by this simple box and therefore the belief will be sent to the conceptual layer which will continue by searching for relevant plans. It is not likely that the agent would need a generic box for its plans. In this case the agent will continue reasoning as if nothing did happen

**Reactive layer.** To test the kind of reactions that can be triggered at the reactive layer we have implemented a scenario in which the agent is faced to an explosion. The explosion is characterized by a sudden and unexpected appearance of a very big source of noise, heat, and light. At the reactive layer we hard-coded the rule:

```
if isIncreased(sound_level, 60dB)
then
  elicit(fear);
  set(fight_or_flight, 20, 80);
```

The agent will, therefore, be scared by the sudden noise and increase the preference for a flight strategy to 80.

**Schematic layer.** To test the kind of behavior that can be triggered at the schematic layer we implemented a scenario in which the agent is driving and is faced to a red traffic light. In this case the agent recognize the traffic light (facing at him and at a predetermined distance) and initiates the actions `break` and `stop_at_traffic_light`.

---

**Deliberative layer.** We implemented a third scenario to test the behavior of the agent when faced to cognitive decision making and appraisal of emotions. For simulating different emotional appraisals we have set the scenario in London at the time of the terroristic attacks and asked ALICIA to react to four different sub-scenarios. In this case ALICIA implements a male character we call Bob.

In the first scenario Bob is comfortably at home in France and learns from the news that there have been terroristic attacks in London. In the second scenario Bob is in London and learns from the news that there have been terroristic attacks in London. In the third scenario Bob is home, but Alex, Bob's girlfriend is in London when he learns from the news that there have been terroristic attacks in London. In the fourth and last scenario Bob is in London and he is faced to a bomb.

Thanks to the appraisal mechanism defined by Scherer [2001] and implemented by us the agent appraises the four situations in very different ways.

In the first case, nor Bob nor anyone Bob knows is in danger; the event is evaluated as novel, unpleasant, but not goal-obstructive and Bob also evaluates that the event could have been avoided (power and control) by the terrorists and English government. The reaction of Bob would be a mild, mitigated anger toward the government and the terrorists.

In the second case, Bob would be directly in mild danger of life; the event evaluation change because death is for Bob an Aversion. Bob will evaluate the event as a novel, goal-obstructive one; he would believe he could not control it, over a certain extent (for example he could leave London or hide in a secure place), and blame the government and the terrorists for the death. Bob will be at the same time angry with the government, and scared for himself.

In the third case, Alex, Bob's girlfriend, would be in direct danger of life; Bob doesn't even know whether she is still alive or not. Bob will therefore be mostly scared for Alex because he knows he cannot control, in any way, the situation.

Finally, in the fourth scenario Bob's life is in direct danger. The reaction to the event would be the one of terror.

## 6.4 Concluding Remarks

In this part we have overviewed a framework for the creation of agents with emotional capabilities and we have shown how they can be used to simulate the appraisal process of emotions.

The resulting architecture take deep inspiration from Leventhal [1979], Scherer [2001], Lisetti and Gmytrasiewicz [2002], and Lisetti and Nasoz [2002] works. In doing this three levels of human mind are simulated, namely:

1. reactive
2. schematic
3. deliberative

We have defined a paradigm for building affective agents, namely BDI+E which mixes the Rao and Georgeff [1991, 1995] BDI architecture with emotions. The Scherer [2001] component process theory of emotions has been applied and the system is able to

---



fully simulate appraisal of surrounding events, elicit emotions, simulate affective phenomena, and simulate the influence of affective states on cognitive functions.

Some novel paradigm have been implemented such as the separation of desires in two databases for desires and aversions leading to a possible algorithm to simulate the effect of mood on perception and evaluation of the surrounding events.

Albeit personality was not used in the demo version of this architecture, different personalities are simulated through different setup of the agent's databases. Furthermore a model of personality has already been developed which is based on the Big Five Factor model (Norman [1963]) of personality and which shall be explicitly used in the future. We suggest that this model other than being effective, can, thanks to its emotional behavior, react in a more natural and human like way.

The model is task independent: it can work in tutoring as well as in domotics, gaming or others environments. The architecture is also platform independent, yet the platform must implement the interfaces we defined between the physical and the reactive layer.

Preliminary experiments on input systems for an intelligent agent have shown the importance of adopting an integrative approach when dealing with different sensor modalities. At a first (reactive) level, we may fuse highly correlated signals (e.g. audio signals coming from different microphones which are coupled to increase signal to noise ratio). At the second (schematic) level, features are extracted and form beliefs which can be fused together by the BRF (e.g. beliefs about recognized phonemes and visemes to increase speech recognition). Finally at the third (conceptual) level pieces of information may be used together to make a decision (e.g. in a domotic environment, information coming from cameras and audio (which detect the position of the user) the status of light switches and the time of the day to plan whether to turn on lights or not.

---

## Chapter 7

# Summary and Conclusion

In this thesis we have described the topic of affective computing. Computers are spreading more and more worldwide and gaining computational power but they still fail to create natural human–computer interactions and to operate intelligently in un–constrained environment. We have overviewed psychological theories suggesting that a possible solution to these issues may be represented by emotional capabilities.

There exist three main families of emotional capabilities: those related to the display and communication of affective states, those connected to the ability of recognizing emotions, and, finally, those linked to the ability to process emotions, think “emotionally”, and to the ultimate extent, “feel” or simulate emotions.

In the three main parts of this thesis we have presented the theories and the technologies which have been proposed in the state of the art and described how we did improve them. In particular our contributions are:

- **Emotion Display:**

1. *Generation of believable, psychologically based, facial expressions.*  
We have generated facial expressions for two different platforms: *Cherry, an avatar by Haptik* and *Cleo, an iCat robot by Philips*. We have followed the guidelines traced by Scherer [2001] component process theory of emotions thus validating the theory. We have demonstrated that the developed facial expressions are both understandable and believable.
2. *Collaboration with psychologists for the definition of the theories of emotional display.*  
While generating the facial expressions we have highlighted few small lacunas in Scherer [2001] theory and we had the chance to discuss these gaps with the psychologists.

- **Emotion Recognition:**

3. *SAMMI: Semantic Affect-enhanced MultiMedia Indexing.*  
We have proposed a framework for the indexing of multimedia excerpts which takes into account emotions among other semantic tags. We have showed that such an approach could solve the issue of the semantic gap and simplify some of the most complex indexing and retrieval tasks.
  4. *AMMAF: Automatic Multilevel and Multimodal Affect cues Fusion paradigm.*  
We have discussed the fact that a multimodal approach increases the chances
-

of correctly indexing emotions. We have, then, proposed AMMAF as a complete framework for the fusion of multimodal signals, features, and emotional decisions. AMMAF makes use of synchronization buffers to build more reliable and precise emotion estimates. AMMAF takes into account different affective phenomena and, in particular, it allows to estimate emotions through the emotional description defined by Scherer [2001], moods defined as an unfocused long-term emotion (Lisetti and Gmytrasiewicz [2002]), and attitudes or personality traits in the sense of longer-term unfocused affective phenomena.

5. *Extensive studies on emotion recognition.*

We have presented some extensive studies on feature extraction, feature selection, feature vector generation, pre and post-processing, classification, and multimodal fusion.

6. *ARAVER: Automatic Real-time Audio-Video Emotion Recognition.*

We have proposed and implemented a system for automatic recognition of emotions exploiting human facial expressions and vocal prosody. The system makes use of an automatic and low-complexity facial feature point tracker which we have proposed. ARAVER combines video and audio estimates with a Bayesian multimodal approach and applies simple pre and post-processing techniques. The results of our studies show that ARAVER obtains recognition rates ranging from 50% to more than 90% depending on the particular settings.

• **Emotion Synthesis:**

7. *ALICIA: Affective Intelligent Agent Architecture.*

We have proposed a framework for the development of artificial intelligent agents capable of taking into account emotions as primary source of decision making. ALICIA updates the existing BDI technology to include emotions therefore defining a BDI+E framework. The framework implements the Scherer [2001] process of appraisal and make decisions which are influenced by its emotions. Emotions have, in the proposed framework, influences at many levels of cognition as suggested by Lisetti and Gmytrasiewicz [2002]. We have implemented a demo version of ALICIA which demonstrates to behave in a natural human way in some dedicated scenarios.

• **Other contributions:**

8. *Biometric People Recognition through Facial Expressions Dynamics.*

We propose a system for biometric people recognition based on the way people express their emotions through facial expressions. We have demonstrated that dynamics of the facial expressions are a source of biometric information. We propose a system which uses biometric information to improve the score of emotion recognition and vice versa.

9. *Automatic Music Transcription.* We have developed a system for the automatic audio-video transcription of guitar music. The system processes the video of a guitar player and returns in real-time a set of possible played notes according to the position of the two hands on the guitar. Using audio to detect attacks and the fundamental frequency of the notes being played, the system is able to extract reliable estimates of the guitar tablature being played.

---

---

The work presented here led to 15 publications in workshops and international conferences, to a best student paper award, and a best poster award.

### Scientific Publications Derived from this Thesis.

1. M. Paleari, C. L. Lisetti, and M. Lethonen. VALERIE: a virtual agent for a learning environment, reacting and interacting emotionally. In *AIED 2005, 12th International Conference on Artificial Intelligence in Education, July 18-22, 2005, Amsterdam, The Netherlands*, July 2005.
  2. A. Grizard, M. Paleari, and C. L. Lisetti. Adaptation d'une théorie psychologique pour la génération d'expressions faciales synthétiques pour des agents d'interface. In *WACA 2006, 2eme Workshop sur les Agents Conversationnels Animés, 26-27 octobre 2006, Toulouse, France*, October 2006.
  3. M. Paleari and C. L. Lisetti. Psychologically grounded avatars expressions. In *1st Workshop on Emotion and Computing at KI 2006, 29th Annual Conference on Artificial Intelligence, June, 14-19, 2006, Bremen, Germany*, June 2006a.
  4. M. Paleari and C. L. Lisetti. Avatar expressions with Scherer's theory. In *3rd HUMAINE EU Summer School, September 22-28, 2006, Genova, Italy*, October 2006b.
  5. M. Paleari and C. L. Lisetti. Toward multimodal fusion of affective cues. In *HCM 2006, 1st International Workshop in Human Centered Multimedia at ACM Multimedia 2006, October 23-27, 2006, Santa Barbara, USA*, October 2006c.
  6. M. Paleari and C. L. Lisetti. Agents for learning environments. In *Demo at 1st Workshop on Emotion and Computing at KI 2006, 29th Annual Conference on Artificial Intelligence, June, 14-19, 2006, Bremen, Germany*, June 2006d.
  7. M. Paleari, B. Duffy, and B. Huet. Using emotions to tag media. In *Jamboree 2007: Workshop By and For KSpace PhD Students, September, 15th 2007, Berlin, Germany (Best Poster Award)*, September 2007a.
  8. M. Paleari, A. Grizard, and C. L. Lisetti. Adapting psychologically grounded facial emotional expressions to different anthropomorphic embodiment platforms. In *FLAIRS 2007, 20th International Florida Artificial Intelligence Research Society Conference, May 7-9, 2007, Key West, USA*, May 2007b.
  9. M. Paleari, B. Huet, and B. Duffy. SAMMI, Semantic affect-enhanced multimedia indexing. In *SAMT 2007, 2nd International Conference on Semantic and Digital Media Technologies, 5-7 December 2007, Genoa, Italy* | Also published as *Lecture Notes in Computer Science Volume 4816*, December 2007c.
  10. M. Paleari and B. Huet. Toward emotion indexing of multimedia excerpts. In *CBMI 2008, 6th International Workshop on Content Based Multimedia Indexing, June, 18-20th 2008, London, UK. (Best Student Paper)*, June 2008.
  11. M. Paleari, B. Huet, A. Schutz, and D. T. M. Slock. Audio-visual guitar transcription. In *Jamboree 2008 : Workshop By and For KSpace PhD Students, July, 25 2008, Paris, France*, July 2008a.
-

12. M. Paleari, B. Huet, A. Schutz, and D. T. M. Slock. A multimodal approach to music transcription. In *1st ICIP Workshop on Multimedia Information Retrieval : New Trends and Challenges, October 12, 2008, San Diego, USA*, October 2008b.
13. M. Paleari, R. Benmokhtar, and B. Huet. Evidence theory based multimodal emotion recognition. In *MMM 2009, 15th International MultiMedia Modeling Conference, January 7-9, 2008, Sophia Antipolis, France*, January 2009.
14. M. Paleari, C. Velardo, J.-L. Dugelay, and B. Huet. Face Dynamics for Biometric People Recognition. In *MMSP 2009, IEEE International Workshop on Multimedia Signal Processing, October 5-7, 2009, Rio de Janeiro, Brazil*, October 2009a.
15. M. Paleari, V. Singh, B. Huet, and R. Jain. Toward Environment-to-Environment (E2E) Affective Sensitive Communication Systems. In *MTDL'09, Proceedings of the first ACM International Workshop on Multimedia Technologies for Distance Learning at ACM Multimedia, Bejin, China*, October 2009b.

### Future Work.

So far, the issues of affective display, emotion recognition, and affect synthesis and display have been extensively discussed. Despite the promise of these research topics, which have been shown in this thesis, these works have some limitations. In this section, we discuss these limitations and possible directions for future research.

We have described the process of generation of facial expressions for different platforms. So far, this process has been performed manually with several steps of ad-hoc refinements which have been made mandatory by some gaps in the psychological theory. A possible solution to this problem would be to analyze data from an emotional database with software similar to ARAVER in order to extract the timings of the activations of the facial muscles in the face. The study of the generated output data could, not only better validate Scherer [2001] theory of emotions, but also give relevant statistic data which may allow to fill the gaps in the theory and facilitate the automation of the facial expression generation process.

In part 5 we have presented a system for the automatic audio-visual recognition of human emotions. We argued that most of the limitations of existing systems performing this kind of task are related to the scarce availability of databases presenting all or most of the characteristics we have listed in section 5.3.1. We believe that the community should spend efforts in building a commonly used database of this kind. The main reasons for that are twofold: from the one hand, having an high quality database comprehensive of different modalities will allow to build better recognition algorithms; from the other hand, having a commonly used database will allow to better compare the results obtained by different systems.

Finally, we believe that an interesting future work would be to select a real-world scenario and to build a system merging the capabilities we developed as three separate modules (i.e. affect recognition, affect display, and affect synthesis and processing). The study of how each separate capability bias the perception of the users and their capabilities to interact effectively with the machine should teach us more about the human-machine interactions and give interesting hints for future research directions.

---

**Part I**  
**APPENDICES**

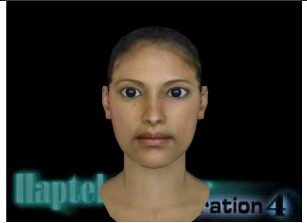
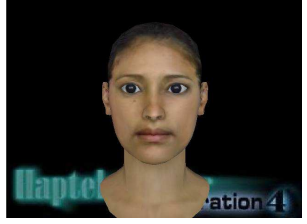








---




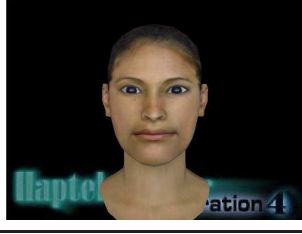


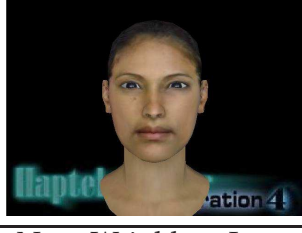


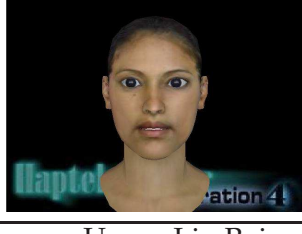


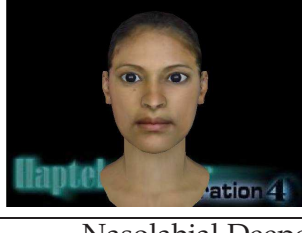


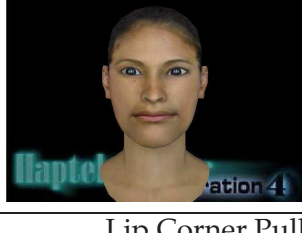
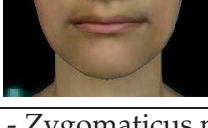





# Appendix A

## AU Table

AU	Haptik (whole face)	Haptik	Human
			
Neutral Face			
1			
Inner Brow Raiser - Frontalis, pars medialis			
2			
Outer Brow Raiser - Frontalis, pars lateralis			
4			
Brow Lowerer - Corrugator supercillii, Depressor supercillii			
Table A.1 – continued on next page			

AU	Haptel (whole face)	Haptel	Human
5			
Cheek Raiser - Orbicularis oculi, pars orbitalis			
6			
Lid Tightener - Orbicularis oculi, pars palpebralis			
7			
Nose Wrinkler - Levator labii superioris alaeque nasi			
10			
Upper Lip Raiser - Levator labii superioris			
11			
Nasolabial Deepener - Zygomaticus minor			
12			
Lip Corner Puller - Zygomaticus major			
Table A.1 – continued on next page			



















AU	Haptel (whole face)	Haptel	Human
13			
Cheek Puffer - Levator anguli oris (a.k.a. Caninus)			
14			
Dimpler - Buccinator			
15			
Lip Corner Depressor - Depressor anguli oris (a.k.a. triangularis)			
16			
Lower Lip Depressor - Depressor labii inferioris			
17			
Chin Raiser - Mentalis			
18			
Lip Puckerer - Incisivii labii superioris et inferioris			

Table A.1 – continued on next page










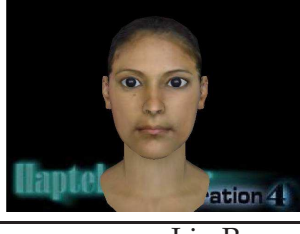

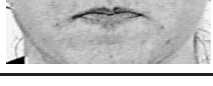




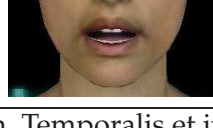
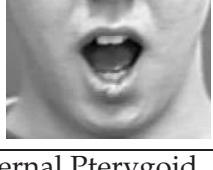
AU	Haptel (whole face)	Haptel	Human
20			
Lip Stretcher - Risorius et platysma			
22			
Lip Funneler - Orbicularis oris			
23			
Lip Tightener - Orbicularis oris			
24			
Lip Pressor - Orbicularis oris			
25			
Lips Part - Depressor labii inferioris or mentalis relaxation			
26			
Jaw Drop - Massete relaxation, Temporalis et internal Pterygoid			

Table A.1 – continued on next page


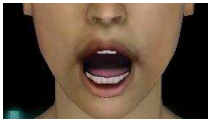

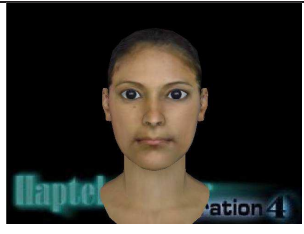





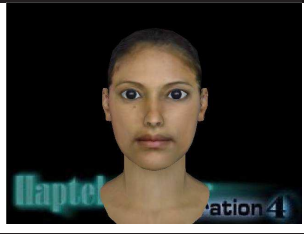








AU	Haptek (whole face)	Haptek	Human
27			
Mouth Stretch - Pterygoids, digastric			
28			
Lip Suck - Orbicularis Oris			
38			
Nostril Dilatator			
39			
Nostril Compressor			
41			
Lid Drop - Levator palpebrae superioris relaxation			
42			
Slit - Orbicularis oculi			

Table A.1 – continued on next page

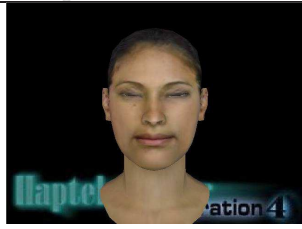


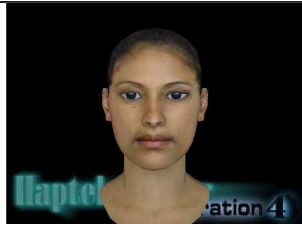




AU	Hapttek (whole face)	Hapttek	Human
43	 A 3D rendered female face with a neutral expression. The text 'Hapttek' and 'ation 4' is visible at the bottom.	 A close-up of the eyes, which are closed.	 A grayscale close-up of human eyes, which are closed.
Eyes Closed - Orbicularis oculi, pars palpebralis			
44	 A 3D rendered female face with a neutral expression. The text 'Hapttek' and 'ation 4' is visible at the bottom.	 A close-up of the eyes, which are squinted.	 A grayscale close-up of human eyes, which are squinted.
Squint - Orbicularis oris et pars palpebralis			
45	 A 3D rendered female face with a neutral expression. The text 'Hapttek' and 'ation 4' is visible at the bottom.	 A close-up of the eyes, which are partially closed.	
Blink - Orbicularis oris et pars palpebralis			

Table A.1: Hapttek showing AUs compared to actors

## Appendix B

# Alternative Feature Point Detection

### B.1 Introduction

In this appendix we present a work we have developed together with Silvia Altare and Sofia Rabellino, two EURECOM Master's students. The purpose of this work is to extract more precisely the feature point position by performing ad-hoc processing on different region of interest. This work takes deep inspiration from the work by Sohail and Bhattacharya [2007].

### B.2 ROI Definition

We have overviewed in section 5.3.5 the steps needed for the detection of the pupils. According to the anthropometric measurements we based on, we defined the following 6 region of interest (ROI) as in figure B.1(a):

1. left eye
2. right eye
3. left eyebrow
4. right eyebrow
5. mouth
6. nose

Every distance is based on the base eye-distance  $ed$  so that the model scales automatically to the subject face. The objective of the complete system is to extract the feature points as in figure B.1(b).

### B.3 Feature Point extraction

In the former section I have overviewed the steps needed for the identification of the region of interests on the face. In this section I will discuss the extraction of the feature point (FP) positions inside the different ROI.

---



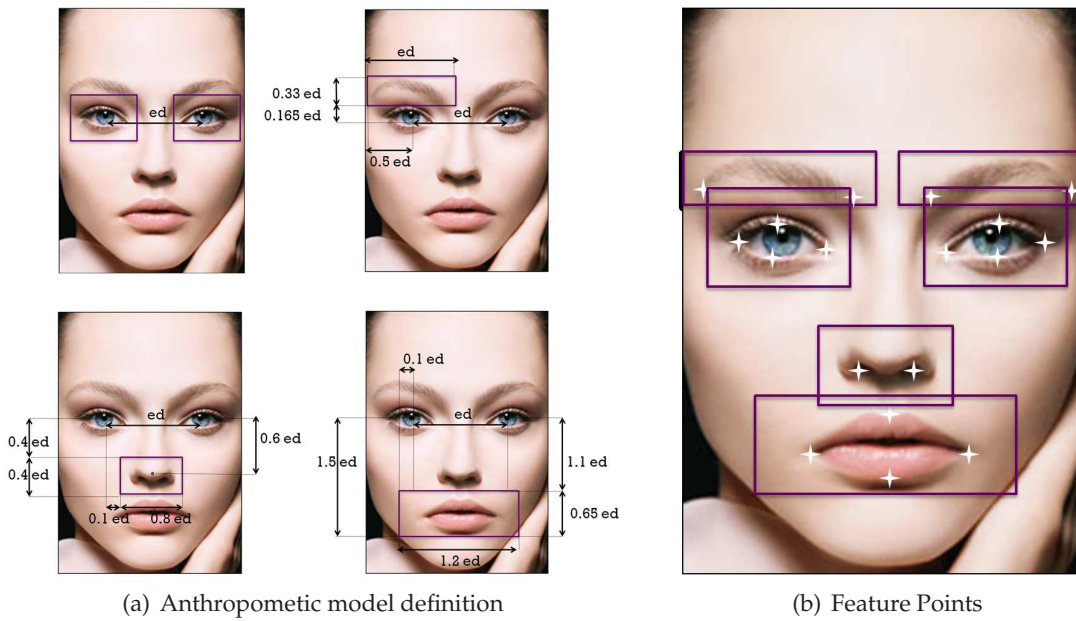


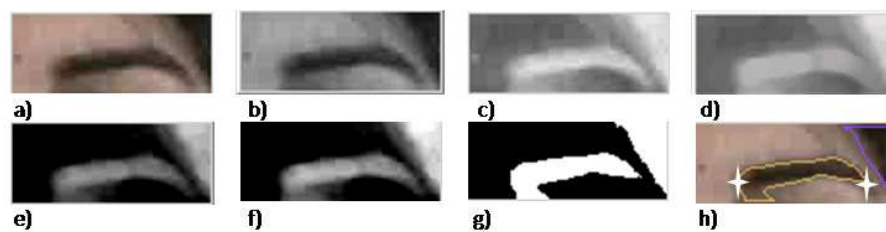
Figure B.1: Anthropometric model and desired feature points

### B.3.1 Eyebrow ROI processing

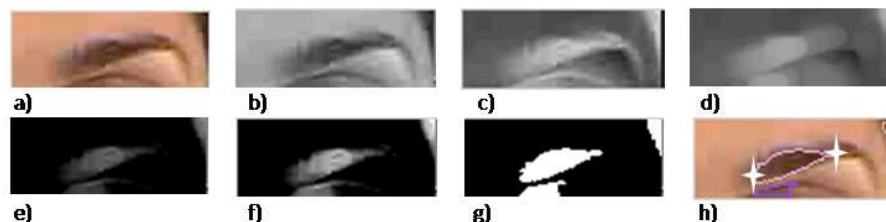
Once we obtained the Eyebrow ROI definition we started to process these regions in order to detect the eyebrow contour. Firstly, we based our experiments on the approach presented by Sohail and Bhattacharya [2007] “Detection of Facial Features Points Using Anthropometric Face Model”.

The method proposed consist of seven phases:

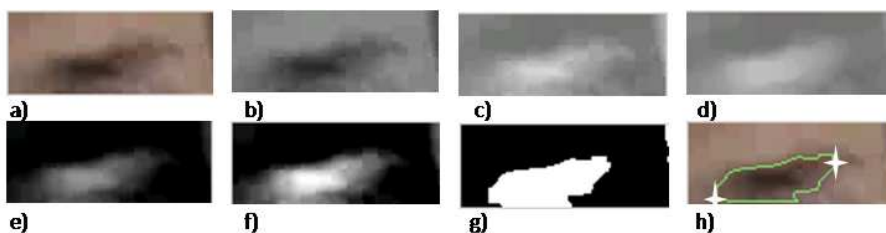
1. **Computation of the complemented eyebrow image:** the dark pixels of the eyebrow, considered as background in many digital imaging technologies, are converted to light colors and vice versa.
2. **Computation of the estimated background:** in order to obtain the background illumination we perform a morphological opening operation with a disk shaped structuring element.
3. **Background subtraction:** we subtract the estimated background from the complemented image to obtain a brighter eyebrow over a more uniform dark background.
4. **Intensity adjustment:** we perform this step on the basis of the pixels’ cumulative distribution.
5. **Binary eyebrow region:** we use the Otsu [1979] method.
6. **Eyebrow contour detection:** we detect the contours on the image and consider the largest obtained as the eyebrow one.
7. **Detection of eyebrow corners:** we consider the two points having the minimum and maximum value of x coordinate as corners.



(a) first example



(b) second example



(c) third example

Figure B.2: Eyebrow results with Sohail and Bhattacharya [2007] algorithms: a)original image, b)black and white image c)complemented image, d)background estimation, e)background subtraction, f)intensity adjustment, g)binarization, h)contour detection and eyebrow corner detection

We performed each step of this method, but we did not reach the expected results. In figure B.3.1 I show three examples of the obtained results.

As it can be seen, in the first case the eyebrow is well defined and uniformly illuminated, resulting in a quite accurate estimation of the feature points. In the other two cases the system cannot define the eyebrow contour in an appropriate manner and the feature point positions are not accurate. In our case the images apparently present more noise and are less defined than the one used by Sohail and Bhattacharya [2007] hence the obtained results were worsen.

Since we were not satisfied by the overall quality of the results we performed several experiments by changing the operations on the ROI and the order in which they are executed.

One step that was apparently important was the binarization step. In this case we have tried several different methods among which the dynamic threshold method provided by OpenCV and the Otsu [1979] method.

We finally decided to use the Otsu [1979] method, setting the value of the threshold to the mean intensity value of the ROI and added few noise-filtering steps. Our final

processing consist of 9 phases:

1. **Computation of the complemented eyebrow image:** to each pixel we assign the complementary gray color (e.g.  $px' = 255 - px$ ).
2. **Intensity adjustment by histogram stretching:** we compute the pixel intensity distribution and stretch it to cover the whole  $[0, 256]$  space.
3. **Morphological eroding operation:** we perform this operation with a disk shaped structuring element with radius of 3 pixels.
4. **Application of a median filter:** we reduce the noise of the image considering a neighborhood of 3 pixels.
5. **Morphological dilating operation:** we perform this operation with a disk shaped structuring element with radius of 3 pixels
6. **Binary eyebrow region:** we used Otsu [1979] method with the mean intensity value of the ROI used as threshold
7. **Morphological eroding operation:** we perform this operation with a disk shaped structuring element with radius of 4 pixels
8. **Eyebrow contour detection:** detect the contours on the image with the dedicated the OpenCV function.
9. **Detection of eyebrow corners:** we consider the 2 points having the minimum and maximum value of x coordinate as eyebrow corners.

figure B.3.1 show our results for the three samples in consideration. The detection of the eyebrows' contours is still far from being perfect but the results are better than the one obtained by following the literature approach.



Figure B.3: Literature results vs. our results

The low precision of the contour detection is due to the presence of many shadows in the eyebrow ROI. A greater resolution of the video and/or a better and more uniform illumination of the subject should overcome some of these limitations.

### B.3.2 Eye ROI processing

The processing for the eye region proposed by Sohail and Bhattacharya [2007] consist of four steps:

1. **Contrast adjustment**
2. **Image Binarization**
3. **Eye contour detection**
4. **Detection of eye feature points** (i.e. the top–most, left–most, central bottom–most, and bottom–most points)

In figure B.4 I show a sample result of such a processing.

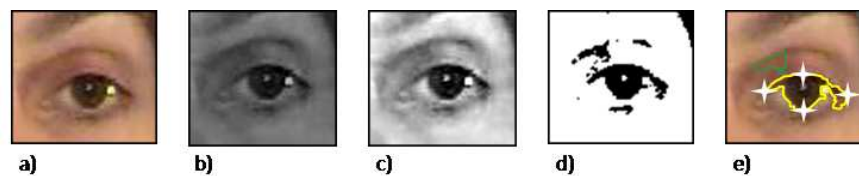


Figure B.4: Eye result with Sohail and Bhattacharya [2007] algorithms: a)original image, b)black and white image c)complemented image, d)intensity adjustment, e)binarization, f)contour detection and eye FP detection

In this case the results were satisfying although not necessarily robust to disturbances such as eyeglasses. Nevertheless, for the detection of the emotion, we are not too much interested about the details of the eye shape unless we could really detect small movements which is, anyway, not always possible with our video definition.

### B.3.3 Mouth ROI processing

Also for the mouth contour definition we based our computations on the paper by Sohail and Bhattacharya [2007].

The proposed steps are:

1. **Intensity adjustment:** we saturate the brighter 50% of the image pixels towards the highest intensity value.
2. **Complemented mouth image:** we complement the image.
3. **Binary mouth region:** we binarize the image with a fixed threshold.
4. **Mouth contour detection:** we detect the contour on the image and consider the largest obtained as the mouth one.
5. **Detection of mouth corners:** we consider four points along the mouth contour: the two extremities on the x axis and the middle points on the upper and the lower contour.

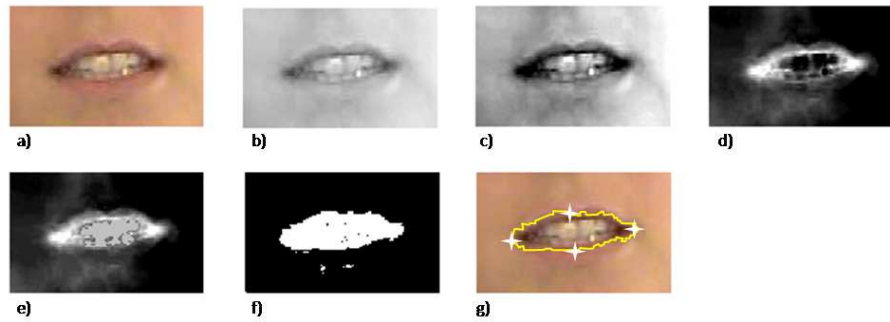


Figure B.5: Mouth result with Sohail and Bhattacharya [2007] algorithms: a)original image, b)black and white image d)intensity adjusted d)complemented image, e)flood fill, f)binarization, g)contour detection and mouth FP detection

We performed the first three steps on three subject obtaining the results in figure B.5.

Although the processing fails to estimate the lips as part of the mouth due to the insufficient contrast between the two regions we were still able to detect the position of the feature points quite accurately. Unfortunately this process is not very stable and, mainly due to different illumination conditions, some parameters have to be changed from one video to another. Future work should concentrate on finding automatic ways to adjust those parameters basing on more complex histogram adjustment.

## B.4 Conclusions

In the previous sections we have presented a work for feature point FP extraction based on Sohail and Bhattacharya [2007] work. We have shown some of the preliminary results we have obtained. The results are encouraging but our study pointed out how these algorithms still need to be made more robust to different illumination and presence of noise in the image.

## Appendix C

# Biometrics

Over the last decades the market of biometric person recognition has gained the attention of both the research community and private investors Luis-Garcia et al. [2003]. More and more private investors and public administrations are asking for systems capable of automatically and unintrusively recognize people in order to assure security of people, objects, and sensible data.

Although biometric solutions are becoming more and more appealing and mature, unconstrained biometric identification remain a largely unsolved problem and state of the art recognition systems are still far away from the capability of the human perception system Zhao et al. [2003].

Albeit several possible biometric information can be extracted unintrusively (e.g. voice, gait and stride, and others soft biometrics traits), automatic computer based face recognition is by far the most studied technique. Indeed, it is natural for humans to recognize a person from his facial appearance.

Most face recognition techniques can be classified into two big families according to the fact that they input still images or video shots. The system belonging to the first category attempt to recognize a subject exploiting only the physiological appearance of the subject; the ones belonging to the second class couple the information about the physiological appearance with information about the dynamic changes of this characteristic over time.

Traditionally, systems dealing with face recognition have to cope with four main challenges: 1) illumination changes, 2) head pose, 3) facial expressions, and 4) variations in facial appearance (e.g. make-up, glasses, beard).

The most well known technique is by far the system called eigen-faces Turk and Pentland [1991]. This technology for recognition of faces from still images creates a lower dimensional space using the faces of the train base and a standard dimensionality reduction technique such it is the principal component analysis (PCA) and recognize a test subject by projecting the test image in the same space and finding the subject linked to nearest train image.

Several techniques have been created to deal with head pose and facial expressions. A first one is known on the name of *Active Appearance Models* (AAM) Cootes et al. [1998], Edwards et al. [1998b]. In AAMs a robust representation of the subject is obtained by using a small set of parameters based on characteristics extracted from the position of a set of landmarks on the input image.

A second important contribution is represented by the *Elastic Graph Matching* (EGM) and the *Elastic Bunch Graph Matching* (EBGM) techniques. Lades et al. [1993], Wiskott

---



et al. [1997] introduce the use of a graph template for the extraction of features. In these techniques a grid template is stretched and deformed to imitate in the best possible way the face in the image. Negative weights and constraints are applied to the deformations of the template in order to set physical boundaries to the deformations of the facial appearance. In the case of EBGM Wiskott et al. [1997] the regular grid is replaced by a 3D graph representation of the human face accordingly to a set of fiducial points.

A first technique involving temporal information make use of *Hidden Markov Models* (HMM) Huang and Trivedi [2002], Liu and Cheng [2003]. HMM by their nature represent the temporal characteristics of signal by modeling the different states which better represent the signal in time, and the probabilities to pass from one state to another. In the face recognition state of the art these signals are represented by either the raw pixel values, eigen coefficients, or discrete cosine transform coefficients Huang and Trivedi [2002], Liu and Cheng [2003].

Perronnin et al. [2005] uses an approach based on HMM and EGM to model the differences between couple of images of the same subject. In this work the features represent both the facial appearance and the grid transformations.

Finally, Chen et al. [2001] exploited the technique known as *Optical Flow* (OF). In OFs a regular grid of image pixel blocks is tracked all along the video. The result is a grid of movement vectors representing the movement of the face inside the video.

In recent years NIST (Phillips et al. [2007]) has promoted an evaluation campaign for facial recognition systems. Different facial expression were depicted as well as different poses and illuminations. However, only two different facial expressions (neutral and smiling) were involved in the challenge while we know that the complexity of the human facial expressions goes far ahead Ekman et al. [1972].

Although several works have been processing face images and video for biometric face recognition, only few Chen et al. [2001] tried to exploit the facial expressions and the facial dynamics themselves as a source of biometric information. Indeed, most of the works in the state of the art still considers facial expressions as noise for the appearance recognition and, therefore, as something to avoid or suppress.

In this work we aim at demonstrating that facial expressions, and in particular the dynamics of emotional facial expressions and speech production, can be seen as a biometric source of information for the recognition of people. We point out that emotional facial expressions, as not dependent to age Ekman and Friesen [1986], are a stable source of information over the years. Furthermore, recognition systems based on facial dynamics have the advantage of being robust to illumination changes and variations of the facial appearance (e.g. beard, glasses, make-up).

This work takes direct inspiration from our system Paleari et al. [2009a] for emotional facial expression recognition and make use of the eNTERFACE'05 multimodal emotional database Martin et al. [2006].

## C.1 System

In this section we briefly overview our solution for the analysis of facial expressions (for further details please refers to Paleari et al. [2009a]). This approach consists in fusing the information coming from the tracking of some feature points (FP) placed in semantically meaningful locations of the subject's face.

We took inspirations from the results of the study by Ekman and Friesen Ekman and

---



Friesen [1986] which assessed that some emotional facial expressions (i.e. anger, disgust, fear, happiness, sadness, and fear) are independent from sex, ethnicity, age, and culture. We have then demonstrated Paleari et al. [2009a] that emotions can be recognized following the dynamic evolution of a facial expressions.

The dynamic analysis is performed for these six expressions in real-time exploiting the tracking of the displacement of the FP. Those are automatically detected and tracked on the subject's face we want to identify.

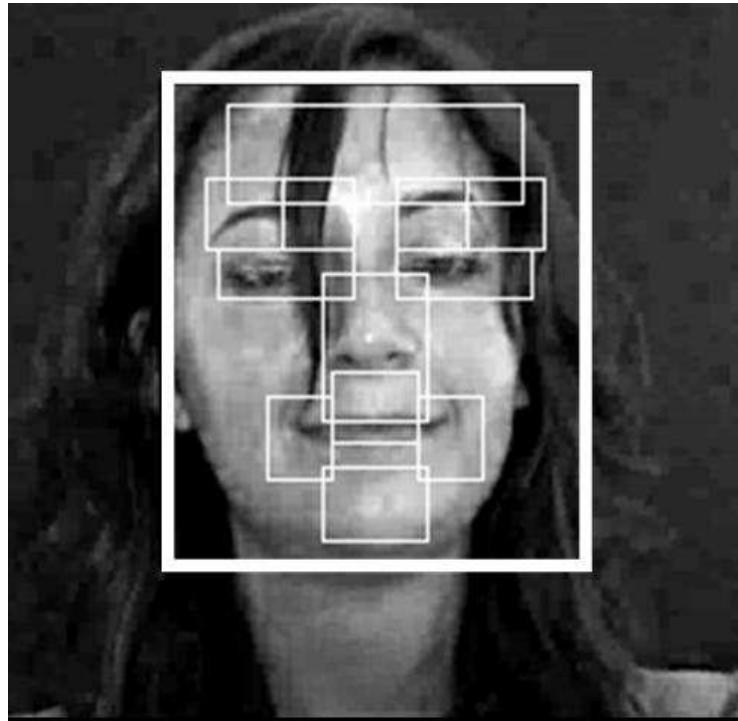


Figure C.1: Anthropometric 2D model

We made use of the eINTERFACE'05 database Martin et al. [2006] for our experiments (see figure C.1 for a reference). The database is the collection of more than 1300 videos regarding non-native English-speaking subjects showing emotions while uttering English sentences. Each sentence is related to one of the six universal emotions from Ekman and Friesen's studies Ekman et al. [1972]; therefore each video shot represent alternatively one emotion among anger, fear, disgust, happiness, sadness, or surprise.

Video shot duration is not constant and ranges between 1.2 to 6.7 seconds ( $2.8 \pm 0.8$  sec). The eINTERFACE'05 database is publicly available on the Internet but presents some drawbacks in term of quality because of several factors, some implicit to the representation of the videos and some to the set up of the database. Compression and interlacing applied to the videos, non-professional performances of the actors, and not English-proficiency of the actors (possibly affecting the quality of the vocal expression) affect the quality of the database Paleari et al. [2009a]. The reference paper itself Martin et al. [2006] admits that some videos are not fully representative of the related emotions. Dealing with the imperfections of such a database raised some difficulties but we argue that some of them are the same that can be found in real application scenarios. Therefore we can look at these imperfections as opportunities for devising more reliable methods.

### C.1.1 Expression analysis

In this section the steps needed for the extraction of facial feature points are overviewed. The system starts identifying the position of the subject face in the shot exploiting the Viola–Jones face detector Viola and Jones [2001]. For this purpose three different detectors are used. Face, mouth, and eyes of a subject are found in a video shot and exploited to estimate the pose and the face orientation angle.

Once the template of the face is found, we proceed by superimposing a simple anthropometric model of the human face (see figure C.2(a)). Using this quasi-rigid 2D model we are able to identify 12 regions of interest (ROIs) on the target face. These ROIs are representative of the following face parts:

1. mouth principal boundaries
2. forehead/nose/chin
3. right/left eyes
4. internal/external eyebrows extremities

For each one of these ROIs a cloud of Lucas–Kanade Tomasi and Kanade [1991] points is searched. The center of mass of these points is tracked in time. This processing results in 24 values (points coordinates) per frame (see figure C.2(a)). These coordinates are used as features representing the movement of the face region belonging to the 12 ROI we identified with the anthropometric template.

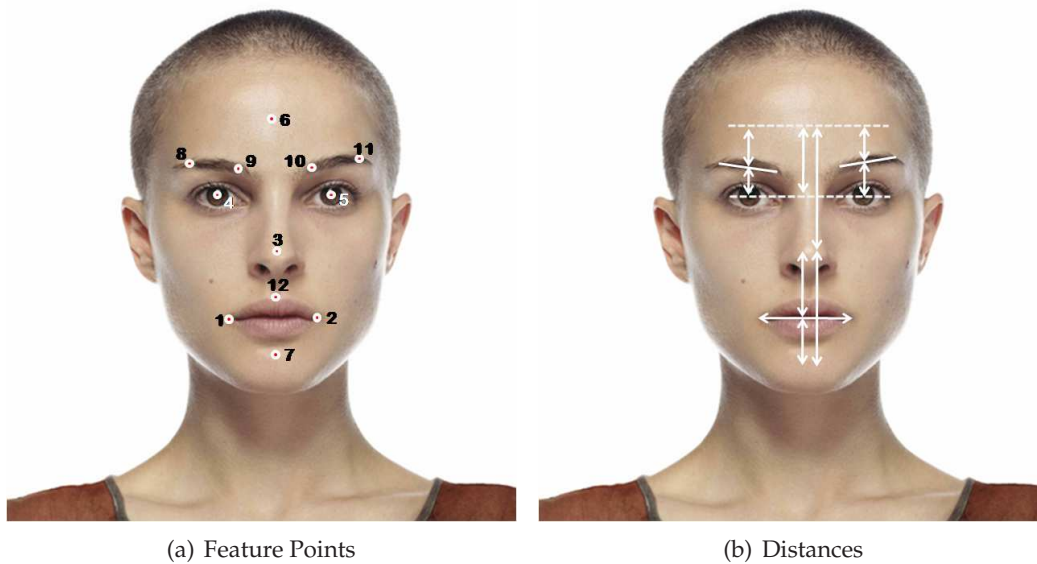


Figure C.2: Video Features

Then, we proceed to the normalization of each coordinate with respect to the nose position. This is done to get rid of the dynamics of the head movement. We keep this information aside, stored in the two variables relative to the nose.

We derive a more meaningful feature set from the 24 signals of the coordinates in order to represent the facial dynamics. These new features are computed as distances or

---

alignments among different points (see figure C.2(b)). This process is performed in order to extract the information of the dynamics evolution of subjects face.

The work of deriving this second feature set is directly inspired to the one done with MPEG-4 Face Definition Parameters (*FDPs*) and Face Animation Parameters (*FAPs*). As result we obtained a reduced set of 14 features  $D(i)$  where  $i = [1, 14]$  is defined hereafter:

1. head x displacement
2. head y displacement
3. normalization factor proportional to head z displacement
4. mouth corner distance
5. chin distance to mouth
6. nose distance to mouth
7. nose distance to chin
8. left eye to eyebrow distance
9. right eye to eyebrow distance
10. left eyebrow alignment
11. right eyebrow alignment
12. left eyebrow to forehead distance
13. right eyebrow to forehead distance
14. forehead to eye line distance

We extract and isolate the global motion information in three separated variables. We still believe this information is representative, as already demonstrated by Matta and Dugelay [2006], and can be efficiently exploited for biometric recognition. Such processing results in a compression of the data, indeed the information lead by 24 variables is now implicitly carried by 14 values.

For each subject we compute averages for the 14 distances. Then, we proceed to remove those values from each distance signal. Thanks to this normalization, we can get rid of the component of information which is linked to the appearance of the subject, thus building up a feature set which only represents the dynamics of the facial expressions.

## C.2 Experimental results and analysis

In this section we are going to present results of our experiments. The tests were conducted using a *GMM* approach. We have also tested approaches based on *HMMs* but the limited gap in terms of performances does not justify the increased complexity of this second technique. We do believe that such a result could be motivated by the not satisfying size of the dataset. The *GMM* used are characterized by one mixture of Gaussians (*MOG*). We also tested an increased number of *MOG* that resulted in comparable performances.

---

For each pair of subject and expression the database contains five repetitions. Each repetition represents the same expression while the uttered sentence changes.

Our results were found by performing complete leave-one-out tests among the five different sentences.

Three approaches were explored:

1. for each subject we have trained six different GMMs (one per emotional facial expression). The tests were carried out using the same expressions;
2. without changing the GMMs trained in the first step, we tested using data coming from different facial expressions (i.e. training on anger we tested on fear, disgust, etc.);
3. for each subject a single GMM was computed mixing all the data available. Similarly, tests were carried using all the available data.

The results are presented in figures C.3 and C.4.

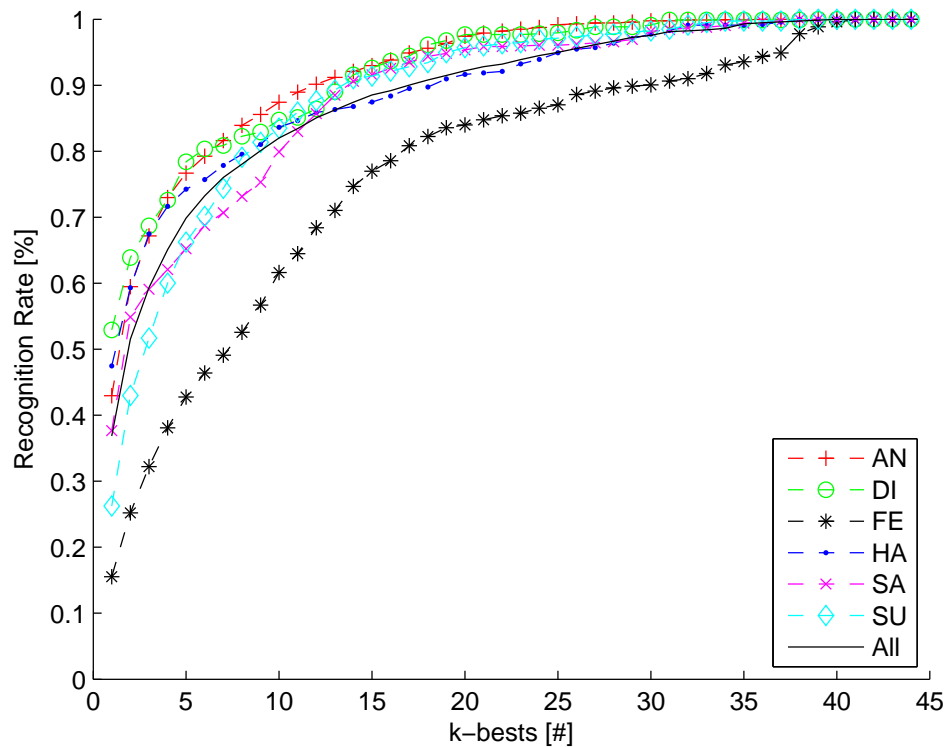


Figure C.3: GMM trained and tested on the same emotional data

Clearly one see from both figures that facial dynamics carry biometric information about the identity of the subjects.

In figure C.3 we show the results of the system of GMMs trained and tested on a specific emotional dynamics (first approach) and we compare it with the result obtained from the analysis conducted without the emotional state information (third approach).

We can observe that emotion specific GMMs perform, in average, slightly better than the output of the mixed approach. Although the distance among curves is little, we point

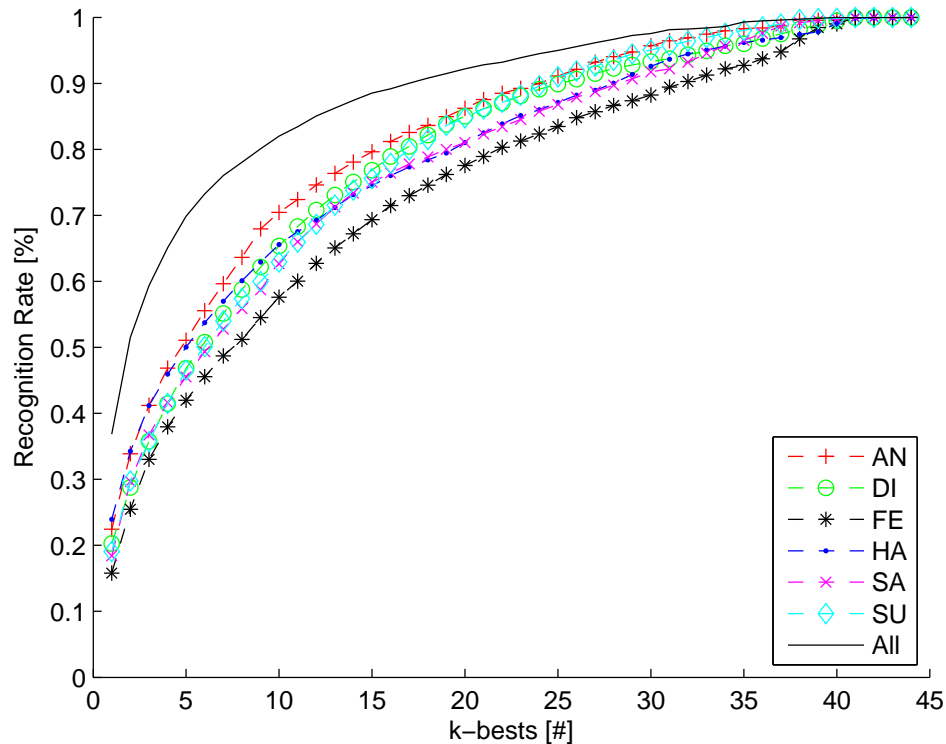


Figure C.4: GMM trained and tested on different emotional data

out that the training data size for emotion specific approach is six time smaller than the amount used for the latter. Given the size of our dataset, we believe that such a disparity could affect the results. In other words increasing the dataset size for the emotion specific GMMs to the one used for training the mixed approach may further improve the first results.

Another conclusion is that some emotional expressions can better discriminate the distances among subjects. In particular, when subjects depict fear their recognition became harder to accomplish. The reasons for that may be two: from one hand 1) it might be that all the subject have a similar emotional dynamics, on the other hand 2) it could be that way a subject represent its fear, changes a lot from one video to the other. In the first case we will observe that the intra-subject standard deviations (STD) of the GMMs mean values is low; in the second possibility the infra-subject STD of the GMMs  $\sigma$  will have low values. From our analysis on the GMMs mean and  $\sigma$  values, we can conclude that the latter possibility is verified.

In figure C.4 we show the results of the system of GMMs trained over an emotional facial expression and tested on the others (second approach). To help the comparison of the results we superimpose here the curve obtained from the analysis conducted without the emotional state information (third approach).

Notwithstanding the fact that performances deteriorated, we observe that part of the biometric information is kept. This suggests that two different components are carried by the facial dynamics: one first component being represented through emotional facial expressions and a second one being specific of the subject facial dynamics in term of vocal

production, twitches, muscles interactions, etc..

### C.3 Conclusions

In this appendix, we have shown a system for biometric people recognition based on facial dynamics. We have extracted these characteristics via a robust, automatic, and real-time point tracker. We have demonstrated that emotional facial expressions, up to now considered as noise by the state of the art, carry enough biometric information to distinguish among different people.

With our analysis we have demonstrated that:

1. facial dynamics carry biometric information
2. two different contributions participate to the recognition:
  - (a) emotional facial expressions
  - (b) subject dependent dynamics

We point out that algorithms exploiting dynamics are less prone to problems due to illumination and day to day facial variations (make-up, glasses, beard, etc.). Furthermore, the dynamics of emotional facial expressions are known to be independent to age, sex, ethnicity, and culture Ekman and Friesen [1986]. Therefore, using such characteristics help to build robust and reliable systems.

We strongly believe that with this work we open a new research path in the study of emotional facial dynamics. Nevertheless, as previously pointed out, a further analysis should be conducted when more data will be available.

---

## Appendix D

# Automatic Music Transcription

### D.1 Introduction

Written music is traditionally presented as a score, a musical notation which includes attack times, duration and pitches of the notes that constitute the song.

When dealing with the guitar this task is usually more complex. In fact, the only pitch of the note is not always enough to represent the movements and the positions that the performer has to execute to play a piece. A guitar can indeed chime the same note (i.e. a note with the same pitch) at different positions of the fretboard on different strings (See Fig. D.1). This is why the musical transcription of a guitar usually takes form of a tablature.

A tablature is a musical notation which includes six lines (one for each guitar string) and numbers representing the position at which the string has to be pressed to perform a note with a given pitch (figure D.2). Special notations are added to represent particular effects like bend ('^'), hammer on ('h'), pull off ('p'), etc.

The information about the movements which one player should execute to perform a piece can also be referred under the name of fingering. Burns and Wanderley Burns and Wanderley [2006] report few attempts that have been done to automatically extrapolate fingering information through computer algorithms:

1. real time processing using midi guitar
2. post processing using sound analysis
3. post processing using score analysis

Verner [1995] retrieves fingering information through the use of midi guitar. Using a midi guitar with different midi channels associated to each different string, Verner can, in real time, extract the complete tablature. The study points out that users of midi guitars reported false note detections, difficulties while playing, and problems of synchronization among the different strings. In any case this approach is not always applicable because it needs expensive equipment which, on top of that, is not usually used by performers on scene.

Traube [2004] suggests a solution based on the timbre of a guitar which only relies on the audio recording of a guitarist. Indeed, even if two notes have the same pitch they can have different timbre; with a-priori knowledge on the timbre of a guitar, it is

---



F3	B3	D#3	D3	C#3	C3	B2	A#2	A2	G#2	G2	F#2	F2	E2
A#4	A4	G#3	G3	F#3	F3	E3	D#3	D3	C#3	C3	B2	A#2	A2
D#4	D4	C#4	C4	B3	A#3	A3	G#3	G3	F#3	F3	E3	D#3	D3
G#4	G4	F#4	F4	E4	D#4	D4	C#4	C4	B3	A#3	A3	G#3	G3
C5	B4	A#4	A4	G#4	G4	F#4	F4	E4	D#4	D4	C#4	C4	B3
F5	E5	D#5	D5	C#5	C5	B4	A#4	A4	G#4	G4	F#4	F4	E4

Figure D.1: Notes on a guitar fretboard

therefore possible to estimate the fingering position. Common issues are precision, needs for a-priori knowledge, and monophonic operation limitation.

Another possibility is to analyze the produced score and to extract the tablature by applying a set of rules based on physical constraints of the instrument, biomechanical limitations, and others philological analysis. This kind of methods can result Radicioni et al. [2004] in tablatures which are similar to the one generated by humans, but hardly deal with situations in which the artistic intention or skill limitations are more important than the biomechanical movement.

Last but not least, Burns and Wanderley [2006] propose to use the visual modality to extract the fingering information. Their approach makes use of a camera mounted on the head of the guitar and extracts fingering information on the first 5 frets by using a circular Hough transformation to detect finger tips. Their system was positively evaluated in some preliminary studies but is not applicable to all cases because it needs ad hoc equipment, configuration, and it only returns information about the first 5 frets. Similarly Zhang et al. Zhang et al. [2007] track finger tips on a violin with a B-spline model of fingers contours.

This paper presents a multimodal approach to address this issue. The proposed approach combines information from video (webcam quality) and audio analysis in order to resolve ambiguous situations.

## D.2 Guitar Transcription

The typical scenario involved in the discussion of this paper involves one guitarist playing a guitar in front of a web-cam (XviD 640x480 pixels at 25 fps). In the work presented here the entire fretboard of the guitar needs to be completely visible on the video.

### D.2.1 Automatic Fretboard Detection

The first frame of the video is analyzed to detect the guitar and its position. The current version of our system presents few constraints: the guitarist is considered to play a right handed guitar (i.e. the guitar face on the right side) and to trace an angle with the horizontal which does not exceed  $90^\circ$ . The background is assumed to be less textured than the guitar. As a final result, this module returns the coordinates of the corner points defining the position of the guitar fretboard on the video (two outermost points for each

detected fret). Guitar frets have some interesting characteristics: they are straight and usually have a different brightness compared to the wood.

The process for obtaining the position of the frets is the following. The Hough transform is employed to find the orientation of the fret board, while the edges are obtained thanks to the Canny algorithm on the original image. The image is then rotated according to the dominant edge orientation in order to align the fret board with the horizontal axis. Wavelet analysis upon the rotated image is performed for enhancing the frets. Then, horizontal projection is performed in order to crop the image to the fretboard only. At this point we have a good estimation of the frets' position but due to some perspective effect the frets may not be straight.

Skewing is applied to the image until the vertical projections are maximized. Candidates (peaks) are chosen on the projection and identified on the original image (by couple). Invalid candidate frets are further filtered out by searching for the maximum energy path between top/bottom and bottom/top extremities. Paths cannot be greater than the distance between the two extremities. Additionally, if the two paths are different then the candidate fret is discarded. At this stage, only valid frets should remain.

## D.2.2 Fretboard Tracking

We have described how the fretboard position is detected on the first frame of the video. We make use of the Tomasi Lukas Kanade algorithm to follow the points along the video.

The coordinates of the end points of each fret are influenced by the movement of the hand. Therefore, some template matching techniques are applied to enforce points to stick to the fretboard. Two constraints were chosen to be invariant to scale, translation or 3D rotations of the guitar: 1) all the points defining the upper (as well as lower) bound of the fretboard must be aligned; 2) the lengths of the frets must comply to the rule  $L_i = L_{(i-1)} * 2^{-1/12}$  where  $L_i$  represent the length of the  $i^{th}$  fret.

To enforce the first constraint a first line is computed that matches the highest possible number of points. The points apart from the line are filtered out and a linear regression (least squares) is computed. All points apart from this second line are filtered out and recomputed.

The second constraint is applied by comparing the positions of the points with a template representing the distances of all the frets from the nut (i.e. the fret at the head of the guitar). The best match is found for having the lowest possible number of errors. Points outside the template are removed and their positions are recomputed.

Every twenty seconds the tracking is re initialized to solve any kind of issues which may arise from a wrongful adrifts of the Lukas Kanade point tracking (see section D.3). Furthermore, sometimes it may happen that no match can be found because too many points are lost at the same time or because the guitar is not facing the camera. In this cases a new match is searched in the following frames trough the algorithms described in section D.2.1.

## D.2.3 Hand Detection

In section D.2.2 the methodology employed to follow the position of the frets along the video has been described. Thanks to these coordinates it is possible to separate the region belonging to the fretboard into  $n\_strings \times n\_frets$  cells corresponding to each string/fret intersection.



Figure D.2: Interface of the Automatic Transcription System

Filtering is done on the frame to detect the skin color and the number of “hand” pixels is counted. A threshold can be applied to detect the presence of the hand (see figure D.3.a).

#### D.2.4 Audio Visual Information Fusion

Thanks to the audio analysis and standard audio processing techniques Serra [1997] we can extrapolate the pitch of the performed notes. This information is converted to a midi file with the information of the note played and the information of the attack time and duration of the note.

For each frame the information about the position of the hand is used to discriminate the correct fret-string couple producing a certain pitch.

Figure D.2 shows an example of the developed interface. We can see the interface incorporate two windows. The windows named “Tablature” shows the resulting tablature. The x axis represents the time and the six horizontal lines represent the six strings of the guitar. The vertical line at around 3/4 of the interface represent  $t = 0$ : at its right the information only comes from the audio analysis; at its left the information is fused together with the video information.

At the right of the line  $t = 0$ , the same note is represented at the same time on several strings to represent the incertitude that audio brings about when dealing with instru-

ments such as the guitar. Indeed that particular pitch can be played though all the tagged strings.

At time 0 (the time represented in the windows named “Original”) video is analyzed, the hand is detected at a certain fret and ambiguity is solved. At the left of the line  $t = 0$  only one note at time is therefore represented. One may notice that all positions represented in the tablature at the left of the line  $t = 0$  generate the same pitch ( $E3 = 164.81Hz$ ).

### D.3 Prototype

We have tested the proposed algorithms on several short videos (around 30 seconds per video). In these videos the guitarist performs different pattern designed to test the algorithms on four different guitars (two classical, one Spanish, and one acoustic). Videos were taken in our laboratories with a DV camera placed on a tripod at less than 2 meters from the guitarist and converted to XviD 640x480 pixels, 25 frames per second at around 250 Kbps. Audio was taken with the integrated camera microphone as well as with a gun zoom microphone to reduce ambient noise.

The guitar tracking algorithm worked correctly all along all the videos. Nevertheless, issues may arise when dealing with fast hand movement which may significantly reduce the number of trackable points and/or slide a consistent number of tracked points in a specific direction. In these cases the two constraints that were described in section D.2.2 may not be sufficient to perform a good tracking.

1) *alignment constraint*. If a significative number of points slide up or down the best fitting line may not be exactly parallel to the strings (see figure D.3 a).

2) *linear template constraint*. When a significative number of points slide horizontally or it is lost it can happen that the template matching matches better the wrong points than the correct ones. This may result in vertical lines which does not anymore match to the frets borders (see figure D.3 b).

With time both these phenomena may be amplified until the tracking is completely lost. We have empirically estimated both these phenomena to be sensible only after 30 to 40 seconds of videos and proceeded to re initialize the tracking algorithm every twenty seconds using the algorithm described in section D.2.

The hand detection was set to detect hand when at least 60% of the cell (i.e. the rectangle defining a fret and a string) contained the hand. This was found to be the minimum percentage allowing to have 0% false positives (which are due to the luminance of frets borders and strings). Setting the threshold at 60% was enough to solve 89% of the note ambiguities (see figure D.2).

In 11% of the cases a note which was played was assigned to two different possible positions. This corresponds to cases in which the played pitch matches with the fundamental pitch of a string (i.e. the pitch the string chime when played without pushing any fret; E2, A2, D3, G3, B3, E4). In this cases both possibilities are actually possible and our system did not disambiguate the note (see figure D.4 a).

In around the 3% of the cases one single long note was transcribed as two or more separate notes. This phenomenon was due to the artistic intention of the guitarist who slightly “bended” the string bringing both the hand and the string outside the cell. This will be addressed in future versions of the algorithm (see figure D.4 b).



(a) Correct Tracking



(b) Vertical adrift



(c) Horizontal adrift

Figure D.3: Example of video errors

## D.4 Future Work

A prototype has been described in the former section which demonstrates how the adoption of simple video analysis can help the process of generation of a tablature for guitar music. The example pieces involved in this first prototype only contained a small subset of the possible techniques involved in guitar music. In this section we list some of the possible improvements upon our system.

In the former section we have seen that our system may lose a note when the guitarist “bends” the string. Future work will solve this issue by applying a probabilistic model for the position of the hand. For each cell on the fretboard a  $P(h)$  will be computed representing the probability that the hand is both present on the cell and used to play (for example, a part from the case of “barre”, only finger tips are used).

Audio analysis will be extended to the polyphonic case allowing for chords and more complex pattern. To help the audio analysis dealing with polyphonic audio we will apply some machine learning techniques to learn prototypical hand positions and shapes (minor chords, major chords and principal variations).

Another system will explicitly perform right hand detection and following to estimate the string attack point to help both the audio and video processing units. Other system may be developed to detect guitar effects such as bending, tapping, slides, hammering on and pulling off, and others.



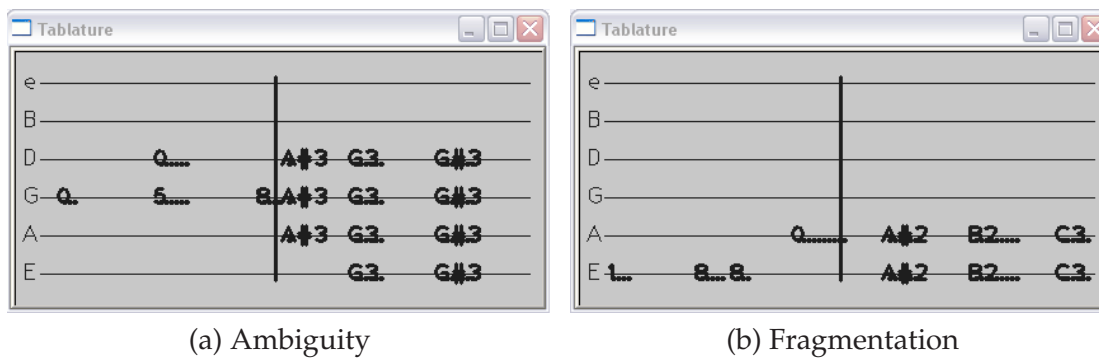


Figure D.4: Transcription Errors

## D.5 Conclusions

In this paper we have overviewed a complete, quasi unconstrained, guitar tablature transcription system which uses low cost video cameras to solve string ambiguities in guitar pieces. A prototype was developed as a proof of concept demonstrating the feasibility of the system with today technologies. Results of our studies are positive and encourage further studies on many aspects of guitar playing.

Applications of this research include computer aided pedagogical system which may significantly help guitar students, automatic indexing of song videos through tablature indexing, computer software which may help guitarist create and share music, and many others.





---

# Bibliography

- E. Adams and F. McGuire. Is laughter the best medicine? A study of the effects of humor on perceived pain and affect. *Activities, Adaptation and Aging*, 8:157–175, 1986.
- N. Amir and S. Ron. Towards an automatic classification of emotions in speech. In *ICSLP '98, Proceedings of the 5th International Conference on Spoken Language Processing*, pages 555–558, Sydney, Australia, 1998.
- M. B. Arnold. *Emotion and Personality*. Columbia University Press, New York, 1960.
- K. Balci. Xface: Mpeg-4 based open source toolkit for 3d facial animation. In *Proceedings of AVIO4, Working Conference on Advanced Visual Interfaces*, Gallipoli, Italy, 2004.
- K. Balci, E. Not, M. Zancanaro, and F. Pianesi. Xface open source project and smil-agent scripting language for creating and animating embodied conversational agents. In *ACM Multimedia*, pages 1013–1016, 2007.
- R. Banse and K. R. Scherer. Acoustic profiles in vocal emotion expression. *J Pers Soc Psychol*, 70(3):614–636, March 1996.
- S. M. Bartlett, G. Littlewort, M. Frank, C. Lainscsek, I. Fasel, and J. Movellan. Fully automatic facial action recognition in spontaneous behavior. In *FGR '06, Proceedings of the 7th International Conference on Automatic Face and Gesture Recognition*, pages 223–230, Washington, DC, USA, 2006. IEEE Computer Society.
- C. Bartneck. Integrating the OCC Model of Emotions in Embodied Characters. In *Proceedings of the Workshop on Virtual Conversational Characters: Applications, Methods, and Research Challenges*, Melbourne, Australia, 2002.
- C. Bartneck, J. Reichenbach, and A. van Breemen. In your face, robot! the influence of a character's embodiment on how users perceive its emotional expressions. In *Fourth International Conference on Design & Emotion*, Ankara, Turkey, July 2004.
- J. Bates. The role of emotion in believable agents. *Communications of ACM*, 37(7):122–125, July 1994.
- A. Batur and M. Hayes. Adaptive active appearance models. *IEEE Transactions on Image Processing*, 14(11):1707–1721, November 2005.
- R. Benmokhtar and B. Huet. Classifier fusion : combination methods for semantic indexing in video content. In *Proceedings of ICANN*, volume 4132, pages 65–74, 2006.
-

- R. Benmokhtar and B. Huet. Neural network combining classifier based on Dempster-Shafer theory for semantic indexing in video content. In *Proceedings of MMM '07, 13th International Conference on MultiMedia Modeling*, volume 4351 of LNCS, pages 196–205, Singapore, January 2007a. Springer Berlin / Heidelberg.
- R. Benmokhtar and B. Huet. Multi-level Fusion for Semantic Video Content Indexing and Retrieval. In *Proceedings of AMR '07, International Workshop on Adaptive Multimedia Retrieval*, volume 4918 of LNCS, pages 160–169, Paris, France, July 2007b. Springer Berlin / Heidelberg.
- C. Besson, D. Graf, I. Hartung, B. Kropfhäusser, and S. Voisard. The Importance of Non-verbal Communication in Professional Interpretation. §Introduction to Interpretation course, 2004. URL <http://www.aiic.net/ViewPage.cfm/page1662.htm>.
- C. M. Bishop. *Neural Networks for Pattern Recognition*. Oxford University Press, November 1995.
- C. M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer, August 2006.
- A. W. Black and P. A. Taylor. The Festival Speech Synthesis System: System documentation. Technical Report HCRC/TR-83, Human Communication Research Centre, University of Edinburgh, Scotland, UK, 1997. Available at <http://www.cstr.ed.ac.uk/projects/festival.html>.
- P. Boersma and D. Weenink. Praat: doing phonetics by computer, January 2008. [<http://www.praat.org/>].
- R. A. Bolt. Put-that-there: Voice and gesture at the graphics interface. In *SIGGRAPH '80, Proceedings of the 7th annual conference on Computer graphics and interactive techniques*, pages 262–270, Seattle, Washington, United States, 1980. ACM Press.
- A. L. Bouhuys, G. M. Bloem, and T. G. G. Groothuis. Induction of depressed and elated mood by music influences the perception of facial emotional expressions in healthy subjects. *Journal of Affective Disorders*, 33(4):215–226, 1995.
- C. L. Breazeal. *Designing Sociable Robots (Intelligent Robotics and Autonomous Agents)*. The MIT Press, September 2004.
- F. Burkhardt. Expressive synthetic speech website. <http://emosamples.syntheticspeech.de/>, May 2009.
- A. Burns and M. M. Wanderley. Visual methods for the retrieval of guitarist fingering. In *NIME '06: Proceedings of the 2006 conference on New interfaces for musical expression*, pages 196–199, Paris, France, 2006.
- C. Busso, Z. Deng, S. Yildirim, M. Bulut, C. M. Lee, A. Kazemzadeh, S. Lee, U. Neumann, and S. Narayanan. Analysis of emotion recognition using facial expressions, speech and multimodal information. In *ICMI'02, Proceedings of the ACM International Conference on Multimodal Interfaces*, pages 205–211, 2004. State College, PA, USA.
- L. Cahill and J. L. McGaugh. A novel demonstration of enhanced memory associated with emotional arousal. *Consciousness and Cognition*, 4:410–421, 1995.
-

- 
- J. Cassell. *Embodied Conversational Agents*. The MIT Press, April 2000.
- G. Castellano, L. Kessous, and G. Caridakis. Multimodal emotion recognition from expressive faces, body gestures and speech. In F. de Rosis and R. Cowie, editors, *Proceedings of the Doctoral Consortium of 2nd International Conference on Affective Computing and Intelligent Interaction*, Lisbon, Portugal, September 2007.
- C. H. Chan and G. J. F. Jones. Affect-based indexing and retrieval of films. In *ACM MM '05: Proceedings of the 13th annual ACM international conference on Multimedia*, pages 427–430, New York, NY, USA, 2005. ACM.
- Y. Chang, C. Hu, and M. Turk. Probabilistic expression analysis on manifolds. In *CVPR'04, Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 2, pages 520–257, Santa Barbara, CA, USA, June 2004.
- P. Charles, N. Good, L. L. Jordan, J. Pal, P. Lyman, H. R. Varian, and K. Swearingen. *How much information 2003?* berkeley.edu, Berkeley, CA, October 2003. <http://www2.sims.berkeley.edu/research/projects/how-much-info-2003/>.
- L. Chen, H. Tao, T. Huang, T. Miyasato, and R. Nakatsu. Emotion recognition from audiovisual information. In *Proceedings, IEEE Workshop on Multimedia Signal Processing*, pages 83–88, Los Angeles, CA, USA, 1998.
- L. Chen, H. Liao, and J. Lin. Person identification using facial motion. In *Proceedings of ICIP, International Conference on Image Processing*, pages 677–680, October 2001.
- C. Clavel and G. Richard. *Reconnaissance acoustique des émotions*, chapter Systèmes d'Interaction Emotionnelle. Hermès, 2009. to appear.
- I. Cohen, N. Sebe, A. Garg, L. S. Chen, and T. S. Huang. Facial expression recognition from video sequences: temporal and static modeling. *Journal of Computer Vision and Image Understanding: Special issue on Face recognition*, 91(1–2):160–187, 2003.
- B. Coker. Freedom to surf: workers more productive if allowed to use the internet for leisure. <http://uninews.unimelb.edu.au/news/5750/>, 2009. Talk at University of Melbourne.
- A. Colmenarez, B. Frey, and T. Huang. Embedded face and facial expression recognition. In *Image Processing, 1999. ICIP 99. Proceedings. 1999 International Conference on*, volume 1, pages 633–637, 1999.
- C. Conati. Probabilistic Assessment of Users Emotions in Educational Games. *Applied Artificial Intelligence*, 16(21):555–575, 2002.
- C. Conati and X. Zhou. Modeling students' emotions from cognitive appraisal in educational games. In *Springer-Verlag Berlin Heidelberg*, pages 944–954, 2002.
- T. Cootes, G. Edwards, and C. Taylor. Active appearance models. *Computer Vision*, 2: 484–498, 1998.
- N. Corporation. Nec personal robot research center. <http://www.nec.co.jp/products/robot/>, 2005.
-

- A. Corradini, M. Mehta, N. Bernsen, and J.-C. Martin. Multimodal Input Fusion in Human-Computer Interaction on the Example of the on-going NICE Project. In *to appear Proceedings of the NATO-ASI conference on Data Fusion for Situation Monitoring, Incident Detection, Alert and Response Management*, Yerevan (Armenia), August 2003.
- G. W. Cottrell and J. Metcalfe. Empath: face, emotion, and gender recognition using holons. In *NIPS-3: Proceedings of the 1990 conference on Advances in neural information processing systems 3*, pages 564–571, Denver, CO, USA, 1990. Morgan Kaufmann Publishers Inc.
- S. R. Covey. *The 7 Habits of Highly Effective People*. Free Press, November 2004.
- R. E. Cytowic. Synesthesia and mapping of subjective sensory dimensions. *Neurology*, 39(6):849–850, June 1989.
- A. Damasio. *Descartes' Error : Emotion, Reason, and the Human Brain*. Penguin (Non-Classics), September 2005.
- C. Darwin. *The Expression of the Emotions in Man and Animals*. Oxford University Press Inc, 3<sup>rd</sup> edition, 1872.
- D. Datcu and L. Rothkrantz. Semantic audio-visual data fusion for automatic emotion recognition. In *Euromedia'08, Proceedings of the European Conference on Multimedia*, Porto, 2008.
- R. Davidson, K. Scherer, and H. Goldsmith. *The Handbook of Affective Science*. Oxford University Press, March 2002.
- B. de Carolis, C. Pelachaud, I. Poggi, and M. Steedman. *Life-Like Characters, Cognitive Technologies*, chapter APML, a mark-up language for believable behavior generation, pages 65–86. Springer, 2004.
- F. de Rosis, C. Pelachaud, I. Poggi, V. Carofiglio, and B. de Carolis. From greta's mind to her face: modelling the dynamics of affective states in a conversational embodied agent. *International Journal of Human-Computer Studies*, 59(1-2):81–118, July 2003.
- F. Dellaert, T. Polzin, and A. Waibel. Recognizing emotion in speech. In *ICSLP '96, Proceedings of the Fourth International Conference on Spoken Language Processing*, volume 3, pages 1970–1973, October 1996.
- N. Dimitrova. Multimedia Content Analysis: The Next Wave. In Springer, editor, *Proceedings of Image and Video Retrieval*, volume 2728 of *Lecture Notes in Computer Science*, pages 415–420, 2003.
- G. B.-A. Duchenne De Boulogne. *The Mechanism of Human Facial Expression*. Studies in Emotion and Social Interaction. Cambridge University Press, New York, 1990 edition, July 1862.
- G. Edwards, C. Taylor, and T. Cootes. Interpreting face images using active appearance models. In *IEEE Proceedings on Automatic Face and Gesture Recognition*, pages 300–305, 1998a.
-

- 
- G. J. Edwards, T. F. Cootes, and C. J. Taylor. Face recognition using active appearance models. In *ECCV '98, Proceedings of the 5th European Conference on Computer Vision*, volume 2, pages 581–595, London, UK, 1998b. Springer-Verlag.
- P. Ekman. Universals and Cultural Differences in Facial Expressions of Emotion. In J. K. Cole, editor, *Proceedings of Nebraska Symposium on Motivation*, volume 19, pages 207–283, Lincoln (NE), 1971. Lincoln: University of Nebraska Press.
- P. Ekman. *Approaches to Emotions*, chapter Expression and the Nature of Emotion, pages 319–343. Lawrence Erlbaum Associates, 1984.
- P. Ekman. Are there basic emotions? *Psychological Review*, 99(3):550–553, July 1992.
- P. Ekman and W. Friesen. *Facial Action Coding System: A Technique for the Measurement of Facial Movement*. Consulting Psychologists Press, Palo Alto, 1978.
- P. Ekman and W. V. Friesen. A new pan cultural facial expression of emotion. *Motivation and Emotion*, 10(2):159–168, 1986.
- P. Ekman and K. R. Scherer. *Approaches to Emotions*, chapter Questions About Emotions: an Introduction, pages 1–8. Lawrence Erlbaum Associates, 1984.
- P. Ekman, W. V. Friesen, and P. Ellsworth. *Emotion in the Human Face*. Oxford University Press, 1972.
- P. Ekman, W. V. Friesen, and P. Ellsworth. *Emotion in the human face*, chapter What emotion categories or dimensions can observers judge from facial behavior?, pages 39–55. Cambridge University Press, New York, 1982.
- P. Ekman, W. V. Friesen, and J. C. Hager. *Facial Action Coding System Investigator's Guide. A Human Face*, 2002.
- I. A. Essa and A. P. Pentland. Coding, Analysis, Interpretation, and Recognition of Facial Expressions. *IEEE Trans. Pattern Anal. Mach. Intell.*, 19(7):757–763, 1997.
- B. Fasel, F. Monay, and D. Gatica-Perez. Latent semantic analysis of facial action codes for automatic facial expression recognition. In *MIR '04, Proceedings of the 6th ACM SIGMM international workshop on Multimedia Information Retrieval*, pages 181–188, New York, NY, USA, 2004. ACM.
- W. A. Fellenz, J. G. Taylor, R. Cowie, E. Douglas-Cowie, F. Piat, S. Kollias, C. Orovas, and B. Apolloni. On emotion recognition of faces and of speech using neural networks, fuzzy logic and the assess system. In *IJCNN '00, Proceedings of the IEEE-INNS-ENNS International Joint Conference on Neural Networks*, volume 2, pages 93–98, Washington, DC, USA, 2000. IEEE Computer Society.
- N. Frijda and J. Swagerman. Can computers feel? Theory and design of an emotional system. *Cognition and Emotion*, 1(3):235–257, 1987.
- N. H. Frijda. *The Emotions: Studies in Emotion and Social Interaction*. Cambridge University Press, April 1986.
- W. Fry. The biology of humor. *Humor: International Journal of Humor Research*, 7:111–126, 1994.
-



- E. Galmar and B. Huet. Analysis of Vector Space Model and Spatiotemporal Segmentation for Video Indexing and Retrieval. In *ACM International Conference on Image and Video Retrieval*, Amsterdam, The Netherlands, 2007.
- H. E. Gardner. *Frames of Mind: The Theory of Multiple Intelligence*. Basic Books, New York, 1983.
- P. J. Gmytrasiewicz and C. L. Lisetti. Emotions and personality in agent design. In *AA-MAS '02: Proceedings of the first international joint conference on Autonomous agents and multiagent systems*, pages 360–361, New York, NY, USA, 2002. ACM.
- D. Goleman. *Emotional Intelligence: 10th Anniversary Edition; Why It Can Matter More Than IQ*. Bantam, September 2006.
- J. Goodman. Laughing matters: taking your job seriously and yourself lightly. *Journal of the American Medical Association*, 267(1858):11–13, 1992.
- J. Gratch. Émile: Marshalling Passions in Training and Education. In C. Sierra, M. Gini, and J. S. Rosenschein, editors, *Proceedings of the Fourth International Conference on Autonomous Agents*, pages 325–332, Barcelona, Catalonia, Spain, 2000a. ACM Press.
- J. Gratch. Socially situated planning. In K. Dautenhahn, editor, *Socially Intelligent Agents: The Human in the Loop (Papers from the 2000 AAAI Fall Symposium)*, pages 61–64. Kluwer Academic Publishers, 2000b.
- J. Gratch and S. Marsella. Tears and fears: modeling emotions and emotional behaviors in synthetic agents. In *AGENTS '01: Proceedings of the fifth international conference on Autonomous agents*, pages 278–285, New York, NY, USA, 2001. ACM Press.
- J. Gratch and S. Marsella. A domain-independent framework for modeling emotion. *Cognitive Systems Research*, 5(4):269–306, December 2004.
- J. A. Gray. The whole and its parts: Behaviour, the brain, cognition and emotion. *Bulletin of the British Psychological Society*, 38:99–122, 1985.
- A. Grizard, M. Paleari, and C. L. Lisetti. Adaptation d’une théorie psychologique pour la génération d’expressions faciales synthétiques pour des agents d’interface. In *WACA 2006, 2eme Workshop sur les Agents Conversationnels Animés, 26-27 octobre 2006, Toulouse, France*, October 2006.
- N. Group. NPD Group website. <http://www.npd.com>, 2009.
- J. B. Halberstadt, P. M. Niedenthal, and J. Kushner. Resolution of lexical ambiguity by emotional state. *Psychological Science*, 6(5):278–282, September 1995.
- T. Halfhill. Parallel processing with CUDA. *Microprocessor Report*, 2008.
- S. B. Hamann, T. D. Ely, S. T. Grafton, and C. D. Kilts. Amygdala Activity Related to Enhanced Memory for Pleasant and Aversive Stimuli. *Nature Neuroscience*, 2(3):289–293, March 1999.
- M. Handel and J. Herbsleb. What is Chat doing in the workplace? In *Proceedings of the ACM CSCW'02, International Conference on Computer Supported Cooperative Work*, pages 1–10, 2002. ACM Press.
-

- 
- Haptik. Haptik website. [www.haptik.com](http://www.haptik.com), 2006.
- C. Harris and M. Stephens. A combined corner and edge detection. In *Proceedings of The Fourth Alvey Vision Conference*, pages 147–151, 1988.
- D. O. Hebb. *A Textbook of Psychology*. W. B. Saunders Co., Philadelphia, 1966.
- R. W. J. Hill, A. W. Hill, J. Gratch, S. Marsella, J. Rickel, W. Swartout, and D. Traum. Virtual humans in the mission rehearsal exercise system. In *Proceedings of KI Embodied Conversational Agents*, 17:32–38, 2003.
- G. Hofer, K. Richmond, and R. Clark. Informed blending of databases for emotional speech synthesis. In *Proc. Interspeech*, September 2005.
- H. Hong, H. Neven, and C. von der Malsburg. Online facial expression recognition based on personalized galleries. In *FG '98, Proceedings of the 3rd. International Conference on Face & Gesture Recognition*, pages 354–359, Washington, DC, USA, 1998. IEEE Computer Society.
- C. W. Hsu, C. C. Chang, and C. J. Lin. A practical guide to support vector classification. Technical report, Department of Computer Science, National Taiwan University, 2003.
- C.-L. Huang and Y.-M. Huang. Facial expression recognition using model-based feature extraction and action parameters classification. *Journal of Visual Communication and Image Representation*, 8(3):278–290, 1997.
- S. Huang and M. Trivedi. Streaming face recognition using multicamera video arrays. In *Proceedings of Pattern Recognition*, pages 213–216, 2002.
- R. Huber, A. Batliner, J. Buckow, E. Nöth, V. Warnke, and H. Niemann. Recognition of emotion in a realistic dialogue scenario. In *ICSLP '00, Proceedings of the 6th International Conference on Spoken Language Processing*, pages 665–668, 2000.
- E. Hudlicka and J.-M. Fellous. Review of computational models of emotion. Technical Report 9612, Psychometrix, Arlington, MA, April 1996.
- Intel Corporation. Open Source Computer Vision Library: Reference Manual, November 2006. [<http://opencvlibrary.sourceforge.net>].
- E. Isaacs, A. Walendowski, S. Whittaker, D. Schiano, and C. Kamm. The Character, Functions, and Styles of Instant Messaging in the Workplace. In *Proceedings of the ACM CSCW'02: International Conference on Computer-Supported Cooperative Work*, pages 11–20, 2002. ACM Press.
- C. E. Izard. *Human Emotions*. Plenum Press, New York, 1977.
- C. E. Izard. Four systems for emotion activation: Cognitive and noncognitive processes. *Psychological Review*, 100(1):68–90, 1993.
- S. j. Garlow and C. B. Nemeroff. *The Handbook of Affective Science*, chapter Neurobiology of Depressive Disorders, pages 1021–1043. Oxford University Press, New York, 2002.
- A. Jaimes and N. Sebe. Multimodal human-computer interaction: A survey. *Computer Vision and Image Understanding*, 108(1-2):116–134, October 2007.
-



- W. James. What is an emotion? *Mind*, 9:188–205, 1884.
- Q. Ji, P. Lan, and C. Looney. A probabilistic framework for modeling and real-time monitoring human fatigue. *Systems, Man and Cybernetics, Part A: Systems and Humans, IEEE Transactions on*, 36(5):862–875, September 2006.
- C. Jones and J. Sutherland. *Affect and Emotion in Human-Computer Interaction: From Theory to Applications*, volume 4868, chapter Acoustic Emotion Recognition for Affective Computer Gaming, pages 209–219. Springer-Verlag, Berlin, 2008.
- R. Junea. Youtube Official Blog. <http://www.youtube.com/blog?entry=on4EmafA5MA>, May 2009.
- S. Kaiser and T. Wehrle. *Appraisal Processes in Emotion: Theory, Methods, Research*, chapter Facial Expressions as Indicators of Appraisal Processes, pages 285–300. Oxford University Press, 2001.
- T. Kang, C. Han, S. Lee, D. Youn, and C. Lee. Speaker dependent emotion recognition using speech signals. In *Proceedings of ICSLP 2000*, pages 383–386, 2000.
- A. Kapoor and R. W. Picard. Multimodal affect recognition in learning environments. In *ACM MM'05: Proceedings of the 13th annual ACM international conference on Multimedia*, pages 677–682, 2005.
- A. Kapoor, S. Mota, and R. W. Picard. Towards a Learning Companion that Recognizes Affect. In *Proceedings from Emotional and Intelligent II: The Tangled Knot of Social Cognition, AAI Fall Symposium*, 2001.
- G. D. Kearney and S. McKenzie. Machine interpretation of emotion: Design of a memory-based expert system for interpreting facial expressions in terms of signaled emotions. *Cognitive Science: A Multidisciplinary Journal*, 17(4):589–622, January 1993.
- S. Kettebekov and R. Sharma. Toward Multimodal Interpretation in a Natural Speech and Gesture Interface. In *Proceedings of IEEE Symposium on Image, Speech, and Natural Language Systems*, pages 328–335. IEEE, November 1999.
- E. Kim, S. Kim, H. Koo, K. Jeong, and J. Kim. Emotion-Based Textile Indexing Using Colors and Texture. In L. Wang and Y. Jin, editors, *Proceedings of Fuzzy Systems and Knowledge Discovery*, volume 3613 of LNCS, pages 1077–1080. Springer, 2005.
- S. Kimura and M. Yachida. Facial expression recognition and its degree estimation. In *CVPR '97, Proceedings of the 1997 Conference on Computer Vision and Pattern Recognition (CVPR '97)*, pages 295–300, Washington, DC, USA, 1997. IEEE Computer Society.
- H. Kobayashi and F. Hara. Facial interaction between animated 3d face robot and human beings. In *Systems, Man, and Cybernetics, 1997. 'Computational Cybernetics and Simulation'*, 1997 IEEE International Conference on, volume 4, pages 3732–3737, October 1997.
- T. Koda and P. Maes. Agents with faces: The effects of personification of agents. In *HCI'96, Proceedings of Human Computer Interactions*, London, UK, 1996.
-

- 
- B. Kort, R. Reilly, and R. Picard. An affective model of interplay between emotions and learning: reengineering educational pedagogy-building a learning companion. In *ALT 2001. Proceedings of the IEEE International Conference on Advanced Learning Technologies*, pages 43–46, 2001.
- F. Kuo, M. Chiang, M. Shan, and S. Lee. Emotion-based music recommendation by association discovery from film music. In *ACM MM'05 Proceedings of ACM International Conference on Multimedia*, pages 507–510, Singapore, 2005.
- M. Lades, J. Vorbruggen, J. Buhmann, J. Lange, C. Malsburg, R. Wurtz, and W. Konen. Distortion invariant object recognition in the dynamic link architecture. *IEEE Transactions on Computers*, 42:300–311, March 1993.
- R. S. Lazarus. *Emotion and Adaptation*. Oxford University Press, New York, August 1991.
- R. S. Lazarus and B. N. Lazarus. *Passion and Reason: Making Sense of Our Emotions*. Oxford University Press, New York, 1996.
- J. E. LeDoux. *Learning and Computational Neuroscience: Foundations of Adaptive Networks*, chapter Information flow from sensation to emotion: Plasticity in the neural computation of stimulus value, pages 3–51. MIT Press, 1990.
- J. E. LeDoux. Emotion, memory and the brain. *Scientific American*, 270(6):50–57, June 1994.
- H. Leventhal. *Perception of emotion in self and others*, volume 5, chapter A perceptual motor processing model of emotion, pages 263–299. New York, Plenum Press, 1979.
- H. Leventhal. A perceptual–motor theory of emotion. *Journal of Advances in Experimental Social Psychology*, 17:117–182, 1984.
- H. Leventhal and K. R. Scherer. The relationship of emotion to cognition: A functional approach to a semantic controversy. *Cognition and Emotion*, 1:3–28, 1987.
- M. Lew, N. Sebe, C. Djeraba, and R. Jain. Content-based multimedia information retrieval: State of the art and challenges. *ACM Transaction on Multimedia Computing, Communications and Applications*, 2(1):1–19, February 2006.
- X. Li and Q. Ji. Active affective State detection and user assistance with dynamic bayesian networks. In *IEEE Transactions on Systems, Man , and Cybernetics. Part A: Systems and Humans*, volume 35, pages 93–105. IEEE, January 2005.
- Y. Li and Y. Zhao. Recognizing emotions in speech using short-term and long-term features. In *ICSLP '98, Proceedings of the 5th International Conference on Spoken Language Processing*, pages 2255–2258, Sydney, Australia, 1998.
- H. Liao. Multimodal Fusion. Master's thesis, University of Cambridge, July 2002.
- C. Lisetti and F. Nasoz. Using noninvasive wearable computers to recognize human emotions from physiological signals. *EURASIP Journal on Applied Signal Processing*, 11: 1672–1687, 2004.
- C. L. Lisetti and P. J. Gmytrasiewicz. Can a rational agent afford to be affectless? a formal approach. *Applied Artificial Intelligence*, 16(7-8):577–609, 2002.
-

- C. L. Lisetti and C. LeRouge. Affective computing in tele-home health: design science possibilities in recognition of adoption and diffusion issues. In *Proceedings 37th IEEE Hawaii International Conference on System Sciences*, volume 7, pages 348–363, Hawaii, USA, 2004.
- C. L. Lisetti and F. Nasoz. Maui: a multimodal affective user interface. In *MULTIMEDIA '02: Proceedings of the tenth ACM international conference on Multimedia*, pages 161–170, New York, NY, USA, 2002. ACM.
- C. L. Lisetti, S. Brown, and A. Marpaung. Cherry, the Little Red Robot with a Mission and a Personality. In *Working Notes of the AAAI Fall Symposium Series on Human–Robot–Interaction*, Menlo Park, CA, 2002. AAAI Press.
- X. Liu and T. Cheng. Video-based face recognition using adaptive hidden Markov models. In *IEEE Proceedings on Computer Vision and Pattern Recognition*, pages 340–345, June 2003.
- Loquendo. Loquendo website. <http://www.loquendo.com/>, 2009.
- B. D. Lucas and T. Kanade. An iterative image registration technique with an application to stereo vision. In *Proceedings of IJCAI '81, International Joint Conferences on Artificial Intelligence*, pages 674–679, Vancouver, Canada, August 1981.
- R. Luis-Garcia, C. Alberola-Lopez, O. Aghzout, and J. Ruiz-Alzola. Biometric identification systems. *Signal Processing*, 83:2539–2557, December 2003.
- M. Lyons, J. Budynek, and S. Akamatsu. Automatic classification of single facial images. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 21(12):1357–1362, December 1999.
- W. Mao and J. Gratch. A utility-based approach to intention recognition. In *In proceedings of AAMAS'04 Workshop on Agent Tracking: Modeling Other Agents from Observations*, 2004.
- G. Mark and P. DeFlorio. An experiment using life-size HDTV. In *Proceedings of IEEE WACE'01, International Workshop on Advanced Collaborative Environments*, 2001.
- G. Mark, S. Abrams, and N. Nassif. Group-to-Group Distance Collaboration: Examining the 'Space Between'. In *Proceedings of ECSCW'03, European Conference of Computer-Supported Cooperative Work*, pages 14–18, 2003.
- S. Marsella and J. Gratch. Modeling coping behavior in virtual humans: don't worry, be happy. In *AAMAS '03: Proceedings of the second international joint conference on Autonomous agents and multiagent systems*, pages 313–320, New York, NY, USA, 2003a. ACM.
- S. Marsella and J. Gratch. Modeling coping behavior in virtual humans: Don't worry, be happy. In *In proceedings of AAMAS'03, Second International conference on Autonomous Agents and MultiAgent Systems*, pages 313–320. ACM Press, 2003b.
- S. Marsella and J. Gratch. Ema: A process model of appraisal dynamics. *Cognitive Systems Research*, 10(1):70–90, March 2009.
-

- 
- S. Marsella, J. Gratch, and J. Rickel. *Expressive Behaviors for Virtual Worlds*, pages 317–360. Springer Cognitive Technologies Series, 2003.
- O. Martin, I. Kotsia, B. Macq, and I. Pitas. The enterface'05 audio-visual emotion database. In *ICDEW '06, Proceedings of the 22nd International Conference on Data Engineering Workshops*, pages 8–16, Washington, DC, USA, 2006. IEEE Computer Society.
- R. Martin. Is laughter the best medicine? Humor, laughter, and physical health. *Current Directions in Psychological Science*, 11:217–219, 2002.
- R. Martin. Sense of humor and physical health: Theoretical issues, recent findings and future directions. *Humor: International Journal of Humor Research*, 17:1–19, 2004.
- K. Mase. Recognition of Facial Expression from Optical Flow. *IEICE Transactions*, E74(10):3474–3483, 1991.
- C. Mathers, T. Boerma, and D. M. Fat. *The Global Burden of Disease: A Comprehensive Assessment of Mortality and Disability from Diseases, Injuries, and Risk Factors in 1990 and Projected to 2020*. Cambridge, World Health Organization Press, Geneva, CH, 1996.
- C. Mathers, T. Boerma, and D. M. Fat. *The Global Burden of Disease: 2004 update*. Cambridge, World Health Organization Press, Geneva, CH, 2008.
- F. Matta and J.-L. Dugelay. A behavioural approach to person recognition. In *ICME 2006, IEEE International Conference on Multimedia & Expo*, Toronto, Canada, July 2006.
- J. Mayer and P. Salovey. The intelligence of emotional intelligence. *Intelligence*, 17:433–442, 1993.
- W. McDougall. *An introduction to social psychology*. Luce, Boston, 1926.
- J. L. McGaugh and L. Cahill. *The Handbook of Affective Science*, chapter Emotion and Memory: Central and Peripheral Contributions, pages 93–116. Oxford University Press, New York, 2002.
- F. McPherson. The Role of Emotion in Memory. <http://www.memory-key.com/NatureofMemory/emotion.htm>, 2004.
- A. Mehrabian. *Silent Messages*. Wadsworth Publishing Company, Inc, Belmont, CA, 1971.
- A. Mehrabian. *Nonverbal Communication*. Aldine-Atherton, Chicago, 1972.
- A. Mehrabian and S. R. Ferris. Inference of attitudes from nonverbal communication in two channels. *Journal of consulting psychology*, 31(3):248–252, 1967.
- A. Mehrabian and M. Wiener. Decoding of inconsistent communications. *Journal of personality and social psychology*, 6(1):109–114, 1967.
- G. Merola and I. Poggi. Multimodality and Gestures in the Teacher s Communication. In *Lecture Notes in Computer Science*, volume 2915, pages 101–111, Feb 2004.
- M. Minsky. *The Society of Minds*. Simon & Schuster, March 1988.
-

- H. Miyamori, S. Nakamura, and K. Tanaka. Generation of views of TV content using TV viewers' perspectives expressed in live chats on the web. In *ACM MM'05 Proceedings of ACM International Conference on Multimedia*, pages 853–861, Singapore, 2005.
- D. Morrison and R. Wang. Real-time spoken affect classification and its application in call-centres. In *ICITA '05, Proceedings of the Third International Conference on Information Technology and Applications*, volume 2, pages 483–487, Washington, DC, USA, 2005. IEEE Computer Society.
- O. H. Mowrer. *Learning theory and behavior*. Wiley, New York, 1960.
- R. Nakatsu, J. Nicholson, and N. Tosa. Emotion recognition and its application to computer agents with spontaneous interactive capabilities. In *ACM MM '99: Proceedings of the seventh ACM international conference on Multimedia*, pages 343–351, New York, NY, USA, 1999. ACM.
- C. Nass and S. S. Is human-computer interaction social or parasocial? Technical Report 100, Stanford Communication Technologies Research Group, 1994.
- C. Nass, J. Steuer, and E. R. Tauber. Computers are social actors. In *Proceedings of CHI '94: Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 72–78, New York, NY, USA, 1994. ACM Press.
- N. Negroponte. *Being Digital*. Vintage, January 1995.
- D. Nguyen and J. Canny. MultiView: Improving Trust in Group Video Conferencing Through Spatial Faithfulness. In *Proceedings of CHI'07, International Conference on Computer Human Interactions*, pages 1465–1474, 2007.
- P. M. Niedenthal and S. Kitayama, editors. *Heart's Eye: Emotional Influences in Perception and Attention*. Academic Press, San Diego, 1994.
- J. Noble. *Spoken emotion recognition with support vector machines*. PhD thesis, University of Melbourne, November 2003.
- D. A. Norman. *Perspectives on Cognitive Science*, chapter Twelve Issues for Cognitive Science, pages 265–295. Erlbaum, Hillsdale, NJ, 1981.
- W. T. Norman. Toward an adequate taxonomy of personality attributes: replicated factors structure in peer nomination personality ratings. *Journal of abnormal and social psychology*, 66:574–583, June 1963.
- T. L. Nwe, F. S. Wei, and L. De Silva. Speech based emotion classification. In *TENCON'01, Proceedings of IEEE Region 10 International Conference on Electrical and Electronic Technology*, volume 1, pages 297–301, 2001.
- K. Oatley and P. N. Johnson-Laird. Towards a cognitive theory of emotions. *Cognition & Emotion*, 1:29–50, 1987.
- S. Ornager. The newspaper image database: empirical supported analysis of users' typology and word association clusters. In *SIGIR'95, Proceedings of the ACM Special Interest Group on Information Retrieval International Conference*, pages 212–218, 1995. Seattle, WA, USA.
-



- 
- A. Ortony and T. J. Turner. What's basic about basic emotions? *Psychological Review*, 97: 315–331, 1990.
- A. Ortony, G. Clore, and A. Collins. *The cognitive structure of emotions*. Cambridge University Press, Cambridge, UK, 1988.
- N. Otsu. A Threshold Selection Method from Gray-level Histograms. *IEEE Transactions on Systems, Man and Cybernetics*, 9(1):62–66, 1979.
- T. Otsuka and J. Ohya. Spotting segments displaying facial expression from image sequences using hmm. In *FG '98, Proceedings of the 3rd. International Conference on Face & Gesture Recognition*, pages 442–447, Washington, DC, USA, 1998. IEEE Computer Society.
- C. Padgett and G. Cottrell. Representing Face Images for Emotion Classification. In *Proceedings of Advances in Neural Information Processing Systems*, pages 894–900, 1996.
- M. Paleari. ALICIA: Modeling Affective Intelligent Agents. In *Master Report*, September 2005.
- M. Paleari and B. Huet. Toward emotion indexing of multimedia excerpts. In *CBMI 2008, 6th International Workshop on Content Based Multimedia Indexing, June, 18-20th 2008, London, UK, June 2008*.
- M. Paleari and C. L. Lisetti. Psychologically grounded avatars expressions. In *1st Workshop on Emotion and Computing at KI 2006, 29th Annual Conference on Artificial Intelligence, June, 14-19, 2006, Bremen, Germany, June 2006a*.
- M. Paleari and C. L. Lisetti. Avatar expressions with Scherer's theory. In *3rd HUMAINE EU Summer School, September 22-28, 2006, Genova, Italy, October 2006b*.
- M. Paleari and C. L. Lisetti. Toward multimodal fusion of affective cues. In *HCM 2006, 1st International Workshop in Human Centered Multimedia at ACM Multimedia 2006, October 23-27, 2006, Santa Barbara, USA, October 2006c*.
- M. Paleari and C. L. Lisetti. Agents for learning environments. In *Demo at 1st Workshop on Emotion and Computing at KI 2006, 29th Annual Conference on Artificial Intelligence, June, 14-19, 2006, Bremen, Germany, June 2006d*.
- M. Paleari, C. L. Lisetti, and M. Lethonen. VALERIE: a virtual agent for a learning environment, reacting and interacting emotionally. In *AIED 2005, 12th International Conference on Artificial Intelligence in Education, July 18-22, 2005, Amsterdam, The Netherlands, July 2005*.
- M. Paleari, B. Duffy, and B. Huet. Using emotions to tag media. In *Jamboree 2007: Workshop By and For KSpace PhD Students, September, 15th 2007, Berlin, Germany, September 2007a*.
- M. Paleari, A. Grizard, and C. L. Lisetti. Adapting psychologically grounded facial emotional expressions to different anthropomorphic embodiment platforms. In *FLAIRS 2007, 20th International Florida Artificial Intelligence Research Society Conference, May 7-9, 2007, Key West, USA, May 2007b*.
-

- M. Paleari, B. Huet, and B. Duffy. SAMMI, Semantic affect-enhanced multimedia indexing. In *SAMT 2007, 2nd International Conference on Semantic and Digital Media Technologies, 5-7 December 2007, Genoa, Italy* | Also published as *Lecture Notes in Computer Science Volume 4816*, December 2007c.
- M. Paleari, B. Huet, A. Schutz, and D. T. M. Slock. Audio-visual guitar transcription. In *Jamboree 2008 : Workshop By and For KSpace PhD Students, July, 25 2008, Paris, France*, July 2008a.
- M. Paleari, B. Huet, A. Schutz, and D. T. M. Slock. A multimodal approach to music transcription. In *1st ICIP Workshop on Multimedia Information Retrieval : New Trends and Challenges, October 12, 2008, San Diego, USA*, October 2008b.
- M. Paleari, R. Benmokhtar, and B. Huet. Evidence theory based multimodal emotion recognition. In *MMM 2009, 15th International MultiMedia Modeling Conference, January 7-9, 2008, Sophia Antipolis, France*, January 2009a.
- M. Paleari, V. Singh, B. Huet, and R. Jain. Toward Environment-to-Environment (E2E) Affective Sensitive Communication Systems. In *MTDL'09, Proceedings of the first ACM International Workshop on Multimedia Technologies for Distance Learning at ACM Multimedia, Beijing, China*, October 2009b. ACM.
- M. Paleari, C. Velardo, J.-L. Dugelay, and B. Huet. Face Dynamics for Biometric People Recognition. In *MMSP 2009, IEEE International Workshop on Multimedia Signal Processing, October 5-7, 2009, Rio de Janeiro, Brazil*, October 2009c.
- J. Panksepp. Toward a general psychobiological theory of emotions. *The Behavioral and Brain Sciences*, 5:407–467, 1982.
- M. Pantic. *Facial expression analysis by computational intelligence techniques*. PhD thesis, Delft University of Technology, October 2001.
- M. Pantic and L. J. M. Rothkrantz. Expert system for automatic analysis of facial expressions. *Image Vision Comput.*, 18(11):881–905, 2000a.
- M. Pantic and L. J. M. Rothkrantz. Automatic analysis of facial expressions: The state of the art. *IEEE Trans. Pattern Anal. Mach. Intell.*, 22(12):1424–1445, 2000b.
- M. Pantic and L. J. M. Rothkrantz. Toward an affect-sensitive multimodal human-computer interaction. *Proceedings of the IEEE*, 91(9):1370–1390, September 2003.
- E.-J. Park and J.-W. Lee. Emotion-based image retrieval using multiple-queries and consistency feedback. In *Proceedings of INDIN'08, the 6th IEEE International Conference on Industrial Informatics*, pages 1654–1659, 2008.
- W. Parrott. *Emotions in Social Psychology*. Psychology Press, Philadelphia, 2001.
- S. Paulmann and S. A. Kotz. Early emotional prosody perception based on different speaker voices. *Neuroreport*, 19(2):209–213, January 2008.
- C. Pelachaud, M. Jean-Claude, E. André, G. Chollet, K. Karpouzis, and D. Pelé, editors. *Intelligent Virtual Agents: 7th International Conference, IVA2007*, volume 4722 of *LNAI: Lecture Notes in Artificial Intelligence*. Springer, Paris, France, September 2007.
-



- 
- F. Perronnin, J.-L. Dugelay, and K. W. Rose. A probabilistic model of face mapping with local transformations and its application to person recognition. *PAMI'05, IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(7), July 2005.
- V. A. Petrushin. Emotion recognition in speech signal: Experimental study, development, and application. In *ICSLP '00, Proceedings of the 6th International Conference on Spoken Language Processing*, pages 222–225, 2000.
- P. J. Phillips, W. T. Scruggs, A. J. O'Toole, P. J. Flynn, K. W. Bowyer, C. L. Schott, and M. Sharpe. FRVT 2006 and ICE 2006 Large-Scale Experimental Results. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 99(1), 2007.
- R. W. Picard. *Affective Computing*. The MIT Press, Cambridge, September 1997.
- R. W. Picard, E. Vyzas, and J. Healey. Toward machine emotional intelligence: analysis of affective physiological state. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 23(10):1175–1191, October 2001.
- R. Plutchik. *Theories of emotion*, volume 1 of *Emotion: Theory, research, and experience*, chapter A general psychoevolutionary theory of emotion, pages 3–33. Academic, New York, 1980.
- I. Poggi, C. Pelachaud, F. de Rosis, V. Carofiglio, and B. de Carolis. *GRETA. A Believable Embodied Conversational Agent*, pages 27–45. Kluwer, 2005.
- T. S. Polzin. Verbal and non-verbal cues in the communication of emotions. In *ICASSP '00: Proceedings of the Acoustics, Speech, and Signal Processing, 2000. on IEEE International Conference*, pages 2429–2432, Washington, DC, USA, 2000. IEEE Computer Society.
- H. Prendinger and M. Ishizuka, editors. *Life-Like Characters : Tools, Affective Functions, and Applications (Cognitive Technologies)*. Springer, January 2004.
- H. Prendinger, J. Lester, and M. Ishizuka, editors. *Intelligent Virtual Agents: 8th International Conference, IVA2008*, volume 5208 of *LNAI: Lecture Notes in Artificial Intelligence*. Springer, Tokio, Japan, September 2008.
- L. R. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989.
- D. P. Radicioni, L. Anselma, and V. Lombardo. A Segmentation-Based Prototype to Compute String Instruments Fingering. In R. Parncutt, A. Kessler, and F. Zimmer, editors, *CIM04: Proceedings of the 1st Conference on Interdisciplinary Musicology*, Graz, Austria, 2004.
- A. S. Rao and M. P. Georgeff. Modeling Rational Agents within a BDI-Architecture. In Allen, Fikes, and Sandewall, editors, *in proceedings of KR'91 the 2nd International Conference on Principles of Knowledge Representation and Reasoning*, pages 473–484, San Mateo, CA, 1991. Morgan Kaufmann.
- A. S. Rao and M. P. Georgeff. BDI agents: From theory to practice. In *Proceedings of ICMAS'95 1st International Conference on Multi-Agent Systems*, page 312–319, San Francisco, CA., 1995.
-

- B. Reeves and C. Nass. *The Media Equation: How People Treat Computers, Television, and New Media Like Real People and Places*. Cambridge, [1st] edition, 1996.
- B. Reeves and C. Nass. *The Media Equation : How People Treat Computers, Television, and New Media Like Real People and Places*. CSLI Lecture Notes. Center for the Study of Language and Information, January 1998.
- A. Roseman, A. Antoniou, and P. Jose. Appraisal determinants of emotions: Constructing a more accurate and comprehensive theory. *Cognition and Emotion*, 10(3):241–277, 1996.
- J. A. Russell and A. Mehrabian. Evidence for a three-factor theory of emotions. *Journal of Research in Personality*, 11(3):273–294, September 1977.
- S. J. Russell and P. Norvig. *Artificial Intelligence: Modern Approach*. Prentice Hall, 1st edition, January 1995.
- Z. Ruttkay, M. Kipp, A. Nijholt, and H. H. Vilhjálmsson, editors. *Intelligent Virtual Agents: 9th International Conference, IVA2009*, volume 5773 of *LNAI: Lecture Notes in Artificial Intelligence*. Springer, Amsterdam, The Netherlands, September 2009.
- P. Saini, B. E. R. de Ruyter, P. Markopoulos, and A. J. N. van Breemen. Benefits of social intelligence in home dialogue systems. In *INTERACT'05, Proceedings of the 9th IFIP Conference in Human-Computer Interaction*, pages 510–521, 2005.
- P. Salovey and J. D. Mayer. Emotional Intelligence. *Imagination, Cognition, and Personality*, 9(3):185–211, 1990.
- A. Salway and M. Graham. Extracting information about emotions in films. In *ACM MM '03: Proceedings of the eleventh ACM international conference on Multimedia*, pages 299–302, New York, NY, USA, 2003. ACM.
- H. Sato, Y. Mitsukura, M. Fukumi, and N. Akamatsu. Emotional speech classification with prosodic parameters by using neural networks. In *Proceedings of The Seventh Australian and New Zealand Intelligent Information Systems Conference*, pages 395–398, November 2001.
- L. P. Schafer, B. Gilmer, and M. Schoen. *Psychology*. Haper and Brothers, New York, 1940.
- K. R. Scherer. Emotion as a process: Function, origin and regulation. *Social Science Information*, 21:555–570, 1982.
- K. R. Scherer. *Approaches to Emotions*, chapter On the Nature and Function of Emotion: A Component Process Approach, pages 293–317. Lawrence Erlbaum Associates, 1984.
- K. R. Scherer. Emotions can be rational. *Social Science Information*, 24(2):331–335, 1985.
- K. R. Scherer. Toward a dynamic theory of emotion: The component process model of affective states. *Geneva Studies in Emotion and Communication*, 1:1–98, 1987.
- K. R. Scherer. *Appraisal processes in emotion: Theory, methods, research*, chapter Appraisal Considered as a Process of Multilevel Sequential Checking, pages 92–120. Oxford University Press, 2001.
-

- 
- K. R. Scherer. Vocal communication of emotion: a review of research paradigms. *Speech Communications*, 40(1-2):227–256, 2003.
- K. R. Scherer and T. Bänziger. Emotional Expression in Prosody: A Review and an Agenda for Future Research. In *Speech Prosody '04*, pages 359–366, Nara, Japan, March 2004.
- K. R. Scherer and P. Ekman. *Approaches to Emotions*. Lawrence Erlbaum Associates, 1984.
- K. R. Scherer, T. Johnstone, and T. Bänziger. Automatic verification of emotionally stressed speakers: The problem of individual differences. In *Proceedings of the International Workshop on Speech and Computer*, St. Petersburg, 1998.
- M. Schröder and J. Trouvain. The german text-to-speech synthesis system mary: A tool for research, development and teaching. In *International Journal of Speech Technology*, pages 365–377, 2003.
- N. Sebe, M. S. Lew, I. Cohen, A. Garg, and T. S. Huang. Emotion recognition using a cauchy naive bayes classifier. In *ICPR'02, Proceedings of the 12th International Conference on Pattern Recognition*, volume 1, pages 17–20, 2002.
- N. Sebe, M. Lew, X. Zhou, T. Huang, and E. Bakker. The State of the Art in Image and Video Retrieval. In Springer, editor, *Image and Video Retrieval*, volume 2728 of *Lecture Notes in Computer Science*, pages 7–12, 2003.
- N. Sebe, E. R. Bakker, I. Cohen, T. Gevers, and T. S. Huang. Bimodal Emotion Recognition. In *Proceedings of the 5th International Conference on Methods and Techniques in Behavioral Research*, Wageningen, The Netherlands, August 2005a.
- N. Sebe, I. Cohen, T. Gevers, and T. S. Huang. Multimodal approaches for emotion recognition: a survey. *Internet Imaging VI*, 5670(1):56–67, 2005b.
- N. Sebe, I. Cohen, and T. Huang. *Handbook of Pattern Recognition and Computer Vision*, chapter Multimodal emotion recognition, pages 387–410. World Scientific, 3rd edition edition, January 2005c.
- A. Sellen, B. Buxton, and J. Arnott. Using spatial cues to improve videoconferencing. In *Proceedings of CHI'92, International Conference on Computer Human Interactions*, pages 651–652, 1992.
- X. Serra. Musical sound modeling with sinusoids plus noise. In C. Roads, S. Pope, A. Piccialli, and G. De Poli, editors, *Musical Signal Processing*, pages 91–122, Lisse, the Netherlands, 1997. Swets & Zeitlinger Publishers.
- M.-K. Shan, F.-F. Kuo, M.-F. Chiang, and S.-Y. Lee. Emotion-based music recommendation by affinity discovery from film music. *Expert System Applications*, 36(4):7666–7674, 2009.
- R. Sharma, V. Pavlovic, and T. Huang. Toward multimodal human-computer interface. *Proceedings of the IEEE*, 86(5):853–869, May 1998.
- J. Shi and C. Tomasi. Good features to track. In *Proceedings of CVPR'94 IEEE International Conference on Computer Vision and Pattern Recognition*, pages 593–600, Seattle, WA, June 1994.
-

- H. A. Simon. Motivational and emotional controls of cognition. *Psychological Review*, 74 (1):29–39, January 1967.
- S. E. Siwek. *Video Games in the 21st Century: Economic Contributions of the US Entertainment Software Industry*. Entertainment Software Association, 2007.
- A. S. M. Sohail and P. Bhattacharya. *Signal Processing for Image Enhancement and Multimedia Processing*, volume 31 of *Multimedia Systems and Applications Series*, chapter Detection of Facial Feature Points Using Anthropometric Face Model, pages 189–200. Springer US, December 2007.
- E. Sony Entertainment Robots. Aibo ers-7 illume-face expresions. <http://support.sony-europe.com/aibo/downloads/en/ledface.pdf>, March 2005.
- E. SONY Entertainment Robots. Sony aibo europe. <http://support.sony-europe.com/aibo/>, 2006.
- H. J. M. Steeneken and J. H. L. Hansen. Speech under stress conditions: overview of the effect on speech production and on system performance. In *ICASSP '99: Proceedings of the Acoustics, Speech, and Signal Processing, 1999. on 1999 IEEE International Conference*, pages 2079–2082, Washington, DC, USA, 1999. IEEE Computer Society.
- K. R. Thorisson and J. Cassell. Why Put an Agent in a Human Body: The Importance of Communicative Feedback in Human-Humanoid Dialogue. In *Proceedings of Lifelike Computer Characters*, pages 44–45, Snowbird, Utah, 1996.
- B. Tiddeman. Face detection demo. <http://morph.cs.st-andrews.ac.uk/fof/haarDemo/index.html>, August 2007.
- C. Tomasi and T. Kanade. Detection and Tracking of Point Features. Technical report, Carnegie Mellon University, April 1991. CMU-CS-91-132.
- S. S. Tomkins. *Approaches to emotion*, chapter Affect Theory, pages 163–195. Erlbaum, Hillsdale, NJ, 1984.
- N. Tosa and R. Nakatsu. Life-like communication agent -emotion sensing character "mic" & feeling session character "muse"-. *Multimedia Computing and Systems, International Conference on*, 0:12–19, 1996.
- C. Traube. *A Interdisciplinary Study of the Timbre of th Classical Guitar*. PhD thesis, McGill University, 2004.
- M. Turk and A. Pentland. Eigenfaces for Recognition. *Journal of Cognitive Neuroscience*, 3 (1):71–86, 1991.
- R. Valenti, N. Sebe, and T. Gevers. Facial Expression Recognition: A Fully Integrated Approach. In *ICIAPW'07, Proceedings of the 14th International Conference of Image Analysis and Processing - Workshops*, pages 125–130, Washington, DC, USA, 2007. IEEE Computer Society.
- M. F. Valstar, H. Gunes, and M. Pantic. How to distinguish posed from spontaneous smiles using geometric features. In *ICMI '07: Proceedings of the 9th international conference on Multimodal interfaces*, pages 38–45, Nagoya, Aichi, Japan, 2007. ACM.
-

- 
- A. van Breemen. Animation engine for believable interactive user-interface robots. In *IROS 2004, Proceedings of the 2004 IEEE/RSJ International Conference on Intelligent Robots and Systems*, volume 3, pages 2873–2878, Sendai, Japan, September 2004a.
- A. van Breemen. Binging robots to life: Applying principles of animation to robots. In *Proceedings of Shapping Human-Robot Interaction workshop held at CHI 2004*, pages 143–144, Vienna, Austria, 2004b.
- A. van Breemen. icat: Experimenting with animabotics. In *AISB'05, Proceedings of the 2005 Artificial Intelligence and the Simulation of Behaviour Convention*, pages 27–32, 2005.
- J. A. Verner. Midi Guitar Synthesis: Yesterday, Today and Tomorrow. *Recording Magazine*, 8 (9):52–57, 1995.
- R. Vertegaal, G. V. der Veer, and H. Vons. Effects of gaze on multiparty mediated communication. In *Proceedings of Graphics Interface*, pages 95–102, 2000.
- O. Villon. *Modeling affective evaluation of multimedia contents: user models to associate subjective experience, physiological expression and contents description*. PhD thesis, Thesis, December 2007.
- P. Viola and M. Jones. Robust Real-time Object Detection. *International Journal of Computer Vision*, 2001.
- D. Vukadinovic and M. Pantic. Fully Automatic Facial Feature Point Detection Using Gabor Feature Based Boosted Classifiers. In *Proceedings of IEEE SMC '05 International Conference on Systems, Man, and Cybernetics*, pages 1692–1698, Waikoloa, Hawaii, October 2005. IEEE.
- W3stfa11. Video-games-sales wiki. <http://vgsales.wikia.com>, 2009.
- J. Wagner, T. Vogt, and E. André. A systematic comparison of different hmm designs for emotion recognition from acted and spontaneous speech. In *ACII '07: Proceedings of the 2nd international conference on Affective Computing and Intelligent Interaction*, pages 114–125, Berlin, Heidelberg, 2007. Springer-Verlag.
- M. Wang, Y. Iwai, and M. Yachida. Expression recognition from time-sequential facial images by use of expression change model. In *FG '98, Proceedings of the 3rd. International Conference on Face & Gesture Recognition*, pages 324–329, Washington, DC, USA, 1998. IEEE Computer Society.
- J. B. Watson. *Behaviorism*. University of Chicago Press, Chicago, 1930.
- T. Wehrle and K. R. Scherer. *Appraisal Processes in emotion: Theory, methods, research*, chapter Toward Computational Modeling of Appraisal Theories, pages 350–365. Oxford University Press, New York, 2001.
- T. Wehrle, S. Kaiser, S. Schmidt, and K. R. Scherer. Studying the dynamics of emotional expression using synthesized facial muscle movements. *Journal of Personality and Social Psychology*, 78:105–119, 2000.
- B. Weiner and S. Graham. *Emotions, cognition, and behavior*, chapter An attributional approach to emotional development, pages 167–191. Cambridge University Press, 1984.
-



- J. Whitehill and C. W. Omlin. Haar features for face recognition. In *FGR '06, Proceedings of the 7th International Conference on Automatic Face and Gesture Recognition*, pages 97–101, Washington, DC, USA, 2006. IEEE Computer Society.
- L. Wiskott, J. Fellous, N. Kruger, and C. Malsburg. Face recognition by elastic bunch graph matching. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19:775–779, July 1997.
- Y. Wu and T. S. Huang. Vision-based gesture recognition: A review. In *GW '99, Proceedings of the 11th International Gesture Workshop on Gesture-Based Communication in Human-Computer Interaction*, volume 1739 of *Lecture Notes in Artificial Intelligence*, pages 103–115, London, UK, 1999. Springer-Verlag.
- M. Yeasin, B. Bulot, and R. Sharma. Recognition of facial expressions and measurement of levels of interest from video. *IEEE Transactions on Multimedia*, 8(3):500–508, 2006.
- R. M. Yerkes and J. D. Dodson. The relation of strength of stimulus to rapidity of habit-formation. *Journal of Comparative Neurology and Psychology*, 18:459–482, 1908.
- Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang. A survey of affect recognition methods: audio, visual and spontaneous expressions. In D. W. Massaro, K. Takeda, D. Roy, and A. Potamianos, editors, *ICMI'07, Proceedings of the 9th International Conference on Multimodal Interfaces*, pages 126–133, Nagoya, Aichi, Japan, November 2007.
- Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang. A survey of affect recognition methods: Audio, visual, and spontaneous expressions. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 31(1):39–58, 2009.
- B. Zhang, J. Zhu, Y. Wang, and W. K. Leow. Visual Analysis of Fingering for Pedagogical Violin Transcription. In *MM '07: Proceedings of the 15th international conference on Multimedia*, pages 521 – 524, Augsburg, Germany, 2007. ACM.
- Z. Zhang, M. Lyons, M. Schuster, and S. Akamatsu. Comparison between geometry-based and gabor-wavelets-based facial expression recognition using multi-layer perceptron. In *FG '98, Proceedings of the 3rd. International Conference on Face & Gesture Recognition*, pages 454–459, Washington, DC, USA, 1998. IEEE Computer Society.
- L. Zhao, W. Lu, Y. Jiang, and Z. Wu. A study on emotional feature recognition in speech. In *ICSLP '00, Proceedings of the 6th International Conference on Spoken Language Processing*, pages 961–964, 2000.
- W. Zhao, R. Chellappa, P. J. Phillips, and A. Rosenfeld. Face recognition: A literature survey. *ACM Computer Surveys (CSUR)*, 35(4):399–458, 2003.
- E. Zovato, A. Pacchiotti, S. Quazza, and S. Sandri. Towards Emotional Speech Synthesis: A Rule Based Approach. In *5th International Speech Communication Association (ISCA) Speech Synthesis Workshop*, pages 219–220, 2004.
-