In *Proceedings of the 2nd International Conference on Semantic and Digital Media Technologies*, Dec, 5-7, 2007, Genoa, Italy.

SAMMI

Semantic Affect-enhanced MultiMedia Indexing

Marco Paleari[1], Benoit Huet[1] and Brian Duffy[2]

[1] Eurecom Institute, B.P. 194,
F-06904 Sophia Antipolis Cedex, France
paleari@eurecom.fr
[2] The SmartLab, University of East London
London, UK

Abstract. Multimedia indexing is about developing techniques allowing people to effectively find media. Content-based methods become necessary when dealing with big 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 present a list of possible applications and advantages that the emotional information can bring about together with a framework called SAMMI and the preliminary results of this newly initiated research work.

1 Introduction

Emotions have been demonstrated to influence many different human cerebral functions and in particular human memory [1]. In this paper we present few possible scenarios involving content-based indexing, retrieval, and summarization of media and we show how a coupled affect and semantic approach can improve results of such a kind of systems. We then detail an architecture combining emotion recognition through multimodal fusion and automatic semantic labeling/tagging of videos for content-based retrieval and summarization.

Even though studies from the indexing and retrieval community 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 [2–6]. [2, 3] analyze the text associated to a film searching for occurrences of emotionally meaningful terms; [4] analyze pitch and energy of the speech signal of a film; [5] canalize features such as tempo, melody, mode, and rhythm to classify music and [6] 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.

State of the art algorithms for emotion recognition usually use the speech signal and/or the facial expression (see [7] for a thorough overview) approaching

a recognition score of 90%. Some limitations are nevertheless usually applied on the training and testing data which makes the data not realistic. Illumination, audio quality, database size, head movement and position, user (in)dependency, or distractions such as beard or glasses represent usually the main challenges for this kind of systems. Only few works have exploited the intrinsically multimodal nature of emotions by using two or more modalities, usually audio and video, and claiming interesting performances (around 90%).

2 Motivations & Case Studies

It seems, in many cases, very reasonable 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. We argue emotions need to be coupled to other content-based semantic tags to build complete and flexible systems.

One example showing the importance of a multi-disciplinary approach could be where one is trying to summarize one action movie: one may look for scenes regarding gunfights and therefore looking for shootings. Supposing there are, in the film, scenes in a shooting range, we may not want to select them. Looking at the content alone would return these scenes together with the real gunfights while only looking for emotionally relevant scenes instead would result in finding scenes which do not contains shootings at all. The combination of the two, however, will be able to return scenes which are emotionally relevant and do contain shootings and that are, therefore, likely to belong to gunfights. 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 surprise, fear, and relief and for having explosion or shooting scenes.

We have seen, so far, how emotions can join other media content descriptors on order to improve upon the performance of content-based retrieval and semantic indexing systems. In the next section we describe SAMMI, a framework we are developing which allows creating such a kind of systems.

3 Semantic Affect-enhanced MultiMedia Indexing

This section describes 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.

There are three main limitations of existing work on emotion-based indexing and retrieval that have been shown: 1) emotion estimation algorithms are very simple and not very reliable, 2) emotions are generally used without being coupled with any other content information and 3) the evaluation of the experiments is preliminary and quite incomplete;

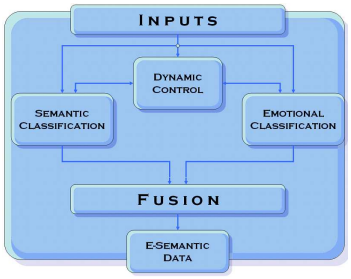

Fig. 1. SAMMI's architecture

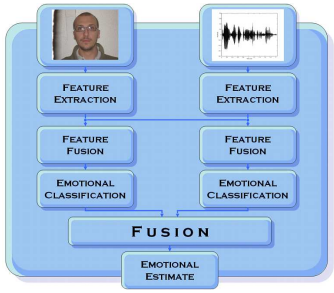

Fig. 2. Bimodal emotion recognition

SAMMI estimates emotions through a multimodal fusion paradigm. Speech is analyzed and different feature sets are extrapolated: pitch, speech formants, energy, MFCC, and Rasta-PLP. Those feature sets are fed to different classification systems (e.g. HMM, GMM, and SVM) to have different emotion estimates to compare. Simultaneously a face is found in the video and the expression is analyzed through motion flow and feature point positions and movements; these features are also fed to different classification systems. Multimodal feature fusion will be experimented, leading to additional emotion estimates.

The different emotion appraisals are fused to extrapolate a single emotion estimate (see Fig. 2). Dynamic control (Fig. 1) 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.

SAMMI couples emotions and other semantic information (Fig. 1). The extraction of different feature sets from the same media, as well as the application of different classification techniques and the use of different modalities are all characteristics which assure good reliability; the use of dynamic control assure stability in presence of noise.

4 Preliminary Results and Concluding Remarks

We have currently developed the automatic and real-time extraction of the feature points from the video. When a face found (Haar classifier) the video is cropped, resized and equalized. Twelve facial zones are considered. For each zone some points are followed along the video (Lukas & Kanade algorithm). The trajectory of the center of mass of the 12 point sets is used as output (Fig. 3).

With the obtained data we trained two classifiers (a NN and a SVM) with different settings and we reached an average 48.4% recognition rate with a strong predominance of the anger and sadness emotions compared to the others (Fig. 4). The analysis of the temporal information reveals that different emotions are recognized according to different temporal patterns. Each point of the graphs in Fig. 4 represents the likelihood to recognize, in a video-shot at a specific time (x axis), the given emotion (column) known the expressed emotion (line).

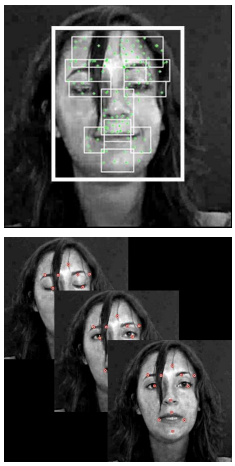

Fig. 3. Video processing

Fig. 4. Confusion Matrix Table

Future work will thus explore possibilities for exploiting this temporal pattern. Additionally, we think some improvements can be reached by using 6 different detectors (one for each emotion) instead of one classifier and by exploiting the multimodality, intrinsic in emotions, and therefore by processing audio.

We believe the examples we have exposed justify the need of such a multidisciplinary approach by making clear its positive impact on tomorrows multimedia indexing and retrieval systems. We argue that this is possible because of the very nature of emotions which facilitates bridging the semantic gap.

References

1. Damasio, A.R.: Descartes' Error: Emotion, Reason, and the Human Brain. Avon books, NY (1994)
2. Salway, A., Graham, M.: Extracting information about emotions in films. In: Proceedings of ACM Multimedia '03. (2003) 299–302 Berkeley, CA, USA.
3. Miyamori, H., Nakamura, S., Tanaka, K.: Generation of views of TV content using TV viewers' perspectives expressed in live chats on the web. In: Proceedings of ACM Multimedia '05. (2005) 853–861 Singapore.
4. Chan, C.H., Jones, G.J.F.: Affect-based indexing and retrieval of films. In: Proceedings of ACM Multimedia '05. (2005) 427–430 Singapore.
5. Kuo, F.F., Chiang, M.F., Shan, M.K., Lee, S.Y.: Emotion-based music recommendation by association discovery from film music. In: Proceedings of ACM Multimedia '05. (2005) 507–510 Singapore.
6. Kim, E.Y., Kim, S.J., Koo, H.J., Jeong, K., Kim, J.I.: Emotion-Based Textile Indexing Using Colors and Texture. In Wang, L., Jin, Y., eds.: Fuzzy Systems and Knowledge Discovery. Volume 3613/2005 of LNCS., Springer (2005) 1077–1080
7. Pantic, M., Rothkrantz, L.: Toward an Affect-Sensitive Multimodal Human-Computer Interaction. In: Proceedings of IEEE. Volume 91. (2003) 1370–1390