An Audio-Visual Saliency Model for Movie Summarization

Konstantinos Rapantzikos, Georgios Evangelopoulos, Petros Maragos, Yannis Avrithis
School of E.C.E., National Technical University of Athens, Athens 15773, Greece
rap@image.ntua.gr, {gevag ,maragos}@cs.ntua.gr, iavr@image.ntua.gr

***Abstract*— A saliency-based method for generating video summaries is presented, which exploits coupled audiovisual information from both media streams. Efficient and advanced speech and image processing algorithms to detect key frames that are acoustically and visually salient are used. Promising results are shown from experiments on a movie database.**

Keywords—saliency; saliency curves; attention modeling; event detection; key-frame selection; video summarization; audiovisual

Topic area—Multimedia:methods and systems (indexing and search of multimedia)

I. INTRODUCTION

The growing availability of video content creates a strong requirement for efficient tools to manipulate multimedia data. Considerable progress has been made in multimodal analysis for accessing and analyzing video content with automatic summarization being one of the main targets of recent research. Summaries are important, since they provide the user with a short version of the video that ideally contains all important information for understanding the content. Hence, the user may quickly evaluate the video as interesting or not. Generally speaking, there are two types of video abstraction: video summarization and video skimming. Video summarization refers to a collection of key-frames extracted from the sequence, while video skimming represents the sequence in the form of a short clip.

Numerous research efforts have been undertaken for automatically generating video summaries. Earlier works were mainly based on processing only the visual input. Zhuang *et al.* extracted salient frames based on color clustering and global motion [4], while Ju *et al.* used gesture analysis in addition to the latter low-level features [5]. Furthermore Avrithis *et al.* represent the video content by a high-dimensional feature curve and detect key-frames as the ones that correspond to the curvature points [6]. Another group of methods uses frame clustering to select representative frames [7][8]. Features extracted from each frame of the sequence form a feature vector and are used in a clustering scheme. Frames closer to the centroids are then selected as key-frames.

Defining what is important in a video stream is quite subjective and therefore several methods in the field, including the ones referred before, suffer from the limitation that evaluation of the summary is quite difficult and subjective. Hence, mapping human perception into an automated abstraction process has become quite common. In an attempt to emulate the multimodal nature of human understanding, the Informedia project and its offsprings, combined speech, image, natural language understanding and image processing to automatically index video for intelligent search and retrieval [9][10][11]. This approach generated interesting results. Going one step further towards human perception, Ma *et al.* proposed a method for detecting the salient parts of video that is based on user attention models [2]. Motion, face and camera attention along with audio attention models (audio saliency and speech/music) are the cues used by the authors to capture the salient information of the multimedia input and identify the video segments to form the final summary.

In this paper we propose a saliency-based method that exploits individual audio and video saliency information by fusing them and generating video summaries. The visual saliency model is based on a feature completion scheme implemented in a regularization framework. Intensity, color and motion features compete in order to form the most visually salient regions. Audio saliency relates with the audio stream microstructure, captured by emerging modulations in small scales. In effect, the amplitude, frequency and instantaneous energy of such modulations is used to quantify the importance of audio events. Preliminary results are obtained on arbitrary videos and on a movie database annotated with respect to dialogue events [12].

The paper is organized as follows: Section II presents the two methods for computing audio and visual saliency and the way to fuse them. Section III presents the experimental results, while conclusions are drawn and future work is discussed in section IV.

II. PROPOSED METHOD

A. Audio Saliency Features

Attention in audio signals is focused perceptually in abrupt changes, transitions and abnormalities in the stream of audio events, like speech, music, environmental noises in real life recordings or sound effects in movies. The salient features that attract attention in an audio stream, are the ones detected more clearly. Biologically, one of the segregations performed by the auditory system in complex channels is in terms of temporal

modulations, while psychophysically modulated carriers seem to be more salient perceptually than stationary by human observers [2][13][14].

Motivated by the above we construct a user attention curve based on measures of temporal modulation in multiple frequencies (scales). The existence of multi-scale modulations during speech production, justifies the AM-FM modulation superposition model for speech [16], according to which speech formants can be modeled by a sum of narrowband amplitude and frequency varying, non-stationary sinusoids $s(t)=\sum a_k(t)\cos\phi_k(t)$. The model is applied here for general audio signals. Demodulation of a real-valued, monocomponent AM-FM

$$x(t)=a(t)\cos\left(\int_o^t \omega(\tau)d\tau\right) \quad (1)$$

with time varying amplitude envelope $a(t)$ and instantaneous frequency $\omega(t)$ signals, can be approached using the non-linear Teager-Kaiser differential energy operator (EO) $\Psi[x(t)]=\dot{x}(t)-x(t)\ddot{x}(t)$, where $\dot{x}(t)=dx/dt$. Applied to narrowband AM-FM signal, EO yields with negligible approximation error the instantaneous source energy, i.e. $\Psi[x(t)]\approx a^2(t)\omega^2(t)$, corresponding to the physical energy of the oscillation-producing source energy. An efficient AM-FM demodulation scheme based on is the energy separation algorithm (ESA) [15] separates the instantaneous energy into its amplitude and frequency components

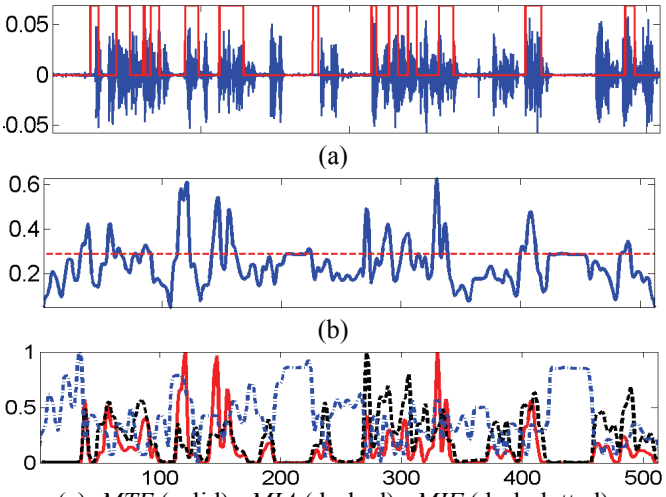
$$\frac{\sqrt{\Psi[\dot{x}(t)]}}{\sqrt{\Psi[x(t)]}}\approx\omega(t), \qquad \frac{\Psi[x(t)]}{\sqrt{\Psi[\dot{x}(t)]}}\approx|a(t)| \quad (2)$$

with a simple, computationally efficient discrete counterpart, of almost instantaneous time resolution. Demodulation through ESA is obtained in the outputs of a set of frequency-tuned, bandpass Gabor filters $h_k(t)=\exp\left(-\sigma_k{}^2t^2\right)\cos(\omega_{kc}t)$ that assume to globally isolate signal modulations [16] in the presence of noise.

By applying the energy operator to the bandpass outputs of a linearly-spaced bank of K filters, a nonlinear energy measurement of dimension K is obtained. For each signal frame m of length N, short-time representations of the dominant modulation component are obtained by tracking, in the multi-dimensional feature space consisting of the filter responses on s, the *maximum average Teager Energy* (MTE)

$$MTE(m)=\max_{1\le k\le K}\frac{1}{N}\sum_{n=1}^{N}\Psi_d(s*h_k) \quad (3)$$

where n is the sample index with (m-1)N+1 ≤ n ≤ mN and h_k the impulse response of the k filter. MTE is considered the dominant signal *modulation energy*, capturing the joint amplitude-frequency information of audio activity [17]. The filter $j(m)=\arg\max\{MTE(m)\}$ is submitted to demodulation via ESA to derive the mean instant amplitude (MIA) and mean instant frequency (*MIF*) features for frame m, leading to a three-dimensional feature vector $A(m)=\{MTE,MIA,MIF\}$ of the mean dominant modulation parameters for each signal frame. An example of the features for an audio stream ("*Jackie Brown*" [12]) can be seen in Fig. 1(c).



(c) *MTE* (solid), *MIA* (dashed), *MIF* (dash dotted)

Fig. 1 (a) audio and saliency indicator, (b) saliency curve with threshold, (c) normalized audio features

B. Visual Saliency

The visual saliency computation module is based on the notion of a centralized saliency map along with an inherent feature competition scheme to provide a computational solution to the problem of Region-Of-Interest (ROI) detection/selection in videos. In this framework, a video sequence is represented as a solid in the three-dimensional Euclidean space, with time being the third dimension. Hence, the equivalent of a spatial saliency map is a spatiotemporal volume where each voxel has a certain value of saliency. This saliency volume is computed with the incorporation of feature competition by defining cliques at the voxel level and use an optimization procedure with constraints coming both from inter- and intra- feature level.

We perform decomposition of the video at a number of different spatiotemporal scales. The final result is a hierarchy of video volumes that represent the input sequence in decreasing spatiotemporal scales. Afterwards, feature volumes for each feature of interest, including intensity, color and 3D orientation (motion) are computed and decomposed into multiple scales. Every volume simultaneously represents the spatial distribution and temporal evolution of the encoded feature. The pyramidal decomposition allows the model to represent smaller and larger "events" in separate subdivisions of the channels.

Feature competition is implemented in the model using an energy-based measure. In a regularization framework the first term of this energy measure may be regarded as the data term (E_D) and the second as the smoothness one (E_S), since it regularizes the current estimate by restricting the class of admissible solutions [18]. The energy involves voxel operations between coarse and finer scales of the volume pyramid, which means that if the center is a voxel at scale $c\in\{2,...,p-d\}$ then the surround is the corresponding voxel at scale $h=c+\delta$ with $\delta\in\{1,2,..d\}$, where d is the desired depth of the center-surround scheme. Hence, if

$F_{0,k}$ corresponds to the original volume of each of the features, with $F = \{ I, RG, BY \}$ and $k \in \{1,...,|F|\}$, each level l of the pyramid is obtained by convolution with an isotropic 3D Gaussian G and dyadic down-sampling:

$$F_{l,k} = (G * F_{l-1,k}) \downarrow 2, \quad l \in \{1,2,...,p\} \tag{4}$$

For each voxel q of a feature volume F the energy is defined as

$$E(F_{c,k}(q)) = \lambda_D \cdot E_D(F_{c,k}(q)) + \lambda_S \cdot E_S(F_{c,k}(q)) \tag{5}$$

where λ_D, λ_S are the importance weighting factors for each of the involved terms. The first term of (2) is defined as

$$E_D(F_{c,k}(q)) = F_{c,k}(q) \cdot \left| F_{c,k}(q) - F_{h,k}(q) \right| \tag{6}$$

and acts as the center-surround operator and the second one as

$$E_S(F_{c,k}(q)) = F_{c,k}(q) \cdot \frac{1}{|N(q)|} \cdot \sum_{\substack{r \in N(q) \\ r \neq q}} \left(F_{c,k}(r) + \tilde{V}_c(r) \right) \tag{7}$$

where $\tilde{V}_c$ is the spatiotemporal orientation conspicuity volume, that may be regarded as an indication of motion activity in the scene.

The motivation behind this feature competition scheme is the experimental evidence of a biological counterpart in the Human Visual System (interaction/competition among the different visual pathways related to motion/depth (M pathway) and gestalt/depth/color (P pathway) respectively) [20]. Shortly, the visual saliency detection module is based on an iterative minimization scheme that acts on 3D local regions and is based on center-surround inhibition regularized by inter- and intra- local feature constraints. The interested user may find a detailed description of the method in [19]. Fig. 1 depicts the computed saliency for three frame of the *"Lord of the rings"* sequence that is included in [12]. Bright values correspond to salient areas (notice Gandalf's moving head and the hobbit's moving arms).

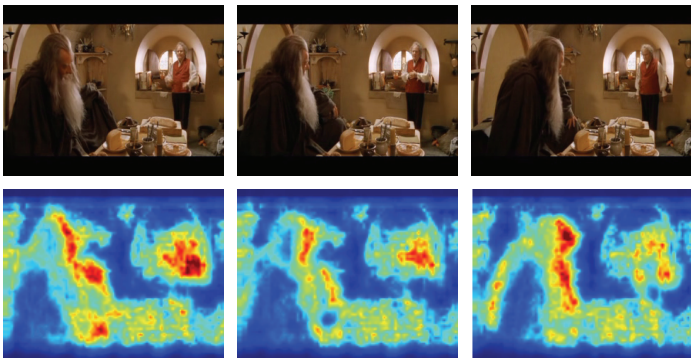


Fig. 2 Original frames and the corresponding saliency maps

C. *Audiovisual Saliency*

Fusing audio and visual information is not a trivial task, since they are computed on modalities of different nature. Nevertheless combining the final output of the two saliency detection modules is straightforward. In this paper we use a simple linear scheme for creating the final audiovisual saliency curve that will provide the key frame index for the summary.

The audio saliency curve is derived by weighted linear fusion of the normalized audio feature vector.

$$S_A = \vec{A} \cdot \vec{w}_A = w_1 \cdot MTE + w_2 \cdot MIA + w_3 \cdot MIF \tag{8}$$

Normalization is done by least squares fit of their individual value ranges to [0, 1] (see Fig.1). In order to provide a global measure of scene change based on saliency we threshold the output of the visual saliency module using a common thresholding technique [21] to discard low saliency areas and compute the average value per frame. Hence, we end up with a 1D vector S_V that describes the change of visual saliency throughout the sequence. The coupled audiovisual curve

$$S_{AV} = w_A \cdot S_A + w_V \cdot S_V \tag{9}$$

serves as an abstract continuous-valued indicator function of salient events, in the audio, the visual or the common audiovisual domain. By detecting simple geometric features of such curves, we can track down important transition and reference points. Such features are the local maxima of the curve (derived by pick-picking), 1D edge transition points (using the zero-crossings of a Derivative-of-Gaussian operator) or regions below certain learned or heuristically defined thresholds. Using the maxima for keyframe selection, and a user-defined skimming index we derive a summarization of the video in terms of its audiovisual saliency.

III. Experimental Results and Discussion

In order to demonstrate the proposed method, we run it both on videos of arbitrary content and on the movie database of A.U.T.H. (Muscle WP5 Movie database v1.1) [12]. This database consists of 42 scenes extracted from 6 movies of different genres. Fig. 3 shows the audio, visual and audiovisual saliency curves for 512 frames of the movie *"Jackie Brown"* included in the previous database. Fig. 4 depicts the same curves with the detected features superimposed, while Fig. 5 shows selected keyframes.

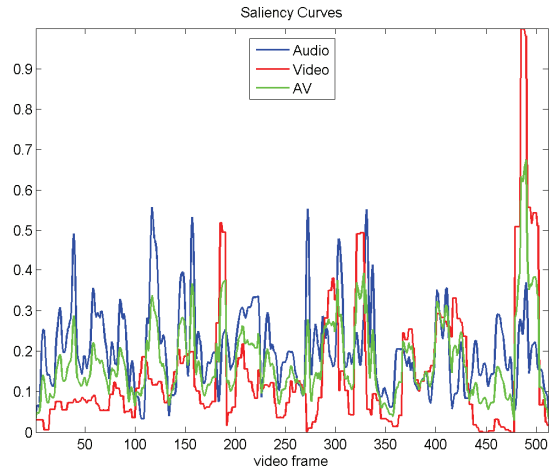


Fig. 3 Superimposed audio, video and audiovisual saliency curves (better viewed in color)

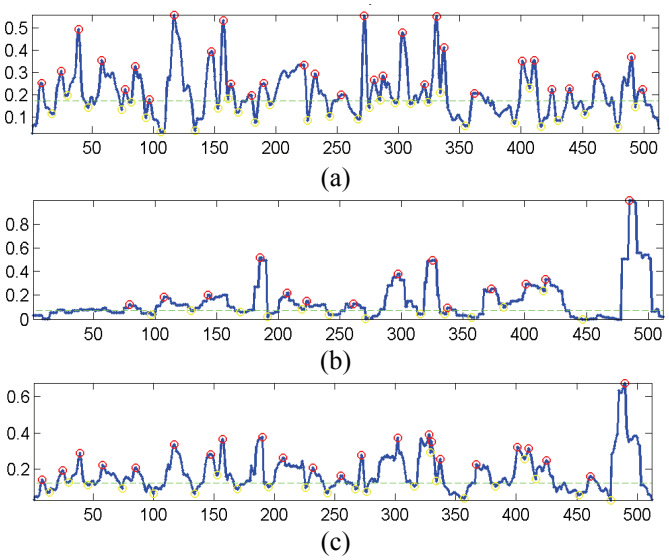


Fig. 4 Curves and detected features for (a) audio saliency, (b) video saliency, (c) audiovisual saliency

Fig. 5 Frames located at the detected points of the audiovisual saliency curve (numbers correspond to frames)

IV. Conclusions and Future Work

In this paper we present two methods for audio and visual saliency computation and explore the potential of their fusion for movie summarization. We believe that movie summarization based on simulated human perception leads to successful video summaries. In the current work we used simple fusion of audiovisual curves to detect key-frames and create the summary. In the future we will examine more fusion methods and extend the technique to create video skims. Video skims are more attractive, since they contain audio and motion information that makes the abstraction more natural and informative.

Acknowledgments

This research work has been supported by the European Network of Excellence MUSCLE. We wish to thank C. Kotropoulos and his group at A.U.T.H. for providing us with the movie database.

References

[1] K. Rapantzikos, Y. Avrithis, "An enhanced spatiotemporal visual attention model for sports video analysis", Proc. CBMI'05, Riga, Latvia, Jun 2005.

[2] Y.-F. Ma, X.-S. Hua, L. Lu, H.-J. Zhang, "A generic framework of user attention model and its application in video summarization", IEEE Trans. on Multimedia, vol. 7, pp. 907-919, Oct 2005.

[3] Y. Li, S.-H. Lee, C.-H. Yeh, C.-C. Jay Kuo, "Techniques for movie content analysis and skimming", IEEE Signal Processing Magazine, pp. 79-89, Mar 2006.

[4] Y. Zhuang, Y. Rui, T.S. Huang, S. Mehrotra, "Adaptive Key Frame Extraction Using Unsupervised Clustering", Proc. ICIP'98, pp. 866-870,Oct 1998.

[5] S.X. Ju, M.J. Black, S. Minneman, D. Kimber, "Summarization of video-taped presentations: Automatic analysis of motion and gestures", IEEE Trans. Circuits Syst. Video Technology, vol. 8, pp. 686-696, Sep 1998.

[6] Y. Avrithis, A. Doulamis, N. Doulamis and S. Kollias, "A Stochastic Framework for Optimal Key Frame Extraction from MPEG Video Databases", Computer Vision and Image Understanding, vol. 75 (1/2), pp. 3-24, Jul 1999.

[7] S. Uchihashi, J. Foote, A. Girgensohn, J. Boreczky, "Video manga: Generating semantically meaningful video summaries", in Proc. ACM Multimedia'99, pp. 383-392, Oct 1999.

[8] K. Ratakonda, M.L. Sezan, R. Crinon, "Hierarchical video summarization", Proc. SPIE, vol. 3653, pp. 1531-1541, Dec 2000.

[9] M.A. Smith, T. Kanade, "Video skimming and characterization through the combination of image and language understanding techniques", Proc. CVPR'97, 1997.

[10] A.G. Hauptmann, "Lessons for the Future from a Decade of Informedia Video Analysis Research", *Lecture Notes in Computer Science*, Volume 3568, pp. 1-10, August 2005.

[11] A.G. Hauptmann, R. Yan, T.D. Ng, W. Lin, R. Jin, D. M., Christel, M. Chen, R. Baron, "Video Classification and Retrieval with the Informedia Digital Video Library System", Proc. TREC'02, Gaithersburg, MD, USA, November 2002.

[12] MUSCLE WP5 Movie Dialogue DataBase v1.1, Aristotle University of Thessaloniki, AIILab, 2007.

[13] C. Kayser, C. I. Petkov, M. Lippert and N. K. Logothetis, **"**Mechanisms for allocating auditory attention: an auditory saliency map", Current Biology, vol. 15, no. 21, pp. 1943-1947, 2005.

[14] N. Tsingos, E. Gallo and G. Drettakis, "Perceptual audio rendering of complex virtual environments"**,** SIGGRAPH 2004.

[15] P. Maragos and J.F. Kaiser and T.F. Quatieri, "Energy Separation in Signal Modulations with Application to Speech Analysis", IEEE Trans. Signal Proc., vol. 41, no. 10, pp. 3024-3051, 1993.

[16] A.C. Bovik and P. Maragos and T.F. Quatieri, "AM-FM Energy Detection and Separation in Noise Using Multiband Energy Operators", IEEE Trans. Signal Proc., vol. 41, no. 12, pp. 3245-3265, 1993

[17] G. Evangelopoulos and P. Maragos, "Multiband Modulation Energy Tracking for Noisy Speech Detection", IEEE Trans. Audio, Speech and Language Proc., vol.14, no.6, pp. 2024-2038, 2006.

[18] K. Rapantzikos. M. Zervakis "Robust optical flow estimation in MPEG sequences", IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Mar 2005

[19] K. Rapantzikos, N. Tsapatsoulis, Y. Avrithis, S. Kollias, "Spatiotemporal saliency for video classification", IEEE Transactions on Multimedia, submitted.

[20] E.R. Kandel, J.H. Schwartz, T.M. Jessell, "Essentials of Neural Science and Behavior", Appleton & Lange, Stamford, Connecticut, 1995

[21] N. Otsu, "A threshold selection method from gray level histograms", IEEE Trans. Systems, Man and Cybernetics, vol. 9, pp. 62-66, 1979