Towards Multimodal Emotion Recognition: A New Approach

Marco Paleari
TEleRobotics and Applications
Italian Institute of Technology
Genoa, Italy
marco.paleari @ iit.it

Benoit Huet
Multimedia Department
EURECOM
Sophia Antipolis, France
benoit.huet @ eurecom.fr

Ryad Chellali
TEleRobotics and Applications
Italian Institute of Technology
Genoa, Italy
ryad.chellali @ iit.it

ABSTRACT

Multimedia indexing is about developing techniques allowing people to effectively find media. Content–based methods become necessary when dealing with large databases as people cannot possibly annotate all available content. Emotions are intrinsic in human beings and are known to be very important for natural interactions, decision making, memory, and many other cognitive functions. Current technologies allows exploring the emotional space by mean of content–based analysis of audio and video, but also thanks to other modalities such as the human physiology.

In this paper, we present the latest development in the emotion recognition part of SAMMI [18] by mean of an extensive study on feature selection and the application of many of the principles we have presented in [17] and [15]. Then, we present the concepts of output thresholding, inverse thresholding, and profiling which we used for improving the results of the recognition. Finally, we present a study on the robustness to rotations and zoom of our feature point tracking system.

Our experiments on the six prototypical emotions by Ekman and Friesen presented in the eNTERFACE'05 database result in as much as 75% correct recognition rate.

Categories and Subject Descriptors

H.1.2 [**User/Machine Systems**]: Human Factor—*Human Information Process*; J.m [**Computer Applications**]: Miscellaneous

General Terms

Human Factors, Algorithms, Experimentation

Keywords

Emotion recognition; facial expressions; vocal expressions; prosody; affective computing; human–centered computing; indexing and retrieval.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

CIVR '10, July 5-7, Xi'an China

Copyright ©2010 ACM 978-1-4503-0117-6/10/07 ...$10.00.

1. INTRODUCTION

Multimedia information indexing and retrieval research is about developing algorithms, interfaces, and tools allowing people to search and find content in all possible forms. Although research in this field have achieved some major steps forward in enabling computers to search texts in fast and accurate ways, difficulties still exists when dealing with different media such as images, audio, or videos.

Current commercial search methods mostly rely on metadata as captions or keywords. On the web this metadata is usually extracted and extrapolated through the text surrounding the media, assuming a direct semantic connection between the two. However, in many cases this information is not sufficient, complete, or exact; in some cases this information is not even available.

Content–based methods are designed to search through the semantic information intrinsically carried by the media themselves. One of the main challenges in content-based multimedia retrieval still remains the bridging of the semantic gap referring to the difference of abstraction which subsists between the extracted low level features and the high level features requested by humans' natural queries.

The ability to recognize emotions, is intrinsic in human beings and is known to be very important for natural interactions, decision making, memory, and many other cognitive functions [21]. Therefore, we argue that the emotion expressed in a piece of media, such as movies or songs, could be used for tasks of indexing and retrieval or automatic summarization.

The information about the emotion that better represents a movie could, for example, be used to index that particular movie by genre-like categories (e.g. happiness ↔ comedy or fear ↔ thriller and horror, etc.). In other cases, such as in adventure or musical movies the link among emotions and film genre is less clear. Still in these cases, there may be, links between the evolution of the emotions in films and their classification [22]. In this sense, action movies could for example be characterized by having an ongoing rotation of surprise, fear, and relief and so on and so forth.

Albeit studies from the indexing and retrieval community [5, 11] 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. These works [3, 9, 10, 13, 22] show the interest for such a kind of emotional content-based retrieval systems, but often lack an appropriate evaluation study. Furthermore, we argue that emotions should not represent the only media characterization and that many other tags about the content of the media shall be used together with emotions to have complete systems; existing literature does not approach this topic.

An 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 or gunfights but the very same explosions will be also present in a documentary about controlled building demolitions and gunfights may be recorded in a shooting range. Another possibility will be to recognize an action movie only through its emotion evolution but this recognition may be very complex. In this case, both these unimodal systems have good chances to fail the task, retrieving non relevant movies. Combining the two systems could facilitate good results with relatively low complexity: videos may be selected which contain explosions and documentaries could be cut off because their general mood and their emotion dynamics are usually very different from the one contained in action movies.

In this work, we present a system for the recognition of emotions which uses audio and video of a subject to classify his emotion. Such a system inside SAMMI interacts with other modules for the extractions of semantic tags facilitating the development of high quality multimedia indexing.

In movies, there are many mean of communicating the emotional states including actors' macro and micro facial expressions, vocal prosody, gestures and postures but also dialogs, music, plot, and even picture colors and setup or camera effects. Of all of these emotional expressions, facial expressions and vocal prosody are, by far, the modalities which are most used, in literature, for the task of emotion recognition.

Typically the state of the art on computer–based automatic facial expressions recognition is based on geometrical facial features such as fiducial face keypoints or shapes of facial components. Pantic [20], who used 25 features as distances and angles from predefined feature points, Sebe [24], who considers 12 facial motion measures, and Essa [7] who uses either a muscle activation model derived from the optical flow of user's video or 2D motion energy maps are typical examples of this principle.

Recently a second stream of the research delineates which is 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 these methods are those of Bartlett [1] and Whitehill [30] who use respectively Gabor wavelets and Haar features, or Fasel [8] who uses the latent semantic statistics of the gray–level intensities.

Literature on audio processing takes, instead, advantage of well known characteristics of the voice such as pitch, energy, harmonicity, speech rate, and mel–frequency cepstral coefficients [4, 14, 17, 31].

Few other modalities, such as physiology, gestures, postures, speech semantics, and others, are thought to carry affective information but are (with the exception of physiology) still only partially exploited and published.

We have overviewed some of the state of the art about recognition of emotions and showed the importance of this particular semantic information in the domain of content–based indexing and retrieval of media. In the following sections we present our approach to emotion recognition together with our methodology of investigation.

2. MULTIMODAL APPROACH

In our approach, emotion recognition is performed by fusing information coming from both the visual and auditory modalities. We are targeting the identification of the six "universal" emotions listed by Ekman and Friesen [6] (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. As several researchers did before us [4, 14, 31], we implicitly make the assumption that these findings are true for emotional prosodic expressions too.

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 will still be able to return an emotional estimation thanks to the other one and 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.

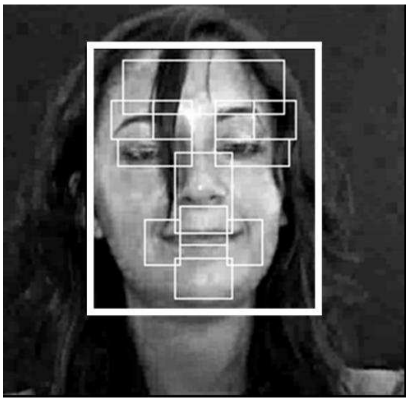


Figure 1: Anthropometric 2D model

Research on multimodal emotion recognition has, nevertheless, a major limitation in the need for high quality multimodal databases. Few databases have been created and fewer are publicly available.

For our experiments the eNTERFACE'05 database [12] (see figure 1) has been selected. This database is composed of over 1300 emotionally tagged videos portraying non-native English speaker displaying a single emotion while verbalizing a semantically relevant English sentence. The 6 universal emotions from Ekman and Friesen [6] are portrayed, namely anger, disgust, fear, happiness, sadness, and surprise. Videos have a duration ranging from 1.2 to 6.7 seconds (2.8 ± 0.8 sec).

This database is publicly available on the Internet but carries few drawbacks mainly due to the low quality of the video compression and actor performances. Furthermore, we could not find other works on this same database, making it impossible to compare the results. Please refer to [17] for an extensive analysis of the database qualities and drawbacks.

In the following sections the techniques for the extractions of emotional features are presented together with an extensive study that was conducted to evaluate the performances of different features and feature vectors. Finally, in section 3, we will present the current implementation of our system for the extraction of emotional estimates.

2.1 Facial Expression Recognition

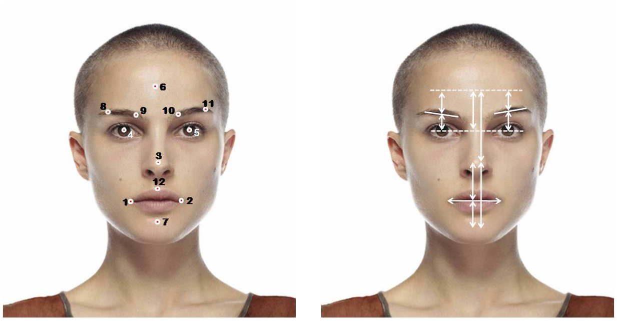
We have developed a system performing real time, user independent, emotional facial expression recognition from still pictures and video sequences [15, 17, 18]. In order to satisfy the computational time constraints required for real-time operation, we employ Tomasi Lucas-Kanade's algorithm [27] to track characteristic face points as opposed to more complex active appearance models [4, 28].

The video is analyzed using the Viola-Jones face detector [29] until a face is found. Multiple consecutive frames are employed to boost the result of the face detection from around 70% to practically 100%. We employ three detectors for the face, the eyes, and the mouth in order to estimate the face orientation. Once the face position and its orientation have been estimated, we proceed to normalize the size of the face and to apply a simple two dimensional anthropometric model of the human face (see figure 1) similarly to what it was done by Sohail and Bhattacharya in [26].

2.1.1 Coordinates Feature Set

Thanks to this model, we can define 12 region of interest as in figure 1 corresponding to the following regions of the face (see also figure 2(a)): right mouth corner, left mouth corner, nose, right eye, left eye, forehead, mouth bottom / chin, external right eyebrow, internal right eyebrow, internal left eyebrow, external left eyebrow, and upper lip / mouth top.

For each one of these 12 regions we search for a cloud of points using the Shi and Tomasi [25] version of the Harris and Stephens algorithm. These points are tracked all along the video using the Tomasi version of the Lucas–Kanade [27] algorithm and for each frame we compute the center of mass of all points belonging to the same region. As a result of this process, 24 features per frame, corresponding to 12 pairs of the feature points (FP) $x(i)$ and $y(i)$ coordinates are created which represent the average movement of the points belonging to the different regions of interest.



(a) Feature Points

(b) Distances

Figure 2: Video Features

2.1.2 Distances Feature Set

This set of 24 coordinate signals represents a first feature set. We have attempted to extract some more meaningful features, from these 24, in a similar way to the one adopted by MPEG-4 Face Definition Parameters (FDPs) and Face Animation Parameters (FAPs) and to the work of Valenti et al. [28]. This process resulted in 14 features $distance(j)$ defined as mouth corner distance, chin distance to mouth, nose distance to mouth, nose distance to chin, left eye to eyebrow distance, right eye to eyebrow distance, left eyebrow alignment, right eyebrow alignment, left eyebrow to forehead distance, right eyebrow to forehead distance, forehead to eye line distance, head x displacement, head y displacement, and normalization factor proportional to head z displacement (see figure 2(b)).

Finally, for each different subject, we divide the quantities found by the relative mean values. In this way, we compress the information which was carried by the 24 signals into 14 relevant movements while also removing the subject's appearance information. This idea has interesting consequences on the domain of biometrics as shown by Paleari et al. [19]. We expect these distances to carry more emotional information than single point coordinates and, therefore, to work better for automatic emotion recognition.

2.2 Prosodic Expression Recognition

Our system for speech emotion recognition, takes deep inspiration from the work of Noble [14]. From the audio signal we extract:

- 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 nine linear predictive coding coefficients (LPC_1 to LPC_9)
- the first ten mel–frequency cepstral coefficients ($MFCC_1$ to $MFCC_{10}$)

These 26 features are collected with the use of PRAAT[1] [2] and downsampled to 25 samples per second to help synchronization with video features. The processing time of the audio analysis is compatible with real-time constrains.

Speech rate and pausing structure are two other features which are thought to carry emotional information but they are related to long term analysis of the speech (several words seconds) and are therefore not compatible with real–time constraints. These two features are, therefore, often not used for emotional speech recognition purposes [4, 14].

2.3 Feature Comparison

We have conducted an extensive study to compare the quality of the features we extract from our database. To these 64 features (24 visual distances, 14 visual moments, and 26 audio) we added some sets of features based on concatenation and grouping.
These sets have been defined as follows:

- sets of variables from coordinates: mouth region, eyes region, nose, and nose and forehead;
- sets of variables from distances: mouth region, eyes region, and head displacements;
- sets of audio variables: pitch and energy, formants, LPC coefficients, MFCC coefficients;

[1] PRAAT is a C++ toolkit written by P. Boersma and D. Weenink to record, process, and save audio signals and parameters. See [2]

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 features individually. 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 needs further development or finer precision.

As a result of these operations 75 sets of one or more features are created. It is expected that affective information is transferred via the dynamics of the facial and prosodic expressions [23]. In order to incorporate dynamics to our framework, we have taken overlapped sliding windows $w(f)$ of the signals changing the size of the window from 1 to 50 frames with a step of one frame; longer time windows carry more information about the dynamics of the signal, shorter better represent the current state of the expression.

In addition to that, we analyzed the system dynamic properties as the signal first and second derivatives: Δ and $\Delta\Delta$.

Some of the statistical characteristic of the signal inside a time windows were investigated too: these are the signal mean values $mean(w(f))$ and standard deviation values $stdev(w(f))$.

We have tested this methodology on the eNTERFACE'05 database under the user–independent condition, i.e. test subjects were never fed to the system during training.

All tests were carried using two-layer feed-forward neural networks with a variable number of neurons for the input layer, 20 neurons as hidden layer, and 6 neurons (one per emotion) for the output layer.

For each possible combination of $feature\_set$, $mode$ and $window\_size$, we have trained a minimum of 3 different Neural Networks (NN) and averaged the different scores. This was done in order to reduce the intrinsic "randomness" effect of NN training.

This results in more than 300,000 different neural networks; i.e. 75 features (and feature sets) by 50 $window_size$s ([1-50]) by 3 $feature\_sets$ (t, Δ, or $\Delta\Delta$) by 3 $modes$ (raw, $mean$, or $stdev$) multiplied by a minimum of 3 different trainings for each setting and 3 different train and test sets. We evaluate the results comparing the CR^+ correct recognition rate of the positive samples. If we summarize the results (please refer to [16] for an extensive review of these results) we observe that depending on the particular emotion and feature different modes should be employed. Generally, we noticed that longer time windows provide slightly better results and that increasing the number of emotionally relevant features does not seem to always improve the result. With the current settings coordinate features work in average better than distances and audio.

More in details the analysis of the results of this study allowed us to conclude that with our dataset and extracted features:

- ***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 $energy$ 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.

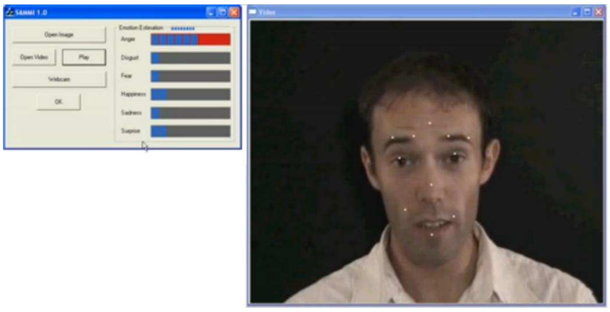


Figure 3: Emotion Recognition System Interface

- ***fear*** is mainly recognized using video features; only $pitch$ 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^{rd} $formant$, the $harmonicity$, and the 5^{th} $MFCC$.

- ***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^{th}, 6^{th} , and 4^{th} $LPCs$ and the 1^{st} $MFCC$

3. EMOTION RECOGNITION SYSTEM

In the former sections we presented the modality of extraction of audio and video features as well as the results of a comparative analysis of their effectiveness for emotion recognition. In this section we overview one of the many possible uses of these result for a multimodal emotion recognition system we have developed (see figure 3).

To evaluate the system as a whole we define a measure called weighted standard deviation $wstd(CR^+) = \frac{std(CR^+)}{m(CR^+)}$. The $wstd$ will be low if all emotions are recognized with the same likelihood and vice versa if some emotions are much better recognized than others, it will be high.

For this experiment we have trained three different neural networks per emotion using data respectively from the audio, the coordinate, and the distances feature sets. In table 1 we list the features which have been selected for this study.

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.

Then, for each feature we have 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.

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 1: Selected features for the different emotions

According to our study longer windows generally returns better results than shorter ones. We have selected this value as a tradeoff 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, having one single window–length allow us to adopt simpler fusion techniques.

We have employed neural–networks with one hidden layer composed of 50 neurons which have been trained on a training set composed of 40 randomly selected subjects from the eNTERFACE'05 database [12]. The extracted data was fed to the networks for a maximum of 50 times (epochs). The remaining 4 subjects were used for test. We have repeated these operations 3 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. . The motivation for such filtering is detailed in our previous works [17].

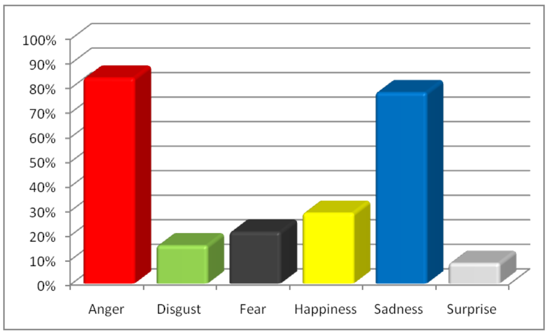


Figure 4: Average CR^+ from the original detectors' outputs

For each emotion we have employed a Bayesian approach to extract a single multimodal emotion estimate per frame o_{emo}. The Bayesian approach has been preferred to other simple decision level fusion approaches and to the NNET approach [15] 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 performs an average recognition rate equal to 45.3%, $wstd(CR^+) = 0.73$ (see figure 4).

The reasons why the *wstd* is so high is because the statistics of the outputs for the six Bayesian emotional detectors are very different. We compute the minimum, maximum, average, and standard deviation values for each one of the detector outputs and proceeded to normalize the outputs to have a minimum estimate equal to 0 and a similar average value.

Performing this operation raise the mean recognition rate to 50.3% while decreasing the $wstd(CR^+)$ to 0.19. In figure 5 we can see the average CR^+ for the six different emotions after this phase of normalization.

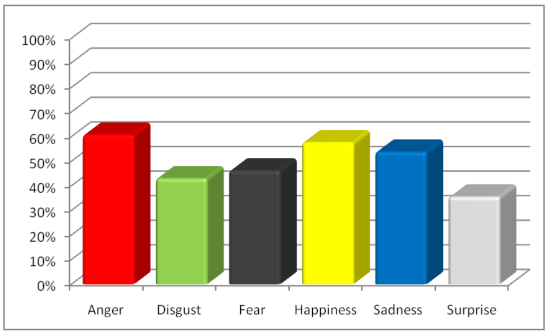


Figure 5: Average CR^+ after outputs normalization

To further boost the results we apply a double thresholding strategy to these results. First, we define a *threshold* below which results are not accepted because they are evaluated as being not reliable enough. By doing this we obviously reduce the number of classified samples therefore reducing the recall of the system. This approach is indeed similar to the one taken in most classification systems in which we generally set a certain recall rate and find the corresponding threshold.

Secondly, we apply a function which we called inverse thresholding. In this case we select more than one estimates for the same audio–video frame in the case in which two (or more) detector outputs are both above a certain $threshold^{-1}$. This operation is somehow similar to using a K–best approach but in this case more estimates are selected only when they are "needed".

Thresholds are 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.

We call the phase of choosing an appropriate couple of thresholds *profiling*. By choosing different profiles the system act differently and we can dynamically adapt the system behavior to the specific need of the system.

We have defined two thresholding profiles returning respectively around 12.5% and 50% of the samples by setting both a lower thresholding and an upper inverse thresholding values.

Almost infinite profiles can be defined which returns about the same number of estimations. Indeed, increasing the threshold or decreasing the inverse threshold have opposite influences on the 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 then no thresholding will be applied.

A second scenario could be the one in which the user need with precise estimations without being too much interested in how often these estimations come. In this case a scenario tagging around 13% of samples will be suitable to increase the recognition while having an average of roughly 3 estimations per second.

It is important to observe that an average of 13% of emotionally tagged samples may turn out not to estimate any samples for few seconds and then to estimate several frames in a row. We argue that lower recall values shall be generally discouraged as they might bring the system to not classify 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.

We applied these thresholding profiles to our outputs. The two newly obtained systems are capable of correctly evaluating respectively 61% and 75% of the recognized samples (see figures 6 and 7). In table 2, we report the specific thresholding settings and the originated results.

As expected, the two systems maintain low weighted standard deviation values while improving the mean recognition rate of the positive samples.

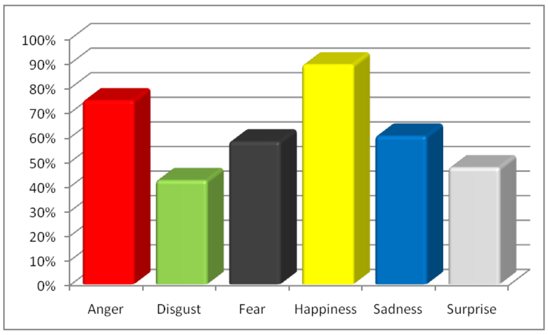


Figure 6: Average CR^+ with the first thresholding profile

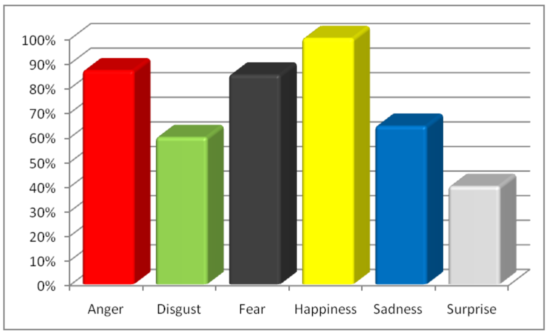


Figure 7: Average CR^+ with the second thresholding profile

In table 3 we report the confusion matrix for the system returning about 3 estimates per second (i.e. about 0.13 of recall).

out / in	ANG	DIS	FEA	HAP	SAD	SUR
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 3: Confusion matrix of the resulting mutimodal system

With the sole exception of surprise which is often confused with anger, fear, and sadness all emotions are recognized in more than 60% of the cases. 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.

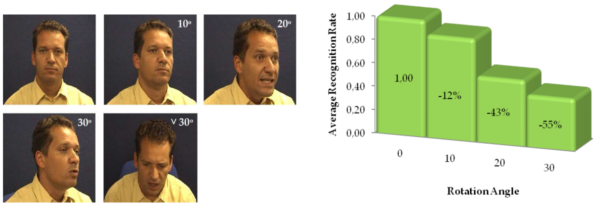
Surprise is the emotion which our system recognizes with most difficulties. This result was to be expected from our previous studies and from the theory. As we pointed out previously, surprise is theoretically hard to distinguish to other emotions as it is most often at least slightly valenced: positive as in sudden happiness or negative as in fear.

4. RELAXING CONSTRAINTS

We have shown how to build up a system performing audio–visual emotion recognition on the eNTERFACE database. One may argue that the constraints of such a database are too strict and therefore that the results cannot be applied to real video setups such as in films and TV series. To overcome this problem we have designed two small experiments:

Viewpoint.

Firstly, the system has been tested on a small database of new videos in which the subject was not frontal to the camera. We setup the angle of the camera by asking three subjects to utter the 30 sentences of the eNTERFACE database while staring frontally to four horizontal positions at 0, 10, 20, and 30^o to the camera. Similarly, a repetition has been tested by moving the camera above the subject to form an angle of about 30^o.



(a) Viewpoints (b) Results

Figure 8: Influence of the viewpoint

The results[2] in figure 8(b) show that the system performances are quite stable to small angles (i.e. $\leq 10^o$ but tend to fall for bigger angles.

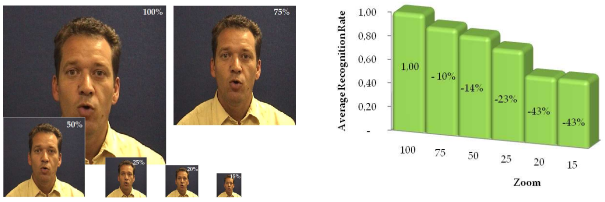
Image Size.

[2] Please note that the results for the vertical or horizontal 30^o angle are the same.

Profile #	Recall	Thresholding Values	Inverse Thresholding Values	$m(CR^{+})$	$wstd(CR^{+})$
1	49.7%	$m(o_{emo}) + 1.2 * std(o_{emo})$	$m(o_{emo}) + 2.0 * std(o_{emo})$	61.1%	0.29
2	12.9%	$m(o_{emo}) + 3.0 * std(o_{emo})$	$m(o_{emo}) + 5.0 * std(o_{emo})$	74.9%	0.29

Table 2: Selected features for the different emotions

Secondly, our system was tested with videos which were reduced in size. For this purpose we reduced the videos of the eNTERFACE database size to 75, 50, 25, 20, and 15% size as in figure 9(a).



(a) Zooms

(b) Results

Figure 9: Influence of the image size

The results in figure 9(b) show that our system performances are quite stable to the influence of size. Below the floor of 25% zoom (i.e. when the face size falls below 90x90 pixels) our face detection module often falls to detect the face and the recognition of users' emotions is practically impossible.

5. CONCLUDING REMARKS

In this paper, we have discussed the topic of multimodal audio–video emotion recognition. Our target is the real–time, user independent, identification of the six "universal" emotions [6] from web–cam quality video and audio of people.

Many different scenarios for human–computer interaction and human-centered computing will profit from an application performing such a task. In particular, we have showed few examples explaining how emotions could be used, together with other content–based information, for tasks such as the indexing, the retrieval, and the summarization of media.

We have briefly overviewed an extensive study on feature selection and feature vector generation for emotion recognition. This study involved the training of more than 300,000 different NN which are compared to evaluate 64 different features and 11 different sets of features in 9 different modalities.

This thorough study lead to two conclusion; individual emotions are generally better recognized by different features and/or modalities (audio or video) and in general, different features need different kind of processing if one wants to effectively extract the emotional information.

Then, our emotion recognition system has been presented and we have discussed the idea of thresholding, inverse thresholding, and profiling. The system we have presented is able to recognize about 75% of the emotions presented by the eNTERFACE'05 database [12] at an average rate of more than 3 estimates per second.

Finally, we have presented two simple studies on relaxation of the constraints regarding the view angle and the size of the face on the reference frame.

Ongoing work consists on testing the presented system on real multimedia excerpts. In this study, about 90 real video sequences from two TV series (namely "How I met your mother" and "The Fringe") were extracted. Albeit the results of this test are not available yet, the preliminary results seem to confirm that our system could work for the detection of emotions in real video sequences.

Future work will focus on the idea, developed in [4], of separating the frames of the video shots into two classes of silence/non silence frames and applying different processing to the two classes.

6. REFERENCES

[1] S. M. Bartlett, G. Littlewort, M. Frank, C. Lainscsek1, 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.

[2] P. Boersma and D. Weenink. Praat: doing phonetics by computer, January 2008. [http://www.praat.org/].

[3] C. H. Chan and G. J. F. Jones. Affect-based indexing and retrieval of films. In *Proc. of ACM MM '05, 13th ACM international conference on Multimedia*, pages 427–430, New York, NY, USA, 2005. ACM.

[4] D. Datcu and L. Rothkrantz. Semantic audio-visual data fusion for automatic emotion recognition. In *Euromedia' 2008*, Porto, 2008.

[5] 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.

[6] P. Ekman and W. V. Friesen. A new pan cultural facial expression of emotion. *Motivation and Emotion*, 10(2):159–168, 1986.

[7] I. Essa and A. Pentland. Coding, Analysis, Interpretation, and Recognition of Facial Expressions. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 19(7):757–763, 1997.

[8] 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.

[9] 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.

[10] 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.

[11] 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 Appllications*, 2(1):1–19, February 2006.

[12] O. Martin, I. Kotsia, B. Macq, and I. Pitas. The eNTERFACEŠ05 Audio-Visual Emotion Database. In *Proc. of the 22nd International Conference on Data Engineering Workshops (ICDEW'06).* IEEE, 2006.

[13] 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 *Proc. of ACM MM'05 ACM International Conference on Multimedia*, pages 853–861, Singapore, 2005.

[14] J. Noble. Spoken emotion recognition with support vector machines. *PhD Thesis*, 2003.

[15] M. Paleari, R. Benmokhtar, and B. Huet. Evidence theory based multimodal emotion recognition. In *MMM '09* 15th *Intl Conference on MultiMedia Modeling*, Sophia Antipolis, France, January 2009.

[16] M. Paleari, R. Chellali, and B. Huet. Features for multimodal emotion recognition: An extensive study. In *Proc. of IEEE CIS'10 Intl. Conf. on Cybernetics and Intelligence Systems*, Singapore, June 2010.

[17] M. Paleari and B. Huet. Toward Emotion Indexing of Multimedia Excerpts. In *CBMI '08 Sixth International Workshop on Content-Based Multimedia Indexing*, London, June 2008. IEEE.

[18] 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*, Dec 2007.

[19] 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 2009.

[20] M. Pantic and L. J. M. Rothkrantz. Expert system for automatic analysis of facial expressions. *Image Vision Comput.*, 18(11):881–905, 2000.

[21] R. Picard. *Affective Computing*. MIT Press, Cambridge (MA), 1997.

[22] A. Salway and M. Graham. Extracting information about emotions in films. In *Proc. of ACM MM '03, 11th ACM international conference on Multimedia*, pages 299–302, New York, NY, USA, 2003. ACM.

[23] K. Scherer. *Appraisal processes in emotion: Theory, methods, research*, chapter Appraisal Considered as a Process of Multilevel Sequential Checking, pages 92–120. New York: Oxford University Press, 2001.

[24] 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.

[25] 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.

[26] A. Sohail and P. Bhattacharya. *Signal Processing for Image Enhancement and Multimedia Processing*, volume 31, chapter Detection of Facial Feature Points Using Anthropometric Face Model, pages 189–200. Springer US, 2007.

[27] C. Tomasi and T. Kanade. Detection and tracking of point features, April 1991. CMU-CS-91-132.

[28] 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.

[29] P. Viola and M. Jones. Robust real-time object detection. *Intl Journal of Computer Vision*, 2001.

[30] J. Whitehill and C. W. Omlin. Haar features for facs au 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.

[31] Z.Zeng, M. Pantic, G. 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, January 2009.