MOVIE SUMMARIZATION BASED ON AUDIOVISUAL SALIENCY DETECTION

G. Evangelopoulos[1], *K. Rapantzikos*[1], *A. Potamianos*[2], *P. Maragos*[1], *A. Zlatintsi*[1], *Y. Avrithis*[1]

[1] School of ECE, National Technical University of Athens, Zografou, 15773 Athens, Greece
[2] Dept. of ECE, Technical University of Crete, Chania 73100, Greece

[gevag,maragos,nzlat]@cs.ntua.gr [rap,iavr]@image.ntua.gr potam@telecom.tuc.gr

ABSTRACT

Based on perceptual and computational attention modeling studies, we formulate measures of saliency for an audiovisual stream. Audio saliency is captured by signal modulations and related multi-frequency band features, extracted through nonlinear operators and energy tracking. Visual saliency is measured by means of a spatiotemporal attention model driven by various feature cues (intensity, color, motion). Audio and video curves are integrated in a single attention curve, where events may be enhanced, suppressed or vanished. The presence of salient events is signified on this audiovisual curve by geometrical features such as local extrema, sharp transition points and level sets. An audiovisual saliency-based movie summarization algorithm is proposed and evaluated. The algorithm is shown to perform very well in terms of summary informativeness and enjoyability for movie clips of various genres.

Index Terms— audio processing, video processing, audiovisual saliency, movie summarization

1. INTRODUCTION

Multimodal analysis, i.e, the concurrent analysis of multiple information modalities, has noted considerable progress for accessing and analyzing video content, with automatic video content summarization being an important application. Summaries provide the user with a short version of the video that ideally contains all important information for understanding the content, serving as a preview, an overview or a query object. Hence, the user may quickly access and evaluate if the video is important, interesting or enjoyable. In [1] video abstraction is classified into *keyframe selection*, which yields a static small set of important video frames and *video skimming* or *summarization*) which results in a dynamic, short subclip.

Earlier works were mainly based on processing only the visual input. Zhuang et al. [2] extracted salient frames based on color clustering and global motion, while Ju et al. [3] used gesture analysis in addition to the latter low-level features. Furthermore Avrithis et al. [4] represent the video content by a high-dimensional feature curve and detect key-frames at the curvature points. Another group of methods is based on frame clustering to select representative frames [5, 6]. Other schemes based on sophisticated temporal sampling [7], hierarchical frame clustering [5, 8], where the video frames are hierarchically clustered by visual similarity, and fuzzy classification [9] have also proposed summarization schemes with encouraging results.

In an attempt to incorporate multimodal or/and perceptual features in the analysis and processing of the visual input, various systems have been designed and implemented within a variety of projects. The Informedia project [10] and its offsprings, e.g. the Video Browsing and Retrieval system (VIRE), [11], MediaMill [12] and IBMs CueVideo [13] combined speech, image, natural language understanding and image processing to automatically index video for intelligent search and retrieval. On a step further towards human perception, Ma et al. [14] proposed a method for detecting the salient parts of a video that is based on user attention models. Motion, face and camera attention along with audio attention models (audio saliency and speech/music) were cues to capture salient information and identify the segments to compose a summary.

We present a saliency-based method to detect important audiovisual segments and focus on the potential benefits of feature-based attention modeling and multi-sensory signal integration. A video summarization algorithm is developed on the basis of salient segments w.r.t. a skimming percentage. As content importance in a video stream is quite subjective, systematic evaluations of summaries is required [15]. In this work, preliminary video summarization results are given on samples from the a movie database, annotated with respect to saliency by human observers [16]. Comparisons of automatic versus manual saliency annotation are presented, as well as subjective user ratings on enjoyability and informativeness of constructed summaries.

2. AUDIO SALIENCY

Streams of audio information may be composed from a variety of sounds, like speech, music, environmental sounds (nature, machines, noises), a result of multiple sources. Aural attention is triggered perceptually by changes in the involved events of an audio stream. These may be changes of the nature/source of events, newly introduced sounds, or transitions and abnormalities in the course of a specific event.Based on biological and perceptual observations, we construct an audio attention-modeling curve from saliency measures of meaningful temporal modulations in multiple frequencies [17].

The AM-FM model for speech [18] is generalized to any source producing oscillating signals and for that purpose it is used here to describe a large family of audio signals. Speech, music, noise, natural and mechanical sounds are the result of resonating sources are modeled as sums of amplitude and frequency (AM-FM) modulated components. The salient structures then are the underlying modulation signals and their properties (i.e., number, scale, importance) define the audio representation.

The audio signal is modeled by a sum of narrowband amplitude and frequency varying, non-stationary sinusoids $s(t) = \sum_{k=1}^{K} a_k(t)\cos\left(\phi_k(t)\right)$, whose demodulation in instantaneous amplitude $a_k(t)$ and frequency $\omega_k(t) = d\phi_k(t)/dt$ is obtained in the output of a set of frequency-tuned Gabor filters $h_k(t)$ using the energy operator Ψ and the ESA. The filters globally separate

This work was supported by grant ΠΕΝΕΔ-2003-ΕΔ866 [cofinanced by E.U.-European Social Fund (75%) and the Greek Ministry of Development-GSRT (25%)], the European Union FP6-IST Network of Excellence 'MUSCLE' and in part by the EU FP6 research program 'HIWIRE'.

modulation components assuming a priori a fixed component configuration. To model a discrete-time audio signal $s[n] = s(nT)$, we use K discrete AM-FM components whose instantaneous amplitude and frequency signals are $A_k[n] = a_k(nT)$ and $\Omega_k[n] = T\omega_k(nT)$, respectively. The model parameters are estimated from the K filtered components using a discrete-time energy operator $\Psi_d(x[n]) \equiv (x[n])^2 - x[n-1]x[n+1]$ and a related discrete ESA.

A representation in terms of a single component per analysis frame emerges by maximizing an energy criterion in the multi-dimensional filter response space [19]. For each frame m of N samples duration, the dominant modulation component is the one with *maximum average Teager energy* (MTE):

$$\mathrm{MTE}[m] = \max_{1\le k\le K} \frac{1}{N} \sum_{n} \Psi_d((s * h_k)[n]), \qquad (1)$$

where $(m-1)N+1 \le n \le mN$, $*$ denotes convolution and h_k the impulse response of the kth filter. The filter $j = \arg\max_k(\mathrm{MTE})$ is submitted to demodulation via ESA and the instantaneous modulating signals are averaged over a frame duration[1] to derive the *mean instant amplitude* (MIA) and *mean instant frequency* (MIF) features:

$$\mathrm{MIA}[m] = (\overline{|A_j[n]|}) \ , \ \mathrm{MIF}[m] = (\overline{\Omega_j[n]}). \qquad (2)$$

Thus, each frame yields average measurements for the source energy, instant amplitude and frequency from the filter that captures the "strongest" modulation signal component.

The resulting three-dimensional feature vector of the mean dominant modulation parameters $\vec{F}_{\mathrm{a}}[m] = [\mathrm{MTE}, \mathrm{MIA}, \mathrm{MIF}]\,[m]$ is a low dimensional descriptor, compared to the potential $3 \times K$ vector from all outputs, of the "average instantaneous" modulation structure of the audio signal involving properties such as level of excitation, rate-of-change, frequency content and source energy. The simplest scenario of an *audio saliency* curve is a weighted linear combination of the normalized audio features

$$S_{\mathrm{a}}[m] = w_1\mathrm{MTE}[m] + w_2\mathrm{MIA}[m] + w_3\mathrm{MIF}[m]. \qquad (3)$$

Here we perform equal weighting $w_i = 1/3$, $i = 1, 2, 3$ and normalization by least squares fit of their individual value ranges to $[0, 1]$.

3. VISUAL SALIENCY

Computation of visual saliency is based on the notion of a centralized saliency map [20] initiated by a feature competition scheme. The motivation behind this scheme is the experimental evidence of a biological counterpart in the Human Visual System. 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 saliency value. This saliency volume is computed with the incorporation of feature competition by defining cliques at the voxel level and use an optimization procedure with both inter- and intra- feature constraints.

The video volume is initially decomposed into a set of feature volumes, namely intensity, color and spatiotemporal orientations. For the intensity and color features, we adopt the opponent process color theory that suggests the control of color perception by two opponent systems: a blue-yellow (BY) and a red-green (RG) mechanism. Spatiotemporal orientations are computed using steerable filters [21], by measuring the filter strength along particular directions θ (the angle formed by the plane passing through the t axis and the $x-t$ plane) and ϕ (defined on the $x-y$ plane). The desired filtering can be implemented using the three dimensional filters $G_2^{\theta,\phi}$ (e.g second derivative of a 3D Gaussian) and their Hilbert transforms $H_2^{\theta,\phi}$, by taking the filters in quadrature to eliminate the phase sensitivity present in each output. This is called the oriented energy:

$$E(\theta,\phi) = [G_2^{\theta,\phi} * I]^2 + [H_2^{\theta,\phi} * I]^2, \qquad (4)$$

where $\theta \in \{0, \frac{\pi}{4}, \frac{\pi}{2}, \frac{3\pi}{4}\}$ and $\phi \in \{-\frac{\pi}{2}, -\frac{\pi}{4}, 0, \frac{\pi}{4}, \frac{\pi}{2}\}$. By selecting θ and ϕ as above, 20 volumes of different spatiotemporal orientations are produced that must be fused together to produce a single orientation volume that will be further enhanced and compete with the rest of the feature volumes. We use an operator based on Principal Component Analysis and generate a single spatiotemporal orientation conspicuity volume. More details can be found in [22].

To obtain a *visual attention curve*, we first 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. Volumes for each feature of interest, including intensity, color and 3D orientation (motion) are then formed 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. The energy involves voxel operations between coarse and finer scales of the volume pyramid, which means that if the center is a voxel at level $c \in \{2, ..., p-d\}$, where p is the maximum pyramid level and d is the desired depth of the center-surround scheme, then the surround is the corresponding voxel at level $h = c + \delta$ with $\delta \in \{1, 2, ..., d\}$. Hence, if we consider the intensity and two opponent color features as elements of the vector $\vec{F}_{\mathrm{v}} = F_{\mathrm{v}_1}, F_{\mathrm{v}_2}, F_{\mathrm{v}_3}$ and if $F^0_{\mathrm{v}_k}$ corresponds to the original volume of each of the features, each level ℓ of the pyramid is obtained by convolution with an isotropic 3D Gaussian G and dyadic down-sampling $F^{\ell}_{\mathrm{v}_k} = \left(G * F^{\ell-1}_{\mathrm{v}_k}\right) \downarrow_2, \quad \ell = 1, 2, ..., p$, where $\downarrow_2$ denotes decimation by 2 in each dimension. For each voxel q of a feature volume F the energy is defined as

$$E_v(F^c_{\mathrm{v}_k}(q)) = \lambda_1 \cdot E_1(F^c_{\mathrm{v}_k}(q)) + \lambda_2 \cdot E_2(F^c_{\mathrm{v}_k}(q)), \qquad (5)$$

where λ_1, λ_2 are the importance weighting factors. The first term, which may be regarded as the *data-term* is defined as

$$E_1(F^c_{\mathrm{v}_k}(q)) = F^c_{\mathrm{v}_k}(q) \cdot |F^c_{\mathrm{v}_k}(q) - F^h_{\mathrm{v}_k}(q)| \qquad (6)$$

and acts as the center-surround operator. The difference at each voxel is obtained after interpolating $F^h_{\mathrm{v}_k}$ to the size of the coarser level. This term promotes areas that differ from their spatiotemporal surroundings and therefore attract attention. The second, *smoothness*, term is defined as

$$E_2(F^c_{\mathrm{v}_k}(q)) = F^c_{\mathrm{v}_k}(q) \cdot \frac{1}{|N(q)|} \cdot \sum_{r\in N(q)} \left(F^c_{\mathrm{v}_k}(r) + V(r)\right), \qquad (7)$$

where V is the spatiotemporal orientation volume that may be regarded as an indication of motion activity in the scene and $N(q)$ is the 26- neighborhood of voxel q. The second energy term involves competition among voxel neighborhoods of the same volume and allows a voxel to increase its saliency value only if the activity of its surroundings is low enough. By iterative energy minimization, a *saliency volume* S is created by averaging the conspicuity feature

[1] Audio analysis frame duration is typically 20 ms. Central frequency steps of the filter design varying between 200-400 Hz, yield filterbanks consisting of 20-40 filters.

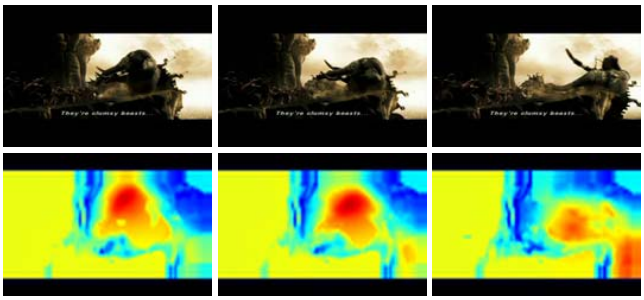

Fig. 1. Frames and corresponding saliency maps from movie "300".

volumes $F^1_{\mathrm{v}_k}$ at the first pyramid level $S(q) = \frac{1}{3} \cdot \sum_{k=1}^{3} F^1_{\mathrm{v}_k}(q)$. Fig. 1 depicts the computed saliency for three frames of "300" movie (from [16]). High values correspond to high salient areas (notice the falling elephant).

In order to create a single value per frame, each of the feature volumes is first normalized to lie in the range $[0, 1]$ and then point-to-point multiplied by the saliency one in order to suppress low saliency voxels. The weighted average is taken to produce a single *visual saliency* value for each frame:

$$S_{\mathrm{v}} = \sum_{k=1}^{3} \sum_{q} S(q) \cdot F^1_{\mathrm{v}_k}(q), \tag{8}$$

where the second sum is taken over all the voxels of a volume at the first pyramid level.

4. AUDIOVISUAL SALIENCY

Audiovisual fusion for modeling multimodal attention can be performed at three levels: i) *low-level* fusion (at the extracted saliency curves), ii) *middle-level* fusion (at the corresponding feature vectors), iii) *high-level* fusion (at the detected salient segments and features of the curves). In a video stream with both aural and visual information present, attention may be modeled by constructing a temporal index of audiovisual saliency. For both modalities, features are mapped to saliency (aural and visual) curve values $(S_{\mathrm{a}}[m], S_{\mathrm{v}}[m])$, and the two curves are integrated to yield an *audiovisual saliency curve* $S_{\mathrm{av}}[m] = \mathrm{fusion}(S_{\mathrm{a}}, S_{\mathrm{v}}, m)$, where m is the the feature sequence time index and $\mathrm{fusion}(\cdot)$ is the process of combining or fusing the two modalities. This is a low-level fusion scheme. Here, we use a straightforward linear memoryless scheme for creating the composite *audiovisual saliency curve* (AVSC) as $S_{\mathrm{av}} = w_{\mathrm{a}} \cdot S_{\mathrm{a}} + w_{\mathrm{v}} \cdot S_{\mathrm{v}}$. This coupled audiovisual curve serves as a continuous-valued indicator function of salient events, in the audio, the video or a common audiovisual domain. Equal weights are used following normalization of the audio and video saliency curves in $[0, 1]$.

Audio-visual events are defined as bounded time-regions of aural and visual activity. The boundaries of events and the activity locus points, correspond to a maximum change in the audio and video saliency curves and the underlying features. Such points are given by the form of the saliency and its various geometric characteristics such as *extrema* points, *peaks*, *edges* and *level sets*, i.e., values of the curve above a learned or heuristic level-threshold. Saliency-based events are tracked on the integrated audiovisual curve. Static video abstracts, i.e., keyframe collections, have been formed based on the local maxima of the saliency curve [17].

5. MOVIE SUMMARIZATION ALGORITHM

Dynamic video summarization is based on the attentional importance given by the associated audiovisual saliency curve. The segment selection and movie rendering algorithm follows the steps:

1. The AVSC is median filtered with a median filter of length $2M + 1$ video frames.
2. A saliency threshold S_c is selected so that the required *percent of summarization* c is achieved. Frames n with AV-saliency value $S_{\mathrm{av}}(n) > S_c$ are selected to be included in the summary. For example, for 20% summarization, $c = 0.2$, the threshold S_c is selected so that the cardinality of the set of selected frames $D = \{n : S_{\mathrm{av}}(n) > S_c\}$ is 20% of the total number of frames[2]. The result from this processing step is a video frame indicator function I_c for the desired level of summarization c. The indicator function equals 1, $I_c(n) = 1$, if frame n is selected for the summary and 0 otherwise.
3. The selected frames are joined into segments. Selected segments that are shorter than N frames are considered too short to be presented to the viewer and are deleted from the summary. This is almost equivalent with the morphological closing of the indicator function I_c with a vector of 1's of length $N + 1$.
4. Neighboring segments that are selected for the summary are joined together if they are less than K frames apart, i.e., at most K non-selected frames that are preceded and followed by selected frames are added to the summary. This is equivalent to the morphological opening of the indicator function resulting from the previous step with a vector of 1's of length $K + 1$.
5. Selected segments are rendered into a summary using simple overlap-add to tailor together neighboring segments. Linear overlap-add is applied on L video frames and the corresponding number of audio samples. Note that video and audio processing is synchronous in all steps.

The evaluated version of the algorithm operates with $M = N = 20$ frames, $K = L = 10$ frames for videos at 25 frames per second.

6. EVALUATION

We demonstrate the proposed methods for audiovisual saliency detection and video summarization on three clips from the MUSCLE movie database [16], i.e., scenes ($\simeq$10 min) from the movies "Lord of the Rings I" (LOTR1), "300" and "Cold Mountain" (CM).

In Fig. 2 we present comparisons of automatically vs. manually derived audiovisual saliency indicator functions for 600 frames of one movie clip. Ground truth, shown in the second row, resulted from manual, human annotation according to the attention the video content attracted. Automatic computation of salient regions was done by thresholding the median-filtered audiovisual curve, using three different filter lengths (7, 19 and 39 frames for a 25 fps video rate). Overall, good agreement was between the curve-based and manual annotation saliency annotation, especially for the longer filter, where correct frame classification (as salient or non-salient) attains 80%.

Summaries obtained for $c = 0.5, 0.33, 0.2$, i.e., skimming 2, 3 and 5 times faster than real times, where subjectively evaluated in terms of informativeness and enjoyability in a scale from 0-100. Note that static, keyframe, abstracts gave less than 5% summarizations, (e.g. $c = 0.45$ for LOTR1). 10 naive subjects were first shown

[2]The saliency threshold is selected globally for short video clips. For long video clips a segment-based threshold might perform better for movie skimming applications.

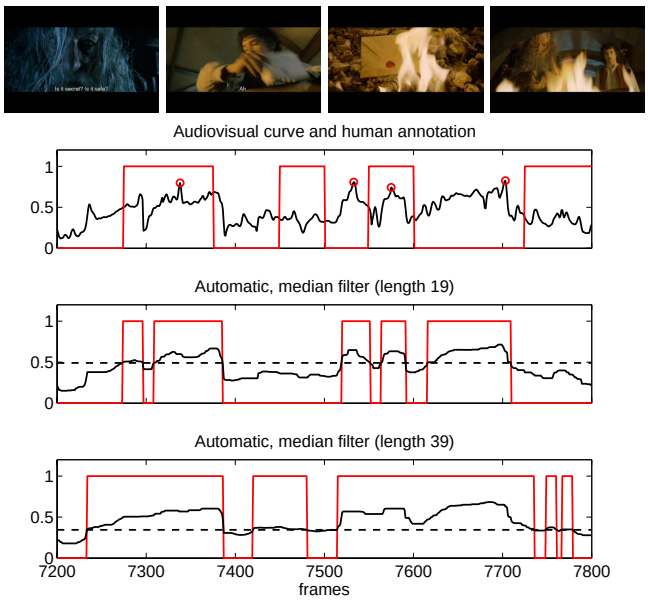

Fig. 2. Main keyframes and human vs. automatic saliency annotations for median-filtered audiovisual curves.

the original clip followed by its three summaries. Each viewer's score was normalized by dividing it with the score given to the original clip. The average results shown next indicate that the summaries obtained by the algorithm are well informative and enjoyable. However, more work is needed to improve the "smoothness" of the summary to increase enjoyability and include syntax constraints to increase comprehensibility.

Movie Clip	**Informativeness** original	$c = 0.5$	$c = 0.33$	$c = 0.2$
LOTR1	100	77.8	66	54.6
300	100	81.8	64.9	53
CM	100	78.8	69.4	55.8
Movie Clip	**Enjoyability** original	$c = 0.5$	$c = 0.33$	$c = 0.2$
LOTR1	100	82	77	65.2
300	100	84.8	77.8	63.7
CM	100	89.9	78.5	68.8

7. CONCLUSIONS

Based on efficient audio and image attention modeling, we presented saliency curves for the aural and visual streams of videos and explored the potential of their integration for a movie summarization application. A simple fusion scheme was employed to create audiovisual saliency curves that were used by a movie summarization algorithm. The algorithm is generic and independent of the video semantics, syntax or genre. Subjective evaluation showed that highly informative video summaries can be obtained using audiovisual saliency indicator functions. The performance of the algorithm is impressive in terms of summary informativeness given that no high-level features, e.g., plot, are used by the movie summarizer. Future work includes improved fusion algorithms for the audio and visual saliency curves, e.g., learning schemes and non-linear feature correlations, the incorporation of high-level features for salient event detection, e.g., movie transcript information, and further experimentation and systematic evaluation of the summarization algorithm.

8. REFERENCES

[1] L. Ying, S.-H. Lee, C.-H. Yeh, and C.-C.J. Kuo, "Techniques for movie content analysis and skimming," in *IEEE Signal Processing Magazine*, Mar 2006, vol. 23, pp. 79–89.

[2] Y. Zhuang, Y. Rui, T.S. Huang, and S. Mehrotra, "Adaptive key frame extraction using unsupervised clustering," in *Proc. IEEE Int'l Conf. Image Processing (ICIP)*, 1998, pp. 866–870.

[3] S. X. Ju, M. J. Black, S. Minneman, and D. Kimber, "Summarization of videotaped presentations: automatic analysis of motion and gesture," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 8, no. 5, pp. 686–696, 1998.

[4] Avrithis Y., Doulamis A., Doulamis N., and Kollias S., "A stochastic framework for optimal key frame extraction from mpeg video databases," *Comp. Vision and Image Understasting*, vol. 75, no. 12, pp. 3–24, 1998.

[5] K. Ratakonda, M.I. Sezan, and R.J. Crinon, "Hierarchical video summarization," in *Proc. SPIE, Visual Comm. and Image Proc. '99*, Dec 1998, vol. 3653, pp. 1531–1541.

[6] S. Uchihashi, J. Foote, A. Girgensohn, and J. Boreczky, "Video Manga: generating semantically meaningful video summaries," in *Proc. 7th ACM MULTIMEDIA*, 1999, pp. 383–392.

[7] X.D. Sun and M.S. Kankanhalli, "Video summarization using r-sequences," *Real-time imaging*, vol. 6, no. 6, pp. 449–459, Dec 2000.

[8] A. Girgensohn, J. Boreczky, and L. Wilcox, "Keyframe-based user interfaces for digital video," *IEEE Computer Magazine*, vol. 34, no. 9, pp. 61–67, Sep 2001.

[9] A. Doulamis, N. Doulamis, Y. Avrithis, and S. Kollias, "A fuzzy video content representation for video summarization and content-based retrieval," *Signal Processing*, vol. 80, no. 6, pp. 1049–1067, Jun 2000.

[10] A.G. Hauptmann, "Lessons for the future from a decade of Informedia video analysis research," in *Proc. Intl. Conf. on Image and Video Retrieval (CIVR), LNCS*, 2005, vol. 3568, pp. 1–10.

[11] M. Rautiainen et al., "TREC 2002 video track experiments at MediaTeam Oulu and VTT," in *Proc. Text Retrieval Conf. (TREC)*, 2002.

[12] S. Raaijmakers, J. Den Hartog, and J. Baan, "Multimodal topic segmentation and classification of news video," in *Proc. Text Retrieval Conf. (TREC)*, 2002, vol. 2, pp. 33–36.

[13] B. Adams et al., "IBM research TREC-2002 video retrieval system," in *Proc. Text Retrieval Conf. (TREC)*, 2002.

[14] Y. Ma, X.S. Hua, L. Lu, and H. Zhang, "A generic framework of user attention model and its application in video summarization," *IEEE Trans. Multimedia*, vol. 7, no. 5, pp. 907–919, Oct 2005.

[15] P. Over, A. F. Smeaton, and P. Kelly, "The TRECVID 2007 BBC rushes summarization evaluation pilot," in *TVS '07*, 2007, pp. 1–15.

[16] "MUSCLE Movie DataBase v3.0," 2007, `http://poseidon.csd.auth.gr/EN/MUSCLE_moviedb`.

[17] G. Evangelopoulos, K. Rapantzikos, P. Maragos, Y. Avrithis, and A. Potamianos, "Audiovisual attention modeling and salient event detection," in *Multimodal Processing and Interaction: Audio, Video, Text*, P. Maragos, A. Potamianos, and P. Gross, Eds. Springer, 2008.

[18] P. Maragos, J.F. Kaiser, and T.F. Quatieri, "Energy separation in signal modulations with application to speech analysis," *IEEE Trans. Signal Processing*, vol. 41, no. 10, pp. 3024–3051, Oct 1993.

[19] G. Evangelopoulos and P. Maragos, "Multiband modulation energy tracking for noisy speech detection," *IEEE Trans. Audio Speech Language Processing*, vol. 14, no. 6, pp. 2024–2038, Nov 2006.

[20] C. Koch and S. Ullman, "Shifts in selective visual attention: towards the underlying neural circuitry," *Human Neurobiology*, vol. 4, no. 4, pp. 219–227, Jun 1985.

[21] W. T. Freeman and E.H. Adelson, "The design and use of steerable filters," *IEEE Trans. PAMI*, , no. 9, pp. 891–906, 1991.

[22] K. Rapantzikos, N. Tsapatsoulis, Y. Avrithis, and S. Kollias, ," *Signal processing: Image Communication*, submitted.