In *Proceedings of the 6th International Workshop on Content Based Multimedia Indexing*, CBMI08, Jun, 18-20, 2008, London, UK.

TOWARD EMOTION INDEXING OF MULTIMEDIA EXCERPTS

Marco Paleari, Benoit Huet

Eurecom Institute
Multimedia Department
2229, route des Crêtes, Sophia Antipolis, France

ABSTRACT

Multimedia indexing is about developing techniques allowing people to effectively find media. Content-based methods become necessary when dealing with large databases. Current technology allows exploring the emotional space which is known to carry very interesting semantic information. In this paper we state the need for an integrated method which extracts reliable affective information and attaches this semantic information to the medium itself. We describe SAMMI [1], a framework explicitly designed to fulfill this need and we present a list of possible applications pointing out the advantages that the emotional information can bring about. Finally, different scenarios are considered for the recognition of the emotions which involve different modalities, feature sets, fusion algorithms, and result optimization methods such as temporal averaging or thresholding.

1. INTRODUCTION

The increasing number of media publicly available on the web encourage the search for effective indexing and retrieval systems.

Current technologies use metadata which are extracted from the text surrounding the media itself supposing a tight link between these two elements. Unfortunately, this is not always the case and the text surrounding a piece of media is often other than its description, furthermore, it is rarely accurate or as complete as we would like.

In recent years, academia has been developing automatic content based methods to extract information about media excerpts. Audio and video are being analyzed to extract both low level features, such as tempo, texture, or color, and abstracted attributes, e.g. person (in an image), genre, and others.

It is well known that in most medias, in most human communications forms, and notably in art expressions, emotions represent a non-negligible source of information.

Even though studies from this community [2] 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 [3, 4, 5, 6, 7] have been done to link emotions to content-based indexing and retrieval of multimedia.

Salway and Miyamory [3, 4] analyze the text associated to a film searching for occurrences of emotionally meaningful terms; Chan et al. [5] analyze pitch and energy of the speech signal of a film; Kuo [6] canalizes features such as tempo, melody, mode, and rhythm to classify music and [7] uses information about textures and colors to extrapolate the emotional meaning of an image.

The evaluation of these systems lack of completeness but when the algorithms are evaluated they allow to positively index as much as 85% of media showing the feasibility of this kind of approach.

Few more researches have been studying algorithms for emotion recognition from humans. Pantic and Rothkrantz [8] provides a thorough state of the art in this field of study. The main techniques involve facial expressions, vocal prosody or physiological signals such as heart rate or skin conductivity and attain as much as 90% recognition rate through unimodal classification algoritms.

Generally, the described algorithms work only under a number of lab condition and major constraints. Few system have analyzed the possibility of using bimodality to improve reliability or to reduce the number of constraints. Busso et al. [9] for example use bimodal audio and visual algorithms to improve the recognition rate reaching as much as 92% on 4 emotions (anger, happiness, sadness, and neutral).

In this paper, we present a general architecture which extracts affective information and attaches this semantic information to the medium itself. Different modalities, feature sets, and fusion schemes are compared in a set of studies.

This paper is organized as follows. Section 2 discusses the audio-video database we use for this study. Section 3 introduces some possible scenarios in which emotions can actively be used to improve media searches. Section 4 discusses the architecture that we designed to extrapolate emotions and link them to the medium itself. Section 5 describes the various experiments we have conducted and their results. Finally Section 6 presents our conclusions and cues for future work.

2. THE ENTERFACE DATABASE

The eNTERFACE database [10] is a publicly available audio-visual emotion database. The base contains videos of 44 subjects coming from 14 different nationalities. 81% of the subjects were men, 31% wore glasses and 17% had a beard, finally one of the subjects was bald (2%).

Subjects were told to listen to six different short stories, each of them eliciting a particular emotion (anger, disgust, fear, happiness, sadness, and surprise), and to react to each of the situation uttering 5 different predefined sentences. Subjects were not given further constraints or guidelines regarding how to express the emotion and head movements were allowed.

The base finally contains 44 (subjects) by 6 (emotions) by 5 (sentences) shots. The average video length is about 3 seconds summing up to 1320 shots and more than one hour of videos. Videos are recorded in a lab environment: subjects are recorded frontal view with studio lightening condition and gray uniform background. Audio is recorded with an high quality microphone placed at around 30 cm from the subject mouth.

The eNTERFACE audio-visual database is a good emotion database and is the only publicly available that we have found for bimodal audio and video; it does, nevertheless present some limitations:

1. The quality of the encoding is mediocre: the 720x576 pixels videos are interlaced and encoded using DivX 5.05 codec at around 230 Kbps using 24 bits color information resulting, sometimes, in some blocking effect.

2. Subjects are not trained actors possibly resulting in a mediocre emotional expression quality.

3. 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.

4. 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.

5. The reference paper acknowledge some videos (around 7.5%) does not represent in a satisfactory way the desired emotional expressions. These videos were, in theory, rejected but this is apparently not the case in the actual database.

This kind of drawbacks introduce some difficulties but it allows us to develop algorithms which should be robust in realistic scenarios. A user study will, nevertheless, be conducted in the future to evaluate the human ability to recognize the emotions presented in the database and generally to evaluate the database quality.

3. SCENARIOS

In many cases, it is very interesting to use emotions for indexing and retrieval tasks. For example one could argue it is simpler to define music as "romantic" or "melancholic" than to define its genre, tempo or melody. Similarly, film and book genres are strongly linked to emotions as can clearly be seen in the case of comedies or horrors. For these reasons we argue that content based semantic tags need to be coupled to emotions to build complete and flexible systems.

One example showing the importance of a multidisciplinary approach, is automatic movie summarization. Indeed when constructing a summary, both specific objects/events (such as explosions, kissing, and others depending on the movie genre), and scenes regarding specific emotions (both elicited in the public or shown by the character) should be selected. For example, we could say that the summary of an action movie should elicit "suspense" or that an horror movie should show fearful people. Fusion between the two modalities (events and emotions) will enable selecting scenes according to both their content and affective meaning (e.g. thrilling gunfights, romantic speech, etc.)

The same principles can be applied to an indexing scenario: an action movie could be, for example, characterized by the fact of having an ongoing rotation of relevant emotions and for having explosion or shooting scenes. A documentary about demolitions through controlled explosions, though, will contain the very same explosions as an horror movie will contain relevant emotions. The presence of emotions will discriminate between the documentary and the action movie and the presence of explosion will discriminate between the action and the horror genres. Using a combined approach should, therefore, correctly index the videos.

In other domains the input about user emotions could help improving the quality of the feedback to the user himself; this is the case of gaming, telemedicine, e-learning, communications, and all human-computer interactions when the affective state plays an important role in the interaction.

We have seen, so far, how emotions can join other media content descriptors in order to improve upon the performance of content-based retrieval and semantic indexing systems. We have briefly listed some other scenarios in which emotions can play an important role. In the next section we describe "Semantic Affect-enhanced MultiMedia Indexing" (SAMMI), a framework we are developing which allows creating such a kind of systems.

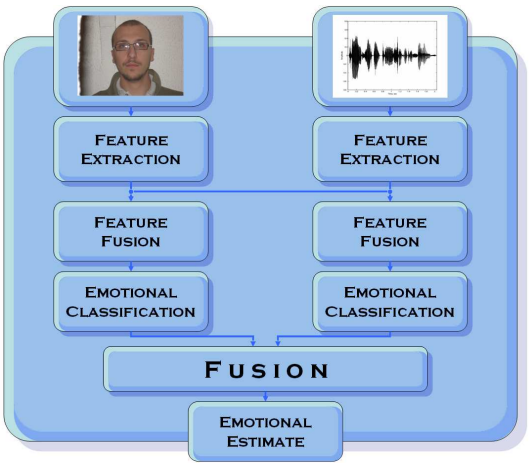

Fig. 1. Bimodal emotion recognition

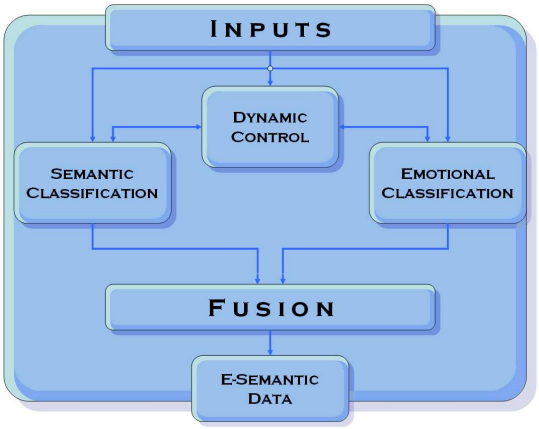

Fig. 2. SAMMI's architecture

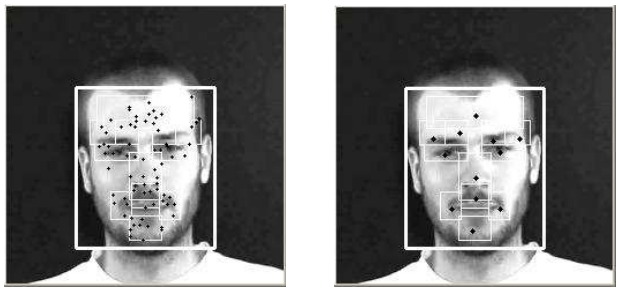

Fig. 3. Followed feature points (FP)

4. THE GENERAL FRAMEWORK

This section describes "Semantic Affect-enhanced MultiMedia Indexing" (SAMMI), a framework explicitly designed for extracting reliable real-time emotional information through multimodal fusion of affective cues and to use it for emotion-enhanced indexing and retrieval of videos.

Two main limitations of existing work on emotion-based indexing and retrieval have been shown: 1) emotion estimation algorithms are very simple and not very reliable and 2) emotions are generally used without being coupled with any other content information.

In SAMMI, emotions are estimated exploiting the intrinsic multimodality of the affective phenomena. Indeed, emotions have a visual component (facial expression, gestures, etc.), an auditory component (vocal prosody, words uttered, etc.), and generally modulates both volitional and un-volitional behaviors (autonomous nervous system, action choices, etc.).

SAMMI (see Fig. 1) takes in account two modalities: the visual and the auditory ones. In particular it analyzes facial expressions and vocal prosody.

Video analysis. The Intel OpenCV library [11] is used to analyze the visual part of the signal. At first, the video is analyzed and a face is searched and eventually found using the Adaboost technique. When the face is found and its position on a video frame is defined 12 regions are defined which corresponds to emotionally meaningful regions on the face (see Fig. 3 (forehead, 2 regions on the left & right brows, eyes, nose, upper & lower central mouth, and 2 mouth corners).

For each region a certain number of points is found which will be easy to follow with the Lukas Kanade algorithm [12]. The position of the points inside one region is averaged to increase the stability of the algorithms in case one point is lost or it is moved outside its true position because of video imperfections (in Fig. 3 the small dots on the left represent the followed points while the bigger dots on the right represent the center of mass of the points belonging to one region). These points are followed along the video and the coordinates of the 12 centers of mass are saved.

This process generates 12[1] couples (x and y components) of signals which are windowed and analyzed to extract meaningful feature vectors. Different window size have been preliminarily analyzed and a length of 25 frames (1 sec) has been selected as optimal. Two possible feature vectors, a statistical and a polynomial representations, will be presented and compared in the next section. The next step consists in training a classifier with the computed feature vectors. Two classifiers, i.e. the Neural Networks (NN) and the Support Vector Machine, will be compared in the next section.

Audio Analysis. The audio is analyzed offline with the help of PRAAT [13], a powerful open source toolkit for audio, and particularly speech, analysis. Thanks to this software pitch, formants, linear predictive coefficients (LPC), mel-frequency cepstral coefficients (MFCC), harmonicity and intensity of the speech are computed. Again these signals are windowed (1 sec windows), 2 different feature vectors are computed (statistical and polynomial), and two different classifiers are trained (NN and SVM). This approach is substantially equivalent to the one described in [14].

[1]Only 11 of the 12 feature points are actually followed because the upper central mouth point was judged not stable enough

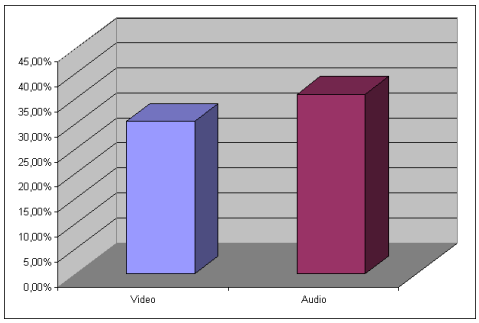

Fig. 4. Video vs. Audio average emotion estimation

As described before, SAMMI (see Fig. 1) takes advantage of multimodality to increase estimation reliability. Emotional information are, thus, fused together and two different levels of fusions are available. These will be described in the next section: 1) the feature level fusion (i.e. the feature vectors are fused and classified together) and 2) the decision level fusion (i.e. the emotional estimates are "averaged" together).

The emotional estimate is part of the general framework described in Fig. 2. Indeed, to develop the applications listed in Section 3, it is necessary to couple the emotional information with other content based semantic information.

Other techniques need to be developed to extract other content-based semantic information and to fuse together these two kinds of information [15, 16].

Dynamic control (Fig. 2) is 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.

We have overviewed SAMMI, a framework for semantic tagging of medias enhanced by affective information. In the next section we presents and compare different the techniques for emotion recognition that we have tested.

5. RESULTS

This section relates the results of several studies we have conducted which give an idea about how different modalities, features sets, fusion techniques, and or other algorithms can influence the results of the emotion estimation.

5.1. Modalities: Video vs Audio

SAMMI can analyze 2 different modalities: 1) video and 2) audio. Figure 4 shows how the two modalities perform on average. It is possible to notice that on average our processing of the audio signal is 17% more reliable than the processing of the video signal.

Observing Figure 5[2] we can observe the behavior of the

[2]Please notice that the perfect emotion detector will be represented by an hexagon having for vertexes the 100% probability; the random generator will be represented by an hexagon crossing at around 17%.

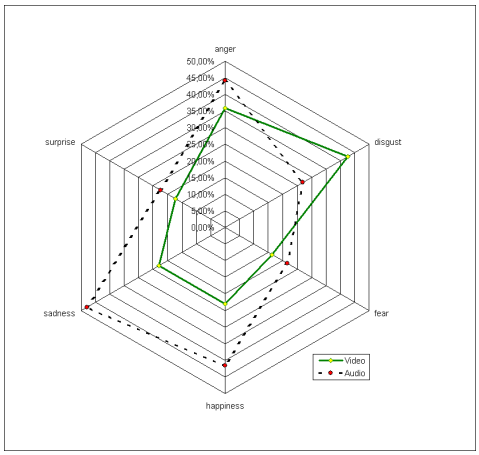

Fig. 5. Video vs. Audio emotion estimation

algorithms for the 6 analyzed emotions. In general, we estimate that, for equivalent average score, a figure resembling more to an hexagon centered in the origin is the optimal (which is also equivalent to say that the best figure is the one maximizing the minimum recognition score or that we want to maximize the inscribed area).

In this case the audio processing figure is the best but we can observe an interesting behavior: both audio and video work better on some particular emotions but these emotions are not the same for the two modalities. In particular, even if the video processing seems less reliable than the audio one, the emotion "disgust" is better recognized from the facial expression than from the vocal prosody. Combining the information coming from the two modalities should therefore improve the overall results.

5.2. Feature Sets

For both modalities we have been testing two feature sets as well as a third resulting from the concatenation of the previous feature vectors. The signals resulting from processing both video and audio were described with 1) a statistical model (based on mean, variance, standard deviation, 5 quantiles, max (and its position), and min (and its position)) and 2) a fifth order polynomial approximation.

In Figures 6 and 7 we can see the effect of the different feature sets on the results for the audio modality. One can observe the polynomial analysis works the worse but still carries some information which can improve the results of the statistical analysis. This is probably due to the choice of using a fifth order polynomial regression which is probably too sensitive to small changes in the original data.

We can see that the particular data we had trained make the system perform very badly on the surprise emotion. A

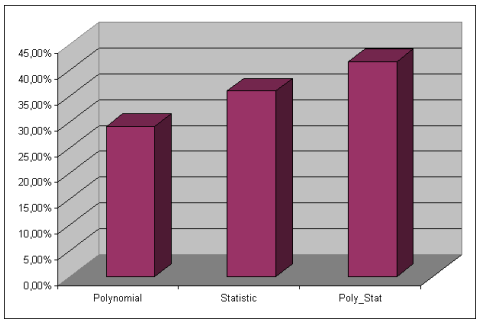

Fig. 6. Average recognition score: different feature sets

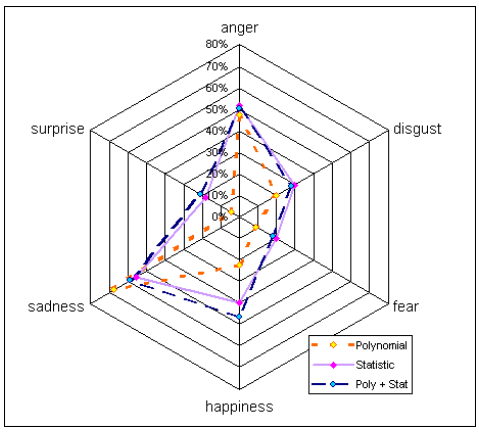

Fig. 7. Emotion recognition using different feature sets

statistical representation of the signals, other than improving the average score, also improves the distribution of the results decreasing the score of the two preferred emotions (sadness and anger) and improving every other score (Fig. 7).

5.2.1. Low Level Features sets: 11 vs 64 points

For the video analysis we have also tested a system other than the one based on feature points. The process is basically the same: the face is found and some regions are defined for which the movement is estimated; this time the regions are result of a regular grid of 8 by 8 cells (See Figures 8 & 9).

This approach is equivalent to consider a dense motion flow instead of a feature point (FP) tracking. We have trained the database with these new data and tested it with the three feature sets. The results (Fig. 10, 11) show a similar average recognition rate but a "nicer shape" resulting from the feature point tracking algorithm. This process could, nevertheless, be used to improve results through fusion; in fact the recognition score for half of the emotions is better recognized from the dense flow algorithms than from the FP tracking.

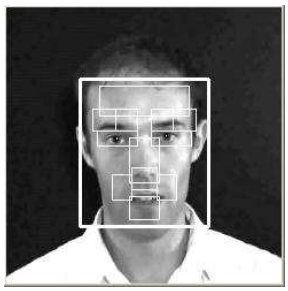

Fig. 8. 12 regions

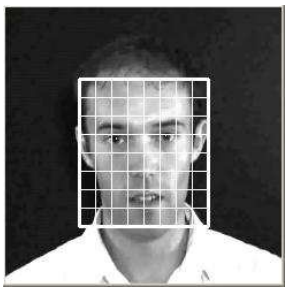

Fig. 9. 64 regions

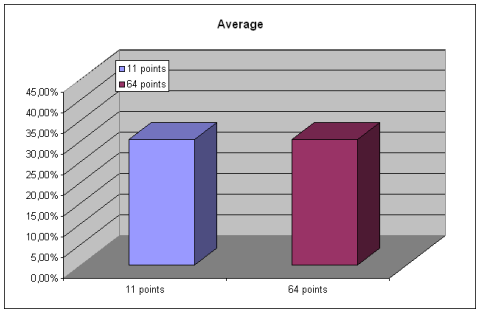

Fig. 10. Average recognition rate: 11 vs. 64 points

5.3. Classifier: SVM vs NN

Another factor which influences the results is the choice of the classifiers. We have been testing neural networks (NN) and support vector machines (SVM). We have employed Matlab to create feed-forward backpropagation neural networks. We did vary the number of neurons between 15 and 100.

We have adopted the libSVM [17] for training and testing SVM. A radial basis function (RBF) has been used as kernel as suggested in [17]. All other parameters have been left to default values.

We can observe in Figure 12 that the average result is quite similar. Nevertheless, we observe, once again, that the distribution of the results (Fig. 13) is quite different suggesting the possibility to exploit fusion of the two results to improve the final recognition score.

5.4. Detectors and Classifiers

Until now the performance of the system have been tested by training one single classifier for the six emotions. However every single emotion has a different temporal behavior. It is therefore possible that the classifier designed for one emotion would work worse on the others and viceversa. This conclusion leads to one solution: substitute 6 detectors to the single classifier. Every detector will be trained to react to one single emotion maximizing the recognition rate.

The results (shown in Figures 14 & 15) show how the adoption of 6 detectors in place of one classifiers for emotion

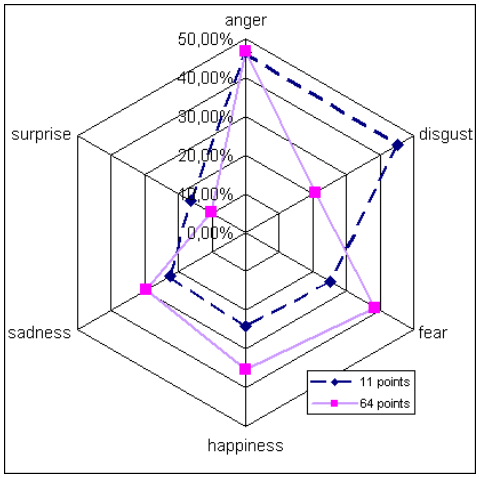

Fig. 11. Emotion estimation: 11 vs. 64 points

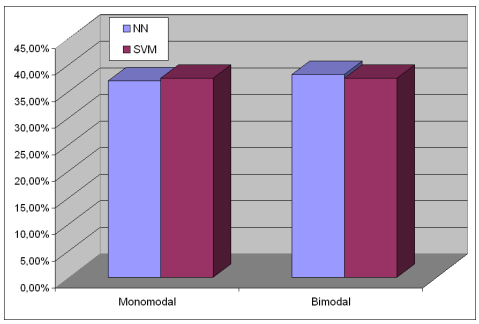

Fig. 12. Average recognition rate: SVM vs. NN

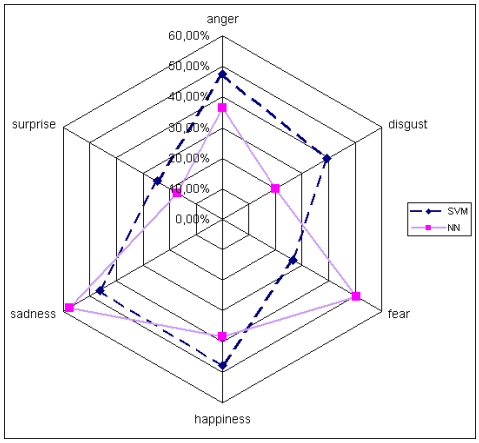

Fig. 13. Emotion estimation: SVM vs. NN

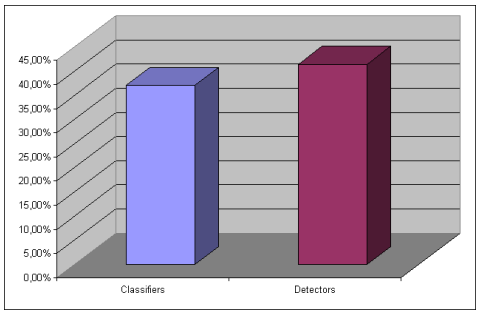

Fig. 14. Average recognition rate: classifier vs. detectors

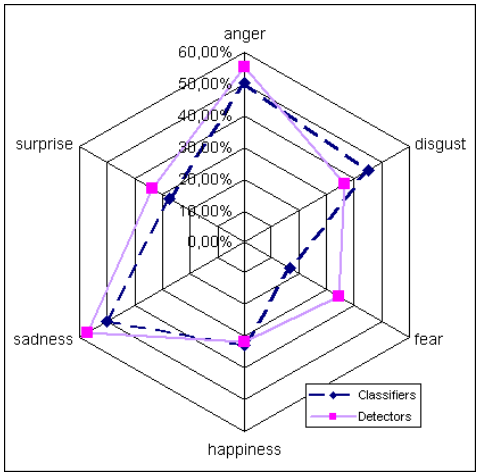

Fig. 15. Emotion estimation: classifier vs. detectors

recognition (from the facial expression) not only improves the average score but also improves the distribution of the results.

5.5. Multimodal Fusion: Features vs Decisions

In this section we want to show how fusion could improve results. In particular we are comparing monomodal systems (Audio and Video) with three different kind of fusion.

The first, namely feature fusion, consist in training the classifier (a NN in Figure 16) with feature vectors resulting from the concatenation of the monomodal video and audio feature vectors. With decision fusion we mean a simple averaging of the outputs of the two monomodal classifiers. With optimized decision fusion we finally mean the weighted average of the two output. The recognition scores for the different emotion of the two monomodal system are used as weights.

Different fusion paradigms leads to different results. We were expecting feature fusion to preserve more information and therefore to work better. Unexpectedly decision fusion (and optimized decision fusion), while employing very simple algorithms works 5% (15%) better than feature fusion. This may suggest that the original data are noisy with respect to emotions and that the emotion estimation is finally noisy

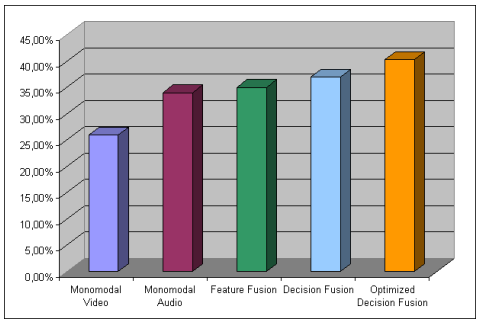

Fig. 16. Average Recognition rate: multimodal fusion

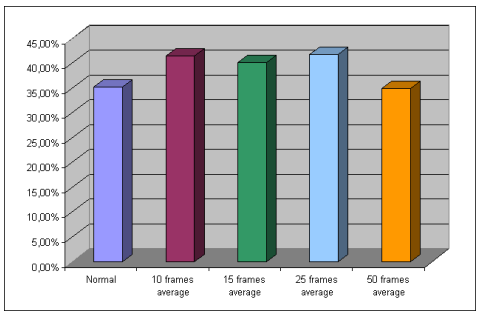

Fig. 17. Average recognition rate: temporal average

too. An average between two modalities does therefore increase the recognition score by reducing the noise (which is statistically independent on the two modalities).

5.6. Temporal Averaging

Starting from the previous conclusions we have tried to perform a temporal averaging of the classification output with the results shown in Figures 17 & 18. The result of such an operation shows a relative improvement of about 18% on the average recognition rate. Furthermore this simple operation does, once again, improve the distribution of the scores improving the recognition rate of the emotions which were less recognized leaving the more recognized emotions untouched. We can notice that averaging windows with lenght between 10 and 25 frames result in similar average score. Bigger averaging windows decreases the score particularly reducing the recognition score of the two emotions fear and happiness.

5.7. Thresholding

Another technique to improve the results is to apply a threshold to the output data. In some scenarios one does not need a frame-by-frame emotion estimation but only to detect when a strong emotion is expressed. In this cases applying a threshold may result in a good technique to improve results in the detriment of the number of emotional estimation. We have tried two different paradigm for thresholding.

Fig. 18. Emotion estimation: temporal average

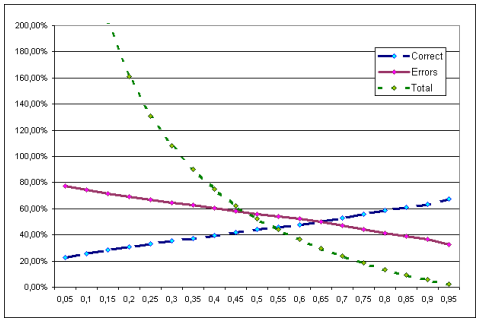

Fig. 19. Effect of thresholding on average recognition

In Figure 19 we show the result of a thresholding of the data for which every detection probability superior to the threshold (x axis) is transformed to 1. By doing this we actually increase the number of detected emotion and we remove the constraint of having one emotion at time, but we detect also the emotions which are not the most likely but are, nevertheless, more probable than the threshold. For low threshold values this technique increases the number of errors but it also improves the score linearly with the threshold up to 67% (from the original 34%). Please note that for low theshold the number of detection is higher than 1 per sample (6 per sample for $th = 0$, 3.6 per sample for $th = 0.05$, etc.).

In Figure 20 we show the result for a thresholding of the data which change to 0 every value smaller than the threshold. In this case the process cuts out emotions which are classified with a low detection probability, compulsorily decreasing the number of detections. The score improves with the threshold up to a maximum of 54.67%.

Both techniques can be exploited to increase the percentage of correctly detected emotion while reducing the total number of detections (note that maximum score is obtained when around 5% frames are tagged).

6. CONCLUSIONS

We have introduced SAMMI: a framework for Semantic Affect-enhanced MultiMedia Indexing. We have overviewed its ar-

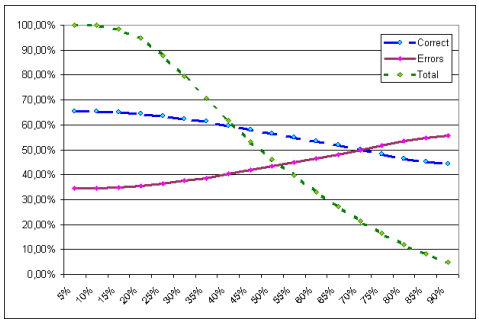

Fig. 20. Effect of thresholding on average recognition (with MAX)

chitecture and presented some scenarios which emphasize the need for a combined affective and semantic tagging.

We have discussed the matter of emotion recognition through speech prosodic features and facial expressions. We have proposed and tested several techniques which can be used to ameliorate the recognition algorithms. Using such techniques we have succeeded to double the average recognition rate. Finally, we have reached with a single technique 67% of average recognition rate. The adoption and the fusion of different techniques can lead to better results.

New feature sets needs to be found to better represents the data. Is our impression that video data should be elaborated more to extract more emotionally meaningful information.

7. REFERENCES

[1] 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.

[2] M.S. 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*, vol. 2, no. 1, pp. 1–19, February 2006.

[3] A. Salway and M. Graham, "Extracting information about emotions in films," in *Proceedings of ACM Multimedia '03*, 2003, pp. 299–302, Berkeley, CA, USA.

[4] 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 *Proceedings of ACM Multimedia '05*, 2005, pp. 853–861, Singapore.

[5] C. Chan and G.J.F. Jones, "Affect-based indexing and retrieval of films," in *Proceedings of ACM Multimedia '05*, 2005, pp. 427–430, Singapore.

[6] F. Kuo, M. Chiang, M. Shan, and S. Lee, "Emotion-based music recommendation by association discovery from film music," in *Proceedings of ACM Multimedia '05*, 2005, pp. 507–510, Singapore.

[7] E.Y. Kim, S. Kim, H. Koo, K. Jeong, and J. Kim, "Emotion-Based Textile Indexing Using Colors and Texture," in *Fuzzy Systems and Knowledge Discovery*, L. Wang and Y. Jin, Eds. 2005, vol. 3613/2005 of *LNCS*, pp. 1077–1080, Springer.

[8] M. Pantic and L.J.M. Rothkrantz, "Toward an Affect-Sensitive Multimodal Human-Computer Interaction," in *Proceedings of IEEE*, 2003, vol. 91, pp. 1370–1390.

[9] 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 *Proceedings of ICMI*, 2004, pp. 205–211, State College, PA, USA.

[10] O. Martin, I. Kotsia, B. Macq, and I. Pitas, "The eNTERFACE05 Audio-Visual Emotion Database," in *Proceedings of the 22nd International Conference on Data Engineering Workshops (ICDEW'06)*. IEEE, 2006.

[11] IntelCorporation, "Open Source Computer Vision Library: Reference Manual," November 2006, [http://opencvlibrary.sourceforge.net].

[12] B.D. Lukas and T. Kanade, "An iterative image registration technique with an application to stereo vision," in *Proceedings of International Joint Conference on Artificial Intelligence (IJCAI)*, 1981, pp. 674–679.

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

[14] J. Noble, "Spoken Emotion Recognition with Support Vector Machines," *PhD Thesis*, 2003.

[15] 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.

[16] R. Benmokhtar and B. Huet, "Multi-level Fusion for Semantic Video Content Indexing and Retrieval," in *5th International Workshop on Adaptive Multimedia Retrieval*, LIP6, Paris, France, 2007.

[17] 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.