Towards Temporal Saliency Detection: Better Video Understanding for Richer TV Experiences

Joël Dumoulin, Elena Mugellini, Omar Abou Khaled
Department of Information Technologies
HES-SO, Fribourg, Switzerland
joel.dumoulin@hes-so.ch
elena.mugellini@hes-so.ch
omar.aboukhaled@hes-so.ch

Marco Bertini, Alberto Del Bimbo
Media Integration and Communication Center (MICC)
University of Florence, Italy
bertini@dsi.unifi.it
delbimbo@dsi.unifi.it

***Abstract*—More and more popular, Smart TVs and set-top boxes open new ways for richer experiences in our living rooms. But to offer richer and novel functionalities, a better understanding of the multimedia content is crucial. If many works try to automatically annotate videos at object level, or classify them, we think that investigating the emotions through the use of digital analysis and processing techniques will allow great TV experience improvements. With our work, we propose a temporal saliency detection approach capable of defining the most exciting parts of a video that will be of the most interest to the users. To identify the most interesting events, without performing their classification (in order to be independent from the video domain), we compute a time series of arousal (excitement level of the content), based on audio-visual features. Our goal is to merge this preliminary work with user emotions analysis, in order to create a multi-modal system, allowing to bridge the gap between users' needs and multimedia contents.**

Keywords-Digital Analysis and Processing; Temporal Saliency; Affective Content Analysis; Arousal Modelling; Emotions

I. INTRODUCTION

Development of Smart TV devices has been ongoing for many years - first patent in 1994, first real attempt at its production (by Microsoft and Thomson) in 2000 - and now it is becoming a commercial reality. Samsung and LG for instance are pushing their new devices, with advanced features such as voice or gesture based control, and embedded recommender systems. Other actors that are not TV manufacturer also propose to bring the Smart TV experience to living rooms without the need to buy a new TV, with set-top boxes. Despite the limited success of the Google TV (launched in 2010), these devices are becoming more and more popular (e.g., Apple TV).

In order to propose to the user richer experiences with these Smart TVs, it is crucial to better understand not only the user itself, considering for example his interests, but also the content. Bridging the gap between users and multimedia content would allow to propose innovative features, and also improve recommender systems [1]. This represents the starting point of our work: we want to create richer TV user experiences. To this end we need automatic multimedia annotation systems, e.g., to create personalized access to video content, according to user preferences. In particular, we want to explore emotions, because we believe that they play a central role in the user TV experience. Detection of emotions can be done with two points of view: analysis of user emotions and analysis of emotions contained in a video [2][3]. Our goal is to build a multi-modal system, capable of combining the two perspectives, because we think that understanding well what the multimedia content represents in terms of emotions, and what the user is feeling while watching the movie, will allow us to close the distance between user needs and the multimedia content, and to provide new experiences into the living rooms.

In this paper, we propose a temporal saliency approach, capable of detecting the exciting parts of a movie, based on an arousal curve. The rest of the paper is organized as follows. A state of the art for the affective content analysis is presented in Section 2. Our temporal saliency approach is introduced in Section 3. Section 4 details the arousal curve generation. Section 5 presents our preliminary results. Finally, we define our research agenda in Section 6, and conclusions are drawn in Section 7.

II. RELATED WORK

Content-based multimedia information retrieval research provides new methods and techniques to search the ever increasing amount of multimedia content [1], and many of them have potential implications in the Smart TV world.

Video summarization is one key aspect in video management. It not only allows to better understand the dynamic of the movie, but also to improve its browsing and navigation. Many works try to focus on the user, for instance by building user centric models [4] or directly by analyzing the user physiological responses [5]. Another approach is to focus on the multimedia content itself, for instance by detecting audiovisual saliency [6].

Computer systems capable of detecting user emotions would have many interesting applications in human-computer interaction area, but also represent an interesting approach to extract the interesting parts of a video, if we define them as the parts that bring emotion to the user. It is possible to focus on the user [3][7], but also on the multimedia content itself. Detecting user emotions is not only very dependent to the user observed, as emotion is a subjective reaction, but it is also an information not available when the video content is produced. To overcome this problem, Hanjalic et

 ISBN: 978-1-61208-324-7

al. [2][8] proposed to extract and model the affective content of the video, and this approach is called "Affective content analysis". The affective content is defined as the intensity and type of emotion expected to arise in the user while watching an image or a video. Wang et al. [9] proposed a set of affective categories and steps for their classification, in order to improve this affective understanding approach in films. Lu et al. [10] investigated how shape features are related to emotions aroused from images in human beings, by analyzing many characteristics such as roundness, angularity, simplicity, complexity, etc. Zhang et al. [11] propose to take advantage of the affective analysis in order to improve movie browsing, and define rich audio-visual features. Recently, Wang et al. [12] achieved soccer highlight extraction by modelling affection arousal based on both visual and audio arousal related features: sound energy, shot cut density, shot intensity and replay, the highlights being extracted based on the arousal curve crests detected. Benini et al. [13] propose to overcome the subjective sphere of emotions, by shifting the representation towards the connotative properties of movies, in a space inter-subjectively shared among users, allowing to define, relate and compare affective descriptions of films.

If affective content analysis approaches are interesting for highlight extraction, particularly for sport videos, they need to be improved and adapted to films, in order to automatically detect the parts that will be of the most interest to the users.

III. TEMPORAL SALIENCY DETECTION

In our work, we want to define important parts in movies as the parts that are salient regarding the others. This idea is inspired from the work of Itti et al. [14] about saliency-based visual attention determination (bottom-up approach), where the idea is to generate a saliency map for a single image - which topographically codes for local conspicuity over the entire visual scene - in order to define where there is something interesting to look at. Our idea is to adapt this approach into a temporal saliency detection specific to video content. The result is not "where" to look at, but "when" to look at. It is similar to Otsuka et al. [15] proposition of a sports highlights detection function by using audio features, based on "commentator's excited speech" identification. While many works are done around the concept of saliency [16], and also spatiotemporal saliency [17], very few address this idea of temporal saliency in videos.

The approach we chose in order to detect this temporal saliency is based on the emotions that these parts could arouse to the user watching the movie, and particularly the excitement, defined as the arousal. We plan to take also in account the type of the emotion (valence). The schema of our system is illustrated in Fig. 1. It shows the different processing steps that are planned to be implemented, in order to extract the needed elements (visual excitement, audio excitement, etc.) from the video and the audio streams, to depict the excitement (arousal) and the type of the emotion (valence).

Our approach is in some aspect close to highlight extraction techniques. But, while highlight extraction techniques try to summarize the movie, with our temporal saliency approach we try to find the parts of the movie that will be exciting for a particular user profile (from an emotional point of view). In regards to this, our approach is different from highlight extraction. As shown in Fig. 1, we use several features in order

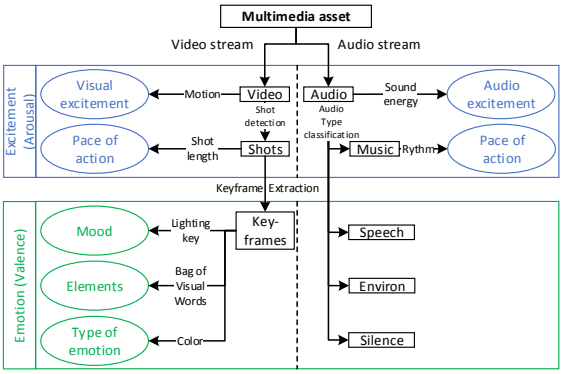

Figure 1. Multimedia asset processing schema. Video and audio streams are processed to extract information about the conveyed excitement and emotion.

to compute the temporal saliency. Some are extracted from the video stream, and the others from the audio stream. It is crucial to use both streams, as they both convey important emotion and excitement information. The processing of those features will allow us to understand the two main affective axes of the multimedia content: the arousal and the valence, needed to define the temporally salient parts of the video.

IV. AROUSAL CURVE GENERATION

As a starting point, we chose two features: shot intensity and sound energy. The shot intensity gives the pace of action, while the sound energy gives the audio excitement. These two features allow us to compute an arousal curve. Finally, we apply a crest generation algorithm on it to detect the most exciting parts of the video.

A. Shot intensity

We detect only hard cuts. As we want to define parts of the movie where there is a high shot intensity, corresponding to high arousal parts, detecting hard cuts is sufficient. As gradual transitions are more used to conclude a scene, it does not means a lot for high shot intensity detection. We chose the Edge Change Ratio (ECR) technique, as it is a good compromise between accuracy and implementation complexity [18]. Our implementation is based on the ECR algorithm proposed by Lienhart [19], but with a small tweak. Usually, a motion compensation step is done in order to compensate small linear movements between two consecutive frames. Instead, we use this idea of calculating the transformation needed to go from one frame to its following, but we calculate it only when the system triggers a shot cut, as a change quantification. Based on a threshold, this change amount value allows us to remove false positives: if the change is small, it is certainly a false positive, but if this value is relatively big, it is certainly a true positive. As we estimate the transformation matrix only when we detect shots and not for every consecutive frames (plus we do not need to apply the transformation), this approach is more efficient, and gives similar results to the standard algorithm. Finally, the shot cut density is computed as proposed by Wang et al. [12]:

$$c(k) = e^{(1-n(k))/r} \quad (1)$$

where *n(k)* is the number of frames including the *k-th* video frame and *r* is a constant determining the distribution of the values.

 ISBN: 978-1-61208-324-7

B. Sound energy

The sound conveys a big part of the emotion of a movie. While several features could be used (music rhythm, etc.), we use the sound energy as our sound feature. We computed the sound energy as described by Wang et al. [12] as:

$$e(k) = \sum_{i=1}^{N} x(i)^2 \quad (2)$$

where *e(k)* is the sound energy at *k-th* frame, *x(i)* is *i-th* sample point value in audio frames, *N* is the number of audio frames.

C. Crests generation

We followed approach proposed by Wang et al. [12] for the crests generation. First, the raw features are normalized and smoothened with a Kaiser window [20]. Then, we used a linear weighted summarization to merge our two features together. After the fusion, the resulting signal is again smoothened and normalized. Finally, the crests are detected with a sliding window based algorithm, and filtered regarding its fluctuant amplitude. We obtained the best results with values of 10 samples for the window size, 3 samples for the window step, and 0.1 for the crest intensity threshold used to filter the insignificant crests. We need to define with future experimentations if these values are depending on the movie type.

V. Preliminary Results

We planned to test our system on the Emotional Movie Database (EMDB) [21], but, unfortunately, it does not come with sound, and we need it in order to compute the sound energy. Instead, we use the Schafer et al. dataset [22], that is composed of a total of 70 film excerpts, representing 7 *a priori* emotional categories: anger, sadness, fear, disgust, amusement, tenderness, and neutral state. It allows us to test our system on different movie styles and categories (black and white, color, horror, comedy, etc.).

Figs. 2, 3, and 4 show the result we obtained by processing one of the movie excerpts. The scene is part of the amusement category. It lasts 2'09", showing "Jacquouille" and Godfroid destroying the postman's car, in the famous French movie "The Visitors". The x-axis corresponds to the frame number. Fig. 2 represents the raw features for the sound and the video streams. The result of the preprocessing step (normalization and smoothing with a Kaiser windows) is shown in Fig. 3. Finally, Fig. 4 is the resulting arousal curve, after the fusion of the two pre-processed features, again normalized and smoothened. The vertical dashed lines are the detected crests, and the two interesting events of the sequence are manually annotated in green.

During this scene, there are two particular events where there is an arousal peak: 1) when the car breaks and stop just in front of "Jacquouille", 2) when "Jacquouille" and Godfroid destroy the postman's car. In these two particular events, there is an increase of the sound energy and the shot cut density, and this is clearly noticeable on the plots. The result is that the arousal curve and the crest detection clearly correspond to these two particular events, and this is a really promising result. We still need to improve our system in order to automatically define these parts around the crests (start and end times).

In order to ease results analysis, we have built a web application, allowing us to navigate through the movie and

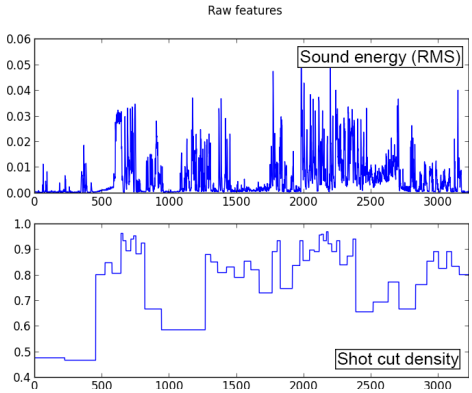

Figure 2. Raw features - Sound energy plot is above (root mean square of the audio stream's sound energy computation), shot cut density is below.

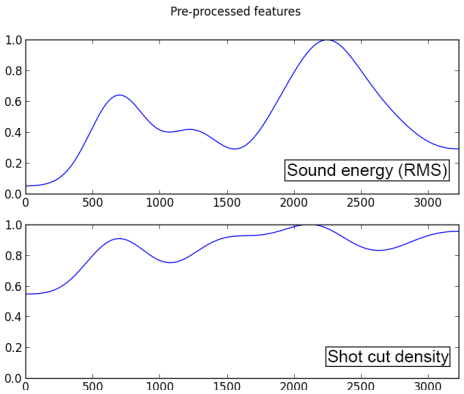

Figure 3. Pre-processed features - Features are normalized and smoothened (with a Kaiser window). Sound energy plot is above, shot cut density is below.

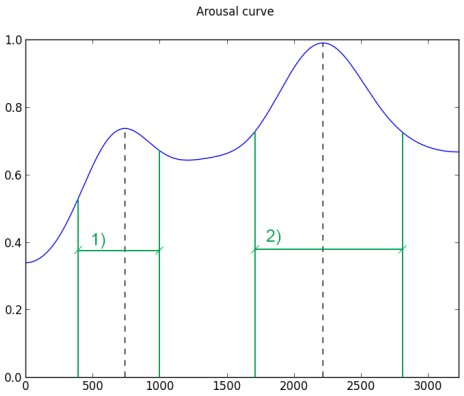

Figure 4. Arousal curve - Arousal curve, resulting from the fusion of the pre-processed features, again normalized and smoothened.

the arousal curve in a convenient way. We can easily go to a particular point of the arousal curve and it automatically seek the movie accordingly. This is helping us to control if the results are coherent with the movie, until we have a ground truth, since it is not available for the dataset we used.

VI. Research Agenda

There are several steps planned next. We first need to validate our system. This step is a challenging one, because, as far as we know, there are currently no datasets containing

 ISBN: 978-1-61208-324-7

the meta data we need to constitute a usable ground truth. Indeed, datasets like EMDB [21] and Schaefer et al. [22] only provide emotional values (arousal, valence) for the whole video sequence in order to classify it in one emotional category. We plan to use MAHNOB-HCI [23], that is a multimodal database recorded in response to affectively stimulating excerpts from movies, over 27 participants. Similarly to the Schafer et al. dataset [22] that we used to conduct our first tests, the movies excerpts cover six main emotional categories: disgust, amusement, joy, fear, sadness, neutral. If different recorded modalities are provided (e.g., facial expressions, audio signals, eye gaze data), we would like to use as a starting point the physiological responses (e.g., electrocardiography, respiration amplitude, skin temperature) in order to generate an arousal curve, that will constitute our ground truth. It will allow us to validate our system, and compare it with other approaches.

Then, in order to improve our system, we plan to add other features. We think that motion analysis will be a good feature to start with, as it depicts well the dynamic of a movie. Once the arousal part validated, we could address the valence part, by analyzing the key-frames allowing to extract the mood, as one of the key aspect that describes the type of the emotion conveyed. Another interesting task would be to have different granularity levels for the emotionally interesting parts of the video. This idea would be to define which parts are interesting at the movie level, or at the chapter level, or at the scene level.

Adapting the curve generation according to the movie type would be an interesting task. This would allow our system to perform well with different movie types and to recognize subcategories of movies (e.g., subtypes of Westerns: Classical Westerns, Spaghetti Westerns, etc.), by using the generated emotional curves as input features for the classification.

VII. Conclusion

We presented our current work, focused on a temporal saliency detection approach in order to define exciting parts of a movie. We think that addressing this task with an emotional point of view can better meet the user needs than with a traditional highlight extraction approach. We presented our method for the arousal curve generation. We also presented a tweak of the Edge Change Ratio (ECR) technique for shot boundary detection. Our goal is to merge our multimedia centric approach, with a user centric approach, based on user emotions detection. This would constitute a multi-modal system, capable of bridging the gap between the users and the multimedia content they access through Smart TVs. Finally, this system will allow us to realize our ideas of new features for Smart TVs, e.g., enhanced recommendation systems matching user mood, and user attention based advanced features like indication to the user that a particularly interesting part of the movie will be missed because he is doing something else.

References

[1] M. S. Lew, N. Sebe, C. Djereba, and R. Jain, "Content-Based Multimedia Information Retrieval : State of the Art and Challenges," ACM Transactions on Multimedia Computing, Communications, and Applications (TOMCCAP), vol. 2, no. 1, 2006, pp. 1–19.

[2] A. Hanjalic, "Extracting moods from pictures and sounds: Towards truly personalized TV," Signal Processing Magazine, IEEE, no. March 2006, 2006, pp. 90–100.

[3] M. Soleymani, G. Chanel, J. J. Kierkels, and T. Pun, "Affective ranking of movie scenes using physiological signals and content analysis," Proceedings of the 2Nd ACM Workshop on Multimedia Semantics, 2008, pp. 32–39.

[4] Y.-f. Ma, X.-s. Hua, L. Lu, and H.-j. Zhang, "A generic framework of user attention model and its application in video summarization," IEEE Transactions on Multimedia, vol. 7, no. 5, Oct. 2005, pp. 907–919.

[5] A. G. Money and H. Agius, "Analysing user physiological responses for affective video summarisation," Displays, vol. 30, no. 2, Apr. 2009, pp. 59–70.

[6] G. Evangelopoulos, "Movie summarization based on audiovisual saliency detection," in Image Processing, 2008. ICIP 2008. 15th IEEE International Conference on, 2008, pp. 2528–2531.

[7] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang, "A survey of affect recognition methods: audio, visual, and spontaneous expressions." IEEE transactions on pattern analysis and machine intelligence, vol. 31, no. 1, Jan. 2009, pp. 39–58.

[8] A. Hanjalic and L. Xu, "User-oriented affective video content analysis," Content-Based Access of Image and Video Libraries, 2001. (CBAIVL 2001). IEEE Workshop on, 2001, pp. 50–57.

[9] H. L. Wang and L.-f. Cheong, "Affective understanding in film," IEEE Transactions on Circuits and Systems for Video Technology, vol. 16, no. 6, Jun. 2006, pp. 689–704.

[10] X. Lu, P. Suryanarayan, R. A. Jr, J. Li, M. G. Newman, and J. Wang, "On shape and the computability of emotions," in Proceedings of the 20th ACM international conference on Multimedia - MM '12, 2012, pp. 229–238.

[11] S. Zhang, Q. Tian, and Q. Huang, "Utilizing affective analysis for efficient movie browsing," in Image Processing (ICIP), 2009 16th IEEE International Conference on, no. 49, 2009, pp. 1853–1856.

[12] Z. Wang, J. Yu, Y. He, and T. Guan, "Affection arousal based highlight extraction for soccer video," Multimedia Tools and Applications, Jul. 2013, pp. 1–28.

[13] S. Benini, L. Canini, and R. Leonardi, "A Connotative Space for Supporting Movie Affective Recommendation," Multimedia, IEEE Transactions on, vol. 13, no. 6, 2011, pp. 1356–1370.

[14] L. Itti, C. Koch, and E. Niebur, "A Model of Saliency-Based Visual Attention for Rapid Scene Analysis," Pattern Analysis and Machine Intelligence, IEEE Transactions on, vol. 20, no. 11, 1998, pp. 1254–1259.

[15] I. Otsuka and K. Nakane, "A highlight scene detection and video summarization system using audio feature for a personal video recorder," Consumer Electronics, IEEE Transactions on, vol. 51, no. 1, 2005, pp. 112–116.

[16] A. Toet, "Computational versus psychophysical bottom-up image saliency: a comparative evaluation study," IEEE transactions on pattern analysis and machine intelligence, vol. 33, no. 11, Nov. 2011, pp. 2131–46.

[17] C. Guo and L. Zhang, "A novel multiresolution spatiotemporal saliency detection model and its applications in image and video compression." Image Processing, IEEE Transactions on, vol. 19, no. 1, Jan. 2010, pp. 185–98.

[18] A. Smeaton, P. Over, and A. Doherty, "Video shot boundary detection: Seven years of TRECVid activity," Computer Vision and Image Understanding, vol. 114, no. 4, 2010, pp. 411–418.

[19] R. Lienhart, "Reliable Transition Detection in Videos : A Survey and Practitioner s Guide," International Journal of Image and Graphics, vol. 01, no. 03, 2001, pp. 469–486.

[20] F. Harris, "On the use of windows for harmonic analysis with the discrete fourier transform," Proceedings of the IEEE, vol. 66, no. 1, Jan 1978, pp. 51–83.

[21] S. Carvalho, J. Leite, S. Galdo-Álvarez, and O. F. Gonçalves, "The Emotional Movie Database (EMDB): a self-report and psychophysiological study." Applied psychophysiology and biofeedback, vol. 37, no. 4, Dec. 2012, pp. 279–94.

[22] A. Schaefer, F. Nils, X. Sanchez, and P. Philippot, "Assessing the effectiveness of a large database of emotion-eliciting films: A new tool for emotion researchers," Cognition & Emotion, vol. 24, no. 7, Nov. 2010, pp. 1153–1172.

[23] M. Soleymani, J. Lichtenauer, T. Pun, and M. Pantic, "A multimodal database for affect recognition and implicit tagging," Affective Computing, IEEE Transactions on, vol. 3, no. 1, 2012, pp. 42–55.

 ISBN: 978-1-61208-324-7