In *Jamboree 2007: One Day Workshop By and For KSpace Ph.D. Students*, Berlin, Germany, 2007

USING EMOTIONS TO TAG MEDIA

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

[(1)]Eurecom Institute, Sophia Antipolis, France [(2)]The SmartLab, University of East London, UK

ABSTRACT

Multimedia information indexing and retrieval is about developing techniques which allow people to effectively find the media they are looking for. Content-based methods become necessary when dealing with big databases due to the limitations inherent in metadata-based systems. Current technology allows researchers to explore the emotional space which is known to carry very interesting semantic information, but emotion recognition systems, however, lack sufficient reliability when dealing with real world data. A possible solution to this problem resides in the multimodal fusion paradigm which aims at improving robustness to real world noise. We state the need for an integrated methodology which extracts reliable affective information through a multimodal fusion system and tags this semantic information to the medium itself. A framework, EMMA, currently under development in our laboratory, will be described.

1. INTRODUCTION

It has been demonstrated that events and objects appraised as emotionally relevant are memorized in more permanent ways but also, that the organization of memory in humans is such that similar remembrances (i.e. which elicit similar emotional reactions) are, linked and stored close to each other. This suggests that emotions are an important characteristic of human memory, by helping us to retrieve the memories we are looking for [1,2].

Considerable efforts have been done in the domain of emotion recognition from different media. Indeed, emotions are mainly recognized from three kind of media: audio, images (still images and video), and physiological signals.

Even though studies from the indexing and retrieval community [3] acknowledge that emotions are an important characteristic of media and that they might be used in interesting ways as semantic tags, only few efforts have been done in using emotions in content-based indexing.

In this paper, we present an architecture that includes emotion recognition through multimodal fusion of affective cues and automatic tagging of videos (with audio) for content-based retrieval and summarization.

2. PREVIOUS WORKS

Research in this field basically developed from work in 2003 with Salway and Graham [4] with the extraction of emotional feature from the transcriptions of audio-descriptors of films for visually-impaired people. 679 different words were considered as emotion tokens belonging to one of the 22 different emotions described in the OCC model. Miyamori et al. [5] perform similar algorithms using blogs' texts. Chan and Jones [6] use films audio and, in particular, pitch and energy of the actors' speech signal. Kuo et al. [7] use films music and algorithms exploiting features such as tempo, melody, mode, and rhythm to classify music. Finally Kim et al. [8] use information about texture and colors of an image to extrapolate the emotion elicited by that picture in humans.

These works show the interest of the community in these approaches. Nevertheless the algorithms used are often not very reliable and the system evaluation lacks completeness.

3. EMMA: EMOTION MULTIMEDIA ANNOTATION

One well known technique to increase reliability involves exploiting multimodal information through fusion. Emotions are intrinsically multimodal (i.e. they affect speech, facial expression, physiology and many other modalities). Indeed, our approach to increase emotion estimation reliability passes through a multimodal fusion framework (Figure 2). EMMA extracts emotions from both the speech (auditory) and the facial expression (visual) signals.

Dynamic control (Figure 1) is used to adapt the fusion algorithms to the quality of the different systems. Indeed if lighting is bad using color information should be limited and the emotion estimate should privilege the auditory modality. Furthermore EMMA couples affective and semantic labeling of the same data.

Future plans consist of increasing the number of modalities starting from the inclusion of emotion recognition through skin conductivity and heart rate (physiology). Pitch, pitch contours, formants, speech energy, mel frequency cepstral coefficients (MFCC), and Rasta-PLP coefficients will be computed for the audio. Feature points positions and movements will be considered together with motion flow as video features.

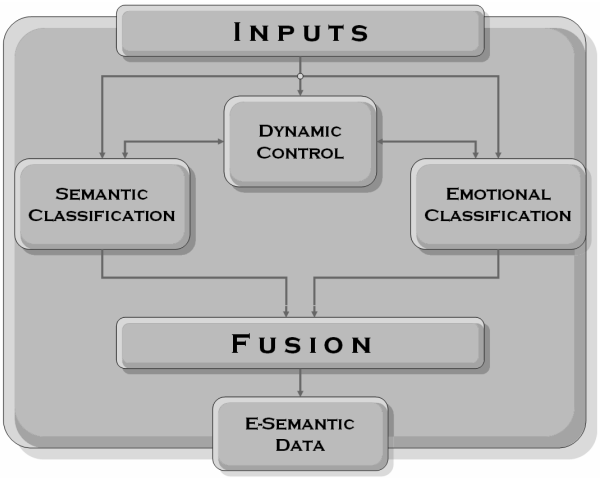

Figure 1: Emotion & Semantic Tagging

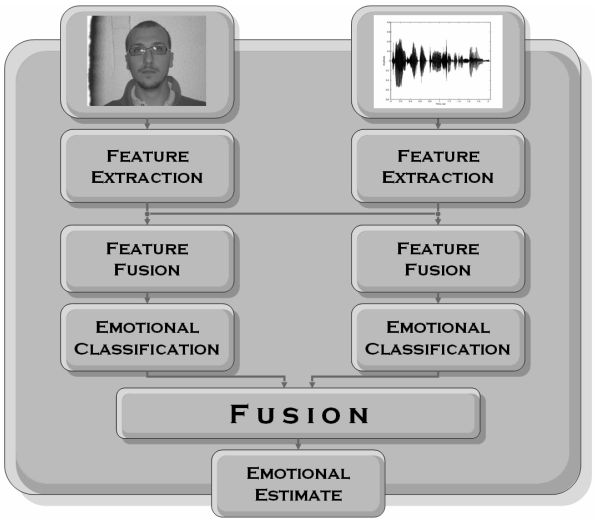

Figure 2: Multimodal Emotional Classification Module

Multimodal feature fusion will be experimented and those different feature sets will be fed into the emotional classification modules. GMM, HMM, ANN, SVM, and other classification techniques will be experimented. Some of them will then be selected and will work in parallel. All the system will then result in an array of emotion appraisals which will become, through fusion, in a single reliable emotion estimate.

Indeed, emotions should not represent the only media characterization. Many other tags about the content of the media may be used together with emotions to have complete systems. In the case of music recommendation it is true that we may want to listen to melancholic or happy music but we may also be interested in a specific music genre, band, or song; in these latter cases emotion-based approaches will not be very useful.

Similarly, while summarizing an action movie, one may look for scenes regarding gunfights and therefore looking for shootings. Supposing there are, in the film, scenes in a shooting range we may not want to select them. Looking at the content alone would return these scenes together with the real gunfights; 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 are, therefore, likely to belong to gunfights.

4. CONCLUSION

One of the main limitations of current content-based indexing and retrieval systems resides in the semantic gap between the high level description of the content searched by humans and the low level descriptors used by computer algorithms to index and then retrieve media.

We have seen, through some examples, that a multidisciplinary affective and content-based approach can help bridge the semantic gap and provide users with interfaces allowing for more natural and effective human-computer interactions.

We presented EMMA, an architecture which allows to index media through their affective and semantic information and therefore to build such a kind of system.

5. REFERENCES

[1] Damasio, A.R.: Descartes' Error: Emotion, Reason, and the Human Brain. *Avon books*, NY (1994)

[2] Lisetti, C.L., Gmytrasiewicz, P.: Can a Rational Agent Afford to be Affectless? A Formal Approach. A*pplied Artificial Intelligence* 16 (Agoust 2002) 577–609(33)

[3] Lew, M.S., Sebe, N., Djeraba, C., Jain, R.: Content-based multimedia information retrieval: State of the art and challenges. *ACM Transaction* 2(1) (February 2006) 1–19

[4] Salway, A., Graham, M.: Extracting information about emotions in films. In *ACM Multimedia 03*, Berkeley, CA, USA.

[5] Miyamori, H., Nakamura, S., Tanaka, K.: Generation of views of TV content using TV viewers' perspectives expressed in live chats on the web. In *ACM Multimedia 05*, Singapore.

[6] Chan, C.H., Jones, G.J.F.: Affect-based indexing and retrieval of films. In *ACM Multimedia 05*, Singapore.

[7] Kuo, F.F., Chiang, M.F., Shan, M.K., Lee, S.Y.: Emotion-based music recommendation by association discovery from film music. In *ACM Multimedia 05*, Singapore.

[8] Kim, E.Y., Kim, S., Koo, H., Jeong, K.: Emotion-Based Textile Indexing Using Colors and Texture. In Wang, L., Jin, Y., eds.: *Fuzzy Systems and Knowledge Discovery*. Vol. 3613/2005 of Lecture Notes in Computer Science., Springer (2005) 1077–1080.