Evidence Theory-Based Multimodal Emotion Recognition

Paleari Marco, Rachid Benmokhtar, and Benoit Huet

EURECOM
2229, route des Crtes,
Sophia Antipolis, FRANCE
{paleari,benmokhtar,huet} @ eurecom.fr

Abstract. Automatic recognition of human affective states is still a largely unexplored and challenging topic. Even more issues arise when dealing with variable quality of the inputs or aiming for real-time, unconstrained, and person independent scenarios. In this paper, we explore audio-visual multimodal emotion recognition. We present SAMMI, a framework designed to extract real-time emotion appraisals from non-prototypical, person independent, facial expressions and vocal prosody. Different probabilistic method for fusion are compared and evaluated with a novel fusion technique called NNET. Results shows that NNET can improve the recognition score (CR$^+$) of about 19% and the mean average precision of about 30% with respect to the best unimodal system.

1 Introduction

It is commonly accepted that in most media, human communications forms, and notably in art expressions, emotions represent a valuable source of information. Changes in people's affective state play a significant role in perception and decision making during human to human interaction. Several studies have, therefore, been investigating how to automatically extract and use the affective information in everyday Human Computer Interactions (HCI) scenarios [1].

The ability to recognize and track the user affective state has the potential to enable a computing system to initiate interactions with the user based on changes in the perceived affective state rather than to simply respond to commands. Furthermore emotions can be used as valuable characteristics to dynamically tag media for future retrieval or summarization which may help to bridge the semantic gap.

In this paper, we explore audio-visual multimodal emotion recognition of acted emotions and a novel fusion technique [2] called NNET which bases on Evidence Theory.

Related works usually focus on three main modalities for the automatic recognition of the affective states; these are:

1. **Physiology**: The affective state is appraised through the modulations emotions exert to the Autonomous Nervous System (ANS). Signals such as heath

beat or skin conductivity are detected through ad hoc input devices. The estimation can be very reliable [3, 4] and it is less sensitive to the acting of emotions than the one extracted from the auditory and visual modalities. The main limitation is related to the intrusiveness of the sensing devices which make this modality impracticable for most HCI scenarios.

2. **Visual**: The affective state is evaluated as a function of the modulations of emotions on facial expressions, gestures, postures, and generally body language. The data are captured through a camera, allowing for non-intrusive system configurations. The systems are generally very sensitive to the video quality both in term of Signal to Noise Ratio (SNR) and in term of illumination, pose, and size of the face on the video and is the most sensitive to false, acted facial expressions. Most of the works use facial expressions [5–8]; only few use gesture, posture, or combinations of gestures with facial expressions [9].
3. **Auditory**: The affective state can finally be estimated as a modulation of the vocal signal. In this case data are captured through a microphone, once again, allowing for non intrusive system configurations [10]. The estimation can be very accurate. The processing needs clean voice data; SNR inferior to 10 dB can severely reduce the quality of the estimation [11]. Furthermore the processing still cannot handle the presence of more than one voice in the audio stream.

Only few works have investigated the possibility to fuse together visual and auditory affective estimation [12, 13, 11]. Most of them only did person dependent affect recognition. Some of them took into account non realistic scenarios with people having dots on their faces to enhance the tracking of the facial movements and none of them, to our knowledge, took into account low or variable quality videos. Furthermore we could not retrieve information about the computational cost of the algorithms involved and in particular we were not able to discern the systems working in real time from the others.

In our scenario a person should be able to sit in front of her computer and have her affective status appraised in real time starting from the video with audio retrieved from a standard web-cam. The video quality is therefore much lower than the one currently used for research, as well as the audio. Furthermore, even though, person dependent evolution of the training set will be available we would like the system to work with any person sitting in front of the computer requiring therefore person in-dependency. In this kind of scenario we also have to foresee issues with the input signals, for example in the case more people are speaking at once or no ambient light is present to enlighten the user face.

In this paper we present SAMMI [14, 15], a system which uses multimodality to overcome modality dependent signal issues and to improve the accuracy of the recognition while in presence of both signals together with different classifier fusion techniques. We, then, present NNET, a tool presented in [2], which uses the Evidence Theory to extract reliable classifications. We used such a system for fusing together the results from the unimodal classifiers; we detail and comment the results of these experiments.

2 eNTERFACE'05

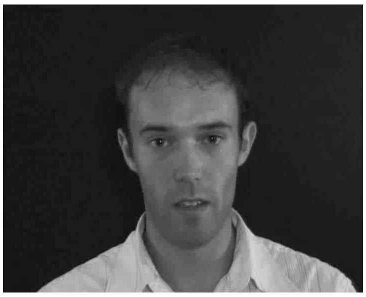

Fig. 1. Example frame.

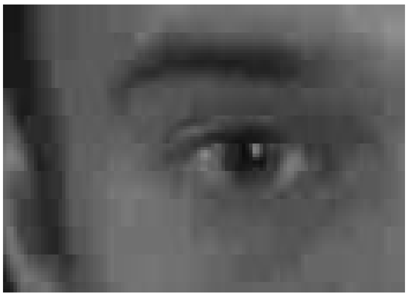

Fig. 2. Zoom on the eye.

One characteristic which seriously affect the performance of an emotion recognition system is related to the quality of the data used for training and testing. In realistic scenarios a system cannot deal with all the kind of data. Having a training database including samples as similar as possible to the true scenario is therefore crucial.

A part from the quality of the samples (both in term of SNR, compression, illumination, etc. and in term of the quality of the acting/spontaneous emotions expressions) a factor which obviously influences the results is the number of considered emotions. Indeed, some databases which were used in literature only consists of 2-3 emotions, some arrive at 16.

Unfortunately, the researching community still lack of available good quality multimodal audio-visual databases. We based our research on the eNTERFACE'05 database. The eNTERFACE database [16] is a publicly available multimodal audio-visual emotion database containing videos of subjects coming from 14 different nationalities.

The base contains 44 subjects presenting the 6 "universal" human emotions (anger, disgust, fear, happiness, sadness, and surprise), through 5 different sentences. The average video length is about 3 seconds summing up to 1320 shots and more than one hour of video. Videos are recorded in a Lab environment: Subjects are in frontal view with studio lightening condition and gray uniform background (see Fig. 1). Audio is recorded with a high quality microphone placed at around 30 cm from the subject mouth. All experiments were driven in English although only about 14% of the subjects were native English speaker. Subjects were not professional actors.

We have previously [15] evaluated the quality of this database pointing out some weaknesses of the base mainly related to compression issues (See Fig. 2) and to the quality of the emotional expression acted by the 44 subjects. In particular, one characteristic influencing our results is the fact that the database has one single tag per shot. Our scenarios often need a real-time, frame-to-frame, evaluation of the performed emotion; we are, therefore, trying to extract information that the one that it is given to us. Furthermore, the lack of neutral samples in the database cause a not irrelevant number of “transitional”/neutral frames to be tagged as emotionally relevant and used for training. An ad-hoc system built up to extract one single evaluation per shot would easily improve the recognition score of our system.

In [15] we also point out how some of these peculiarities motivate us to develop algorithms which should be robust in realistic scenarios.

3 SAMMI

We have overviewed the eNTERFACE bimodal database. In this section, we detail SAMMI: Semantic Affect-enhanced MultiMedia Indexing, a framework explicitly designed for extracting reliable real-time emotional information through multimodal fusion of affective cues.

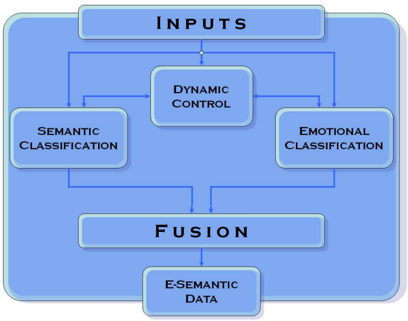

Fig. 3. SAMMI’s architecture.

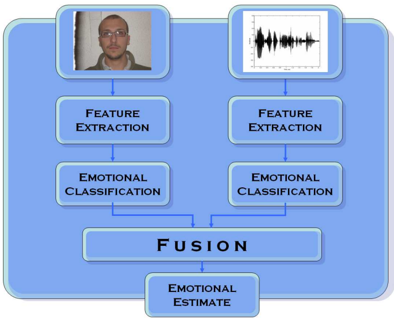

Fig. 4. Multimodal emotion recognition.

The main objective of SAMMI is to emotionally tag video. In this contest, emotions can play an important role, but we claim that they cannot play a sole role and should be coupled with other semantic tag to build effective HCIs. SAMMI (see Fig. 3) does this by coupling affective information with other content information extracted with state of the art techniques [17, 18].

A module called “Dynamic Control” in Fig. 3 is committed to adapt the various fusion algorithms and content based concept extractors to the quality of the signals in input. For example, if the sound quality is detected to be low then, the relevance of the vocal emotional estimation with respect to the one of the video emotional estimation will be reduced. This is very important in order to make the system more reliable and to loose some constraints.

For the purpose of this paper, we will concentrate on the modules committed to extract the emotional appraisal. These modules work on two inputs: The video for the facial expressions and the audio for the prosodic vocal expressions.

3.1 Feature Extraction

Facial Expressions: For our system we have chosen to base our system on facial feature points (FP) (see Fig. 5) and tested two approaches: The first based on the facial FP absolute movements and the second based on relative movements of couples of facial FP[1].

We want SAMMI to run in real-time, setting a constraint in term of complexity of the algorithms we could use. We are therefore using the Tomasi implementation of the Lukas Kanade (LK) algorithm which is embedded in the Intel OpenCV library [19].

We have built a computationally cheap feature point detector for the first frame. This module works as follows: It analyzes the first frame and detect the position of the face, of the eyes and of the mouth with three classifiers based on Haar-features; it detects 12 FP regions by applying a 2D face template to the found face; for each found region we search points[2] which will be tracked with the Lukas Kanade algorithm; finally for each region it computes the center of mass of the points belonging to that particular region finally obtaining the coordinates of 12 points which will be used for the emotional estimation.

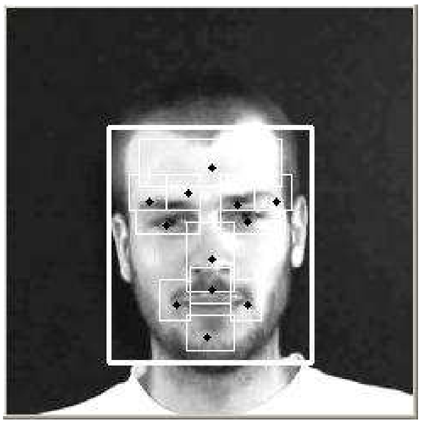

Fig. 5. Facial Feature Points.

This algorithm presents some limitations. Firstly, because the Haar-based classifier demonstrates not to be precise enough. In particular, even though the face is very often correctly recognized, the surrounding bounding box can sensibly change between two consecutive frames. We tried to partially overcome this

[1] For example, we take into account the openness of the mouth or the shape of the eyebrows simply by computing the distance between the top and bottom point of the mouth, or by looking at the relative vertical position of two points on the eyebrows.

[2] The number of these points may vary according to lightening conditions, on the particular region, and on the subject. We imposed a maximum at 50 points per region.

issue by computing multiple positions in subsequent frames and by intelligently choosing the most probable bounding box; this approach does nevertheless make us lose some information about the first frames.

Secondly, our 2D model works with scaling/zooming of the face (the Haar-based classifiers deal with this transformation), and, thanks to the fact that the LK points tend to attach to the position with big intensity changes in the luminance of the image, it also deals with small rotations around the z axis (i.e. the axis perpendicular to the screen plane) and x axis (i.e. the horizontal axis) but does not work properly for the rotation around the y axis (i.e. the vertical one). A 3D model will be more appropriate but it will also need some more computational power to be computed. Further works will investigate the possibility to apply such a model.

Thirdly, although in average the founded points should follow the movements inside the feature point regions, the estimation of the movement may not be precise enough. A different scheme for FP extraction is being currently tested in our laboratory which directly computes the position of the relevant FPs by analyzing the vertical and horizontal histograms of the mouth and eyes regions [20] and by applying some simple morphological operations [21].

Vocal Expressions: Our approach to the extraction of affective appraisals from the voice is similar to the one described by Noble in [10]. The approach bases on the PRAAT open source software [22]. Through PRAAT we extract the fundamental frequency f_0, the first 5 formants f_1, f_2, f_3, f_4, f_5, the signal intensity, the harmonicity (i.e. the degree of acoustic periodicity, also called Harmonics-to-Noise Ratio), ten Mel-Frequency Cepstral Coefficients (MFCC) and ten Linear Predictive Coefficients (LPC).

This approach has one main limitation; indeed the same information about the spectral envelope of the vocal tract is represented three different times with three different methods. The quantity of overlapping information is therefore significant. Although SVM should be able to handle big feature vectors, it is often better to reduce the size of the training vector to reduce the complexity of the model.

3.2 Emotional Classification

Two different classifiers are currently used for testing which are Support Vector Machines (SVM) and Neural Networks (NN).

We want to take into account the temporal information brought by the evolution of the video and audio signals to the emotions. For this purpose, we temporally window the signals (the FP positions or the audio features). One second windows are currently being used for every signal. For each windowed signal we compute two models: One statistical evaluation based on mean value, standard deviation, variance, 5 quantiles, minimum, and maximum (with the relative positions inside the 1 second window); and one polynomial evaluation based on a first grade polynomial regression coupled with the number of the zero crossings.

3.3 Classifier Fusion

Classifier 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 [23].

SAMMI 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. From the other side classification boosting, with too many classifiers, it is not an available option.

Multiple classifier fusion strategies have been tested and evaluated including Max, Vote, Mean, Bayesian combination and the new NNET which will be discussed in the following section.

4 Evidence Theory

Probabilistic methods have an inherent limitation: Most treat imprecision but ignore the uncertainty and ambiguity of the system (information to be fused). Evidence theory allows dealing with the uncertain data.

4.1 Applications to Fusion

The objective is to associate for each object x (frame), one class from the set of classes $\Omega = \{w_1, .., w_M\}$. This association is given via a set of training of N samples. Each sample can be considered as a part of belief for one class of Ω. This belief degree can be assimilated to evidence function m^i, with 2 focal elements: The class of x^i noted w_q, and Ω. So, if we consider that the object x^i is near to x, then a part of belief can be affected to w_q and the rest to Ω. The mass function is obtained by decreasing function of distance as follow:

$$\begin{cases} m^i(\{w_q\}) = \alpha^i \phi_q(d^i) \\ m^i(\Omega) = 1 - \alpha^i \phi_q(d^i) \end{cases} \tag{1}$$

Where $\phi(.)$ is a monotonically decreasing function such as an exponential function $\phi_q(d^i) = \exp\left(-\gamma_q (d^i)^2\right)$, and d^i is an Euclidean distance between the vector x and the i^{th} vector of training base. $0 < \alpha < 1$ is a constant which prevents a total affectation of mass to the class w_q when x and i^{th} samples are equal. γ_q is a positive parameter defining the decreasing speed of mass function. A method for optimizing parameters (α, γ_q) has been described in [24].

We obtain N mass functions, which can be combined into a single one using the equation (Eq. 2).

$$m(A) = (m^1 \oplus ... \oplus m^N) = \sum_{(B_1 \bigcap ... \bigcap B_N) = A} \prod_{i=1}^{N} m^i(B_i) \tag{2}$$

4.2 Neural Network based on Evidence Theory

We propose to resume work already made with the evidence theory in the connectionist implementation [2, 24], and to adapt it to classifier fusion. For this aim, an improved version of Radial Basis Function neural network based on evidence theory [2] which we call NNET, with one input layer L_{input}, two hidden layers L_2 and L_3 and one output layer L_{output} (Fig. 6) has been devised. Each layer corresponds to one step of the procedure as briefly described below:

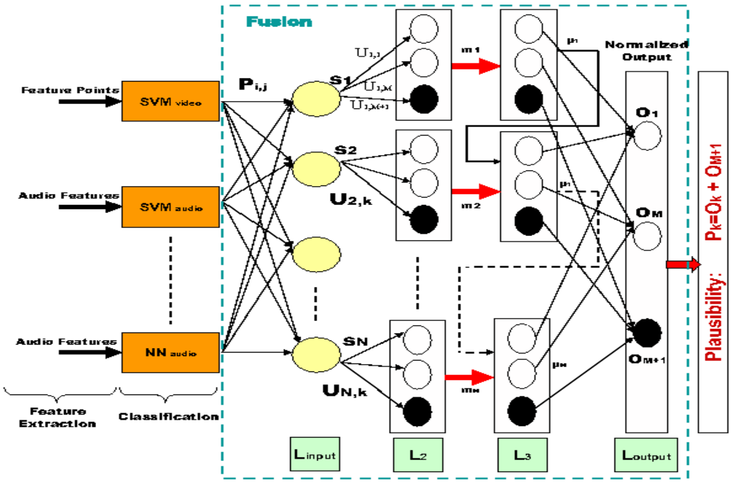

Fig. 6. NNET classifier fusion structure.

Layer L_{input}: Contains N units (prototypes). It is identical to an RBF network input layer with an exponential activation function ϕ. d is a distance computed using training data and dictionary created (clustering method). K-means is applied on the training data in order to create a ”visual” dictionary of the frames.

$$s^i = \alpha^i \phi(d^i) \tag{3}$$

Layer L_2: Computes the belief masses m^i (Equ. 4) associated to each prototype. It is composed of N modules of $M+1$ units each (Equ. 5). The units of module i are connected to neuron i of the previous layer. Note that each frame can belong to only one class.

$$\begin{cases} m^i(\{w_q\}) = \alpha^i u^i_q \phi(d^i) \\ m^i(\Omega) = 1 - \alpha^i \phi(d^i) \end{cases} \tag{4}$$

$$\begin{aligned} m^i &= (m^i(\{w_1\}), ..., m^i(\{w_{M+1}\})) \\ &= (u_1^i s^i, ..., u_M^i s^i, 1 - s^i) \end{aligned} \tag{5}$$

where u_q^i is the membership degree to each class w_q, q class index $q = \{1, ..., M\}$.

Layer $\boldsymbol{L_3}$: The Dempster-Shafer combination rule combines N different mass functions in one single mass. It's given by the conjunctive combination (Eq. 2). For this aim, the activation vector $\overrightarrow{\mu^i}$ can be recursively computed using the following formula:

$$\begin{cases} \mu^1 = m^1 \\ \mu_j^i = \mu_j^{i-1} m_j^i + \mu_j^{i-1} m_{M+1}^i + \mu_{M+1}^{i-1} m_j^i \\ \mu_{M+1}^i = \mu_{M+1}^{i-1} m_{M+1}^i \end{cases} \tag{6}$$

Layer $\boldsymbol{L_{output}}$: In [24], the output is directly obtained by $O_j = \mu_j^N$. The experiments show that this output is very sensitive to the number of prototype, where for each iteration, the output is purely an addition of ignorance. Also, we notice that a small change in the number of prototype can change the classifier fusion behavior. To resolve this problem, we use normalized output (Eq. 7). Here, the output is computed taking into account the activation vectors of all prototypes to decrease the effect of an eventual bad behavior of prototype in the mass computation.

$$O_j = \frac{\sum_{i=1}^{N} \mu_j^i}{\sum_{i=1}^{N} \sum_{j=1}^{M+1} \mu_j^i} \tag{7}$$

The different parameters (Δu, $\Delta \gamma$, $\Delta \alpha$, ΔP, Δs) can be determined by gradient descent of output error for an input pattern x. Finally, the maximum of plausibility P_q of each class w_q is computed.

$$P_q = O_q + O_{M+1} \tag{8}$$

5 Results

In this section, we will see how this novel approach performs on our database and we explain the results. We used roughly 60% of the data (i.e. 912 shots) for training the classifiers and the NNET and the remaining (i.e. 378 shots) for the evaluation. The results are obtained by testing the system on the remaining 5 subjects (the train base is rather unbalanced presenting 20% samples for the emotion anger, 18% for disgust, 14% for fear, 14% for happiness, 19% for sadness, and finally 13% for surprise).

The performance has been measured using the standard precision and recall metrics, in particular the Mean Average Precision (MAP) for the first 33% of the responses and the Positive Classification Rate (CR$^+$).

	Anger	Disgust	Fear	Happiness	Sadness	Surprise	**CR$^+$**	**MAP**
Video NN	0.420	0.366	0.131	0.549	0.482	0.204	0.321	0.205
Audio NN	0.547	0.320	0.151	0.496	0.576	0.169	0.354	0.234
Video SVM	0.342	0.342	0.193	0.592	0.426	0.244	0.320	0.211
Audio SVM	0.627	0.220	0.131	0.576	0.522	0.162	0.361	0.253
Max	0.612	0.378	0.120	0.619	0.586	0.185	0.384	0.260
Vote	**0.666**	0.422	0.142	0.622	0.495	0.161	0.391	0.296
Mean	0.635	0.406	0.150	0.721	0.600	0.206	0.415	0.331
Bayesian	0.655	**0.440**	0.159	**0.743**	0.576	0.235	**0.430**	0.335
NNET	0.542	0.388	**0.224**	0.633	**0.619**	**0.340**	0.428	**0.337**

Table 1. Classification accuracy for all emotions.

In the Table 1, we compare the results obtained from the NNET classifier fusion with the NN and SVM classifier outputs as well as with the Max, Vote, Mean, Beyesian combination fusion systems.

Firstly, we notice that Bayesian combination and NNET approaches outperform other systems in term of CR$^+$ and MAP. Both systems improve the CR$^+$ of a relative 19% and the MAP of a relative 32% with respect to the best unimodal system.

Secondly, some interesting points can be discussed regarding the performances per concept:

- NNET improves the CR$^+$ of the emotions which are usually classified with the worst scores (i.e. fear and surprise). This phenomenon can be explained by the positive impact of the evidence theory in the conflicting situations, where the incertitude is taken into account.
- A less positive impact can be observed for the NNET on some other emotions due to the limitation of our trains set when classification rates of the unimodal systems are already good (e.g. anger or happiness).
- The Bayesian combination fusion system achieves better improvements on the emotions which are usually better recognized by the unimodal systems (e.g. anger or happiness). This is normal because the product between the evaluations, intrinsic in the Bayesian combination returns much higher results for the emotions which are consistently recognized by the unimodal systems.

These results give us the impression that the NNET tool could be used in some scenarios when the objective is to maintain the average classification score while reducing the margin between the best recognized and the worst recognized emotions.

Indeed, NNET demonstrates to be a valuable tool, when dealing with incertitude and specifically when wanting to discern positive samples from negative samples; if for example we had in our testing base neutral samples, then NNET will probably do a good job in cutting these samples out. In our particular case, we observe that the Evidence Theory characteristics, boosting up results when multiple concepts are concurrently present in one scene, cannot be exploited.

NNET presents, nevertheless, some drawbacks: Firstly, it needs a training step which is not demanded by the other classifier fusion techniques discussed in this paper; Secondly, NNET are more computationally expensive than the other fusion techniques.

6 Concluding Remarks

In this paper, we have presented SAMMI, a framework for semantic affect-enhanced multimodal indexing of multimedia excerpts, and NNET, a classifier fusion technique based on the evidence theory. We have therefore seen how these two systems works together and we have commented the results.

NNET improves the CR^+ of about the 19% and the MAP of about the 32% with respect to the best unimodal classification system. Average results are similar to the one obtained by a classifier fusion system based on the Bayesian combination, but we have discussed how the evidence theory allows to improve the score of the worst recognized emotions thus reducing the gap between the best and the worst recognized emotions.

Future works will investigate new classification method such as Gaussian Mixture Models (GMM) and Hidden Markov Model (HMM). HMMs are very promising in the case of emotions since it allows us to better take into account the temporal information of the affective expressions [25].

A dimensionality reduction techniques such as Linear Discriminant Analysis (LDA), Principal Component Analysis (PCA), or Non-negative Matrix Factorization (NMF) will be tested, to reduce the size of the feature vectors used as input of the classifiers [26].

References

1. Picard, R.: Affective Computing. MIT Press, Cambridge (MA) (1997)
2. Benmokhtar, R., Huet, B.: Neural network combining classifier based on Dempster-Shafer theory for semantic indexing in video content. In: Proceedings of MMM. Volume 4351. (2007) 196–205
3. Lisetti, C., Nasoz, F.: Using noninvasive wearable computers to recognize human emotions from physiological signals. EURASIP Journal on ASP **11** (2004) 16721687
4. Villon, O., Lisetti, C.L.: Toward Building Adaptive Users Psycho-Physiological Maps of Emotions using Bio-Sensors. In: Proceedings of KI. (2006)
5. Mase, K.: Recognition of facial expression from optical flow. In: Proceedings of IEICE Transactions. Volume E74. (1991) 3474–3483
6. Essa, I.A., Pentland, A.P.: Coding, Analysis, Interpretation, and Recognition of Facial Expressions. IEEE Transactions PAMI **19**(7) (1997) 757–763

7. Cohen, I., Sebe, N., Garg, A., Lew, S., Huang, T.: Facial expression recognition from video sequences. In: Proceedings of ICME 2002. (2002) 121,124
8. Pantic, M., Rothkrantz, L.: Toward an Affect-Sensitive Multimodal Human-Computer Interaction. In: Proceedings of IEEE. Volume 91. (2003) 1370–1390
9. Gunes, H., Piccardi, M.: Bi-modal emotion recognition from expressive face and body gestures. Journal NCA **30**(4) (2007) 1334–1345
10. Noble, J.: Spoken Emotion Recognition with Support Vector Machines. PhD Thesis (2003)
11. Zeng, Z., Hu, Y., Liu, M., Fu, Y., Huang, T.S.: Training combination strategy of multi-stream fused hidden Markov model for audio-visual affect recognition. In: ACM MM. (2006) 65–68
12. Busso, C., Deng, Z., Yildirim, S., Bulut, M., Lee., C., Kazemzadeh, A., Lee, S., Neumann, U., Narayanan, S.: Analysis of emotion recognition using facial expressions, speech and multimodal information. In: Proceedings of ICMI. (2004) 205–211
13. Audio-Visual Affect Recognition through Multi-Stream Fused HMM for HCI. In: CVPR. Volume 2. (2005)
14. Paleari, M., Huet, B., Duffy, B.: SAMMI, Semantic Affect-enhanced MultiMedia Indexing. In: SAMT. (2007)
15. Paleari, M., Huet, B.: Toward Emotion Indexing of Multimedia Excerpts. In: CBMI. (2008)
16. Martin, O., Kotsia, I., Macq, B., Pitas, I.: The eNTERFACE05 Audio-Visual Emotion Database. In: Proceedings of ICDEW. (2006)
17. Galmar, E., Huet, B.: Analysis of Vector Space Model and Spatiotemporal Segmentation for Video Indexing and Retrieval. In: ACM CIVR. (2007)
18. Benmokhtar, R., Huet, B.: Multi-level Fusion for Semantic Video Content Indexing and Retrieval. In: Proceedings of AMR. (2007)
19. IntelCorporation: Open Source Computer Vision Library: Reference Manual (November 2006) [http://opencvlibrary.sourceforge.net].
20. Vukadinovic, D., Pantic, M.: Fully automatic facial feature point detection using Gabor feature based boosted classifiers. In: Proceedings of IEEE ICSMC. (2005) 16921698
21. Sohail, A.S.M., Bhattacharya, P.: Detection of Facial Feature Points Using Anthropometric Face Model. In: Proceedings of SPIEMP. Volume 31. (2006) 189–200
22. Boersmal, P., Weenink, D.: Praat: doing phonetics by computer (January 2008) [http://www.praat.org/].
23. Benmokhtar, R., Huet, B.: Classifier fusion : combination methods for semantic indexing in video content. In: Proceedings of ICANN. Volume 4132. (2006) 65–74
24. Denoeux, T.: An evidence-theoretic neural network classifer. In: Proceedings of IEEE SMC. Volume 31. (1995) 712–717
25. Cohen, I., Garg, A., Huang, T.S.: Emotion recognition from facial expressions using multilevel HMM. In: NIPS. (2000)
26. Benmokhtar, R., Huet, B.: Low-level feature fusion models for soccer scene classification. In: 2008 IEEE ICME. (Jun 2008)