PREDICTING AUDIO-VISUAL SALIENT EVENTS BASED ON VISUAL, AUDIO AND TEXT MODALITIES FOR MOVIE SUMMARIZATION

P. Koutras, A. Zlatintsi, E. Iosif, A. Katsamanis, P. Maragos and A. Potamianos

School of E.C.E., National Technical University of Athens, 15773 Athens, Greece
Email: {pkoutras, nzlat, nkatsam, maragos}@cs.ntua.gr, {iosife,potam}@central.ntua.gr

ABSTRACT

In this paper, we present a new and improved synergistic approach to the problem of audio-visual salient event detection and movie summarization based on visual, audio and text modalities. Spatio-temporal visual saliency is estimated through a perceptually inspired frontend based on 3D (space, time) Gabor filters and frame-wise features are extracted from the saliency volumes. For the auditory salient event detection we extract features based on Teager-Kaiser Energy Operator, while text analysis incorporates part-of-speech tagging and affective modeling of single words on the movie subtitles. For the evaluation of the proposed system, we employ an elementary and non-parametric classification technique like KNN. Detection results are reported on the MovSum database, using objective evaluations against ground-truth denoting the perceptually salient events, and human evaluations of the movie summaries. Our evaluation verifies the appropriateness of the proposed methods compared to our baseline system. Finally, our newly proposed summarization algorithm produces summaries that consist of salient and meaningful events, also improving the comprehension of the semantics.

***Index Terms*—** Visual saliency, auditory saliency, affective text analysis, audio-visual salient events, movie summarization

1. INTRODUCTION

Summarization task refers to producing a shorter version of a video, which contains all the necessary information required for context understanding without sacrificing much of the original informativeness and enjoyability. Automatic summaries can be created with: 1) key-frames, which correspond to the most important video frames and represent a static storyboard, or 2) by video skims that include the most descriptive and informative video segments. Movies consist of visual, audio and textual streams, and many computational models have been proposed to estimate their multimodal saliency [1, 2, 3]. Moreover, movies contain a lot of semantic events, whose modeling is not always easy employing only bottom-up and data-driven techniques. In this paper, we deal with the prediction of audio-visual salient events, meaning salient event estimation in a multimodal stream. This task is closely related to visual saliency estimation, however we show that the two additional modalities of audio and text, are not only important but also necessary for producing informative and enjoyable movie summaries with smooth scene transitions.

Designing a complete summarization system requires multimodal saliency models. The early methods for video skimming were mainly based on visual features such as color or motion [4] and techniques from statistical pattern classification [5, 6, 7]. Methods for video synopsis are also based on a cost minimization [8, 9]. The most recent movie summarization systems usually employ multimodal saliency models, which in addition to visual saliency they also exploit auditory saliency and semantics. In [10], attention models for video summarization are used, i.e., motion, face, camera and static attention models for visual attention and energy based features for auditory saliency estimation, since loudness attracts people's attention. Both visual and audio feature extraction is performed in [11] for the analysis of objects and events and consequently semantically oriented video summarization. An audiovisual model is also used in [12] so as to assist visual saliency in predicting eye movements in dynamic conversations. For more general reviews about movie summarization see also [1, 2, 13, 14].

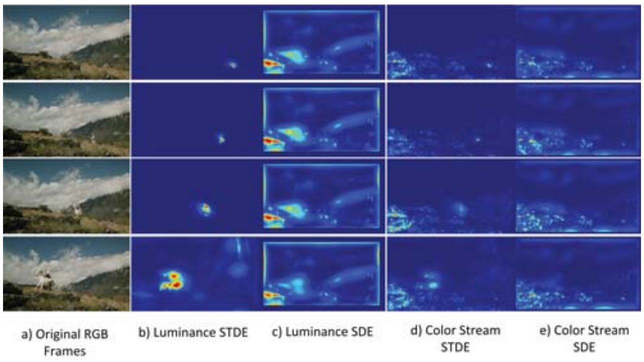
Analysis of text to estimate affect or sentiment is a relatively recent research topic that has attracted great interest with application to numerous domains spanning from tweet analysis [15] to dialogue systems [16]. Text can be analyzed at different levels of granularity: from single words to entire sentences. In [17], the affective ratings of unknown words were estimated using the affective ratings for a small set of words (seeds) and the semantic relatedness between the unknown and the seed words. A sentence-level approach was proposed in [18] applying techniques from n-gram language modeling.

In this work, we present an extension of our baseline multimodal saliency-based movie summarization algorithm [14]. In addition to the proposed spatio-temporal visual saliency model described in Sec. 2, we introduce new frameworks for both the audio and text modality so as to enhance the detection of salient events. Section 3 describes the audio features which are based on energy tracking and other perceptual features that correlate to the human perception of sound. In Sec. 4 we present the text analysis which is based on part-of-speech tagging and affective modeling of single words found on the subtitles of the movies. The problem of audio-visual salient event detection is approached using a machine learning technique, Sec. 5, where a frame-wise classification of salient vs. non salient events is taking place in order to validate the efficiency of the proposed algorithms. In Sec. 5 we also introduce a new movie summarization algorithm which resulted, according to subjective evaluations, to both informative and enjoyable summaries.

2. VISUAL ANALYSIS

Our energy-based model for spatio-temporal visual saliency estimation is more relevant to the cognition-inspired saliency methods, based on Koch & Ullman theory [19], as it was implemented in [20]. It uses biologically plausible spatio-temporal filters, like oriented 3D Gabor filters, in order to extract visual features. In a first phase the initial RGB video volume is transformed into Lab space

This research work was supported by the project "COGNIMUSE" which is implemented under the "ARISTEIA" Action of the Operational Program Education and Lifelong Learning and is co-funded by the European Social Fund and Greek National Resources.



a) Original RGB Frames b) Luminance STDE c) Luminance SDE d) Color Stream STDE e) Color Stream SDE

Fig. 1: Example frames of the four energy volumes computed using our visual frontend on the *Lord of the Rings*.

and split into two streams: luminance and color stream. We use the CIE-Lab color space because in this space luminance and chromaticity components can be well separated while it has the additional property to be perceptually uniform [21]. In the resulting video volume $\mathbf{I}_{Lab}(x,y,t)$ the L^* component expresses the perceptual response to luminance, while a^*, b^* describe differences between red-green and yellow-blue colors respectively. We have also employed an approach that models the double color opponent cells that exist in primary visual cortex V1 and has been used in color constancy applications [22]. Instead of using the R, G, B components, we use the chromaticity components (a^*, b^*) that indirectly include the $R-G$ and $B-Y$ differences. The resulting color stream that expresses both the color intensity and contrast is given by: $C_{ab}(x,y,t) = \sqrt{(a^*(x,y,t))^2 + (b^*(x,y,t))^2}$. Then follows the filtering process [23] which is applied both on luminance and color stream channels.

3D Gabor Filtering: For the filtering process of the video's luminance we choose to use oriented Gabor filters in a spatio-temporal version, due to their biological plausibility and their uncertainty-based optimality [24, 25]. Specifically, we apply quadrature pairs of 3D (spatio-temporal) Gabor filters with identical central frequencies and bandwidth. These filters can arise from 1D Gabor filters [24] in a similar way as Daugman proposed 2D Oriented Gabor Filters [26]. An 1D complex Gabor filter consists of a complex sine wave modulated by a Gaussian window. Its impulse response with unity norm has the form:

$$g(t) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{t^2}{2\sigma^2}\right) \exp(j\omega_{t_0} t) = g_c(t) + jg_s(t). \quad (1)$$

The above complex filter can be split into one odd(sin)-phase ($g_s(t)$) and one even(cos)-phase ($g_c(t)$) filters, which forms a quadrature pair filter.

The 3D Gabor extension [27] yields an *even (cos)* 3D Gabor filter whose impulse response is:

$$g_c(x,y,t) = \frac{1}{(2\pi)^{3/2}\sigma_x\sigma_y\sigma_t} \exp\left[-\left(\frac{x^2}{2\sigma_x^2} + \frac{y^2}{2\sigma_y^2} + \frac{t^2}{2\sigma_t^2}\right)\right] \cdot \cos(\omega_{x_0} x + \omega_{y_0} y + \omega_{t_0} t), \quad (2)$$

where $\omega_{x_0}, \omega_{y_0}, \omega_{t_0}$ are the spatial and temporal angular center frequencies and $\sigma_x, \sigma_y, \sigma_t$ are the standard deviations of the 3D Gaussian envelope. Similarly for the impulse response of *odd (sin)* filter which we denote by $g_s(x,y,t)$.

The 3D filtering is a time consuming process due to the complexity of all required 3D convolutions. However, Gabor filters are separable [27], which means that we can filter each dimension separately using an impulse response having the form of (1). In this way, we apply only 1D convolutions instead of 3D [23].

For the spatio-temporal filterbank we used $K_G = 400$ Gabor filters (isotropic in the spatial components) which are arranged in five spatial scales, eight spatial orientations and ten temporal frequencies. The spatial scales and orientations are selected to cover a squared 2D frequency plane in a similar way to the design by Havlicek et al. [28]. We also use ten temporal Gabor filters, five at positive and five at negative center frequencies due to the 3D spectrum symmetries. For the static (only spatial) filterbank we use the same spatial parameters with zero temporal frequency ($L_G = 40$ filters). The spatio-temporal filterbank can detect motion activities, while the static one can find significant image regions which may attract human attention such as specific texture or strong edges.

Postprocessing: After the filtering process, for each filter i we obtain a quadrature pair output $(y_s^{3D}(x,y,t), y_c^{3D}(x,y,t))$ which corresponds to the even- and odd-phase 3D filter outputs. For each filter we can compute the total Gabor energy $E(\cdot)$, which is invariant to the phase of the input, by taking the sum of the squared energy of these two outputs: $E(y_s^{3D}, y_c^{3D}) = \left(y_s^{3D}(x,y,t)\right)^2 + \left(y_c^{3D}(x,y,t)\right)^2$.

After applying the above energy operator to each filter we have K_G *energy volumes* for the spatio-temporal part (STE_i) and L_G for the static part (SE_i). In order to form one volume for each of these independent filtering parts we apply the first step of *Dominant Component Analysis* [28, 29] both to spatio-temporal and static energy volumes. Specifically, for each voxel (x,y,t) we keep its maximum value between all existing energy volumes: $STDE = \max_{1\leq i\leq K_G} STE_i$, $SDE = \max_{1\leq i\leq L_G} SE_i$. Instead of keeping only the dominant energy we can keep the $N_B = 6$ highest spatio-temporal energies for each voxel and afterwards compute the min value of them. Finally, we have two raw energy volumes for each luminance and color stream: spatio-temporal dominant energy $STDE$ (see Fig. 1b,d) and static dominant energy SDE (see Fig. 1c,e). These energy volumes can become further smoothed by applying a *temporal moving average* (TMA). Thus, each frame energy is computed as the mean inside a temporal window which includes N_T successive frames whose total duration is 1 second. In this way, we integrate visual events which take place close in time, in a similar way that humans are believed to do.

The produced energy maps can be mapped to a 1D map giving time-varying saliency features. We employed a simple 3D to 1D mapping by taking the mean value for each 2D frame slice of each 3D energy volume. The resulting temporal sequence of feature vectors, each corresponding to the 4 different TMA energies, along with its first and second time derivatives comprise the feature set for the visual modality.

3. AUDIO ANALYSIS

The issue of saliency computation in the audio stream is approached as a problem of assigning a measure of interest to audio frames, based on spectro-temporal cues. The importance of amplitude and frequency changes for audio saliency has motivated various studies where subject responses are measured with respect to tones of modulated frequency or loudness [30, 31, 32].

Extensive experimentation with different configurations, for the analysis of the audio stream, leaded to an energy-based feature set for the saliency-modeling of the audio stream, which was approached using the nonlinear differential energy operator proposed by Teager [33] and further investigated by Kaiser [34]. The Teager-Kaiser Energy Operator (TEO), which can track the instantaneous energy of a source, is given by

$$\Psi[x] = \dot{x}^2 - x\ddot{x}, \text{ where } \dot{x} = dx/dt. \quad (3)$$

Since Teager energy is only meaningful in narrowband signals, the application of the operator is preceded by bandpass filtering; specifically filtering of the signal with a mel-spaced Gabor filterbank consiting of 25 filters with 50% overlap of the successive filters [35]. The energy features that are used in this paper are the mean instantaneous energies derived for each Gabor filter.

Moreover, we computed two additional perceptual features which are assumed to correlate to the functioning of the human auditory system. The first one is roughness proposed in [36] and reported to be associated with human attention [37]; which is an estimation of the sensory dissonance of a sound. It expresses a sense of roughness of a sound due to rapid fluctuations in its amplitude and is related to the beating phenomenon whenever pairs of sinusoids are close in frequency. An estimation of roughness can be given by computation of the peaks of the spectrum followed by averaging among all possible pairwise combinations of peaks [38]. In this paper, we use a variant model that uses a more complex weighting [39]. The second perceptual feature used in this work is loudness, also associated to attention, which corresponds to the perceived sound pressure level. For the computation of loudness the model proposed in [37] was used, which is based on the calculation of the excitation on the basilar membrane taking into account phenomena such as the temporal frequency masking.

4. AFFECTIVE TEXT ANALYSIS

In this work, we extend the text analysis of [3] and we include affective modeling of single words extracted from the subtitles information available with each movie distribution. Our baseline text analysis is summarized in the following steps: (i) extraction of the movie transcript from the english subtitle file, ii) part-of-speech tagging, (iii) audio-text alignment, and (iv) assignment of a text saliency value to each frame based on the parser tag assigned to the corresponding word.

A word w is characterized regarding its affective content in a continuous (within the $[-1, 1]$ interval) space consisting of three dimensions (affective features), namely, valence (v), arousal (a), and dominance (d). For each dimension, the affective content of w is estimated as a linear combination of its semantic similarities to a set of K seed words and the corresponding affective ratings of seeds (for the corresponding dimension), as follows [18]:

$$\hat{u}(w) = \lambda_0 + \sum_{i=1}^{K} \lambda_i \; u(t_i) \; S(t_i, w), \tag{4}$$

where $t_1 ... t_K$ are the seed words, $u(t_i)$ is the affective rating for seed word t_i with u denoting one of the aforementioned dimensions, i.e., v, a, or d. λ_i is a trainable weight corresponding to seed t_i. $S(t_i, w)$ stands for a metric of semantic similarity between t_i and w. This model is based on the assumption that *"semantic similarity can be translated to affective similarity"* [18]. The $S(.)$ metric can be computed within the framework of (corpus-based) distributional semantic models that rely on the hypothesis that *"similarity of context implies similarity of meaning"* [40]. A contextual window of size $2H+1$ words is centered on the word of interest w_i and lexical features are extracted. For every instance of w_i in the (text) corpus the H words left and right of w_i formulate a feature vector x_i. For a given value of H the semantic similarity between two words, w_i and w_j, is computed as the cosine of their feature vectors: $Q^H(w_i, w_j) = \frac{x_i \cdot x_j}{||x_i|| \; ||x_j||}$. The elements of feature vectors can be weighted according to various schemes.

In this work, the context-based Q^H metric was applied with $H = 1$ over a web-harvested corpus, while the contextual features were weighted using a binary scheme. The word affective ratings were estimated using as seeds 600 entries selected from the ANEW lexicon [41]. More details about the corpus, seed selection, and the training of λ weights can be found in [18].

5. MACHINE LEARNING EVALUATION AND MOVIE SUMMARIZATION ALGORITHM

MovSum Database: We have evaluated our framework on seven movies from the Movie Summarization (MovSum) database [14], which consists of half-hour continuous segments (three and a half hours in total), namely: "Beautiful Mind" (BMI), "Chicago" (CHI), "Crash" (CRA), "The Departed" (DEP), "Gladiator" (GLA), "Lord of the Rings" (LOR) and the animation movie "Finding Nemo" (FNE). The movies were annotated by three expert viewers considering both monomodal and multimodal levels of saliency (incl. audio (A), visual (V) and audio-visual sensory (AV) level) and the sensory/semantic (AVS) level. Additionally, they were manually segmented into shots and scenes. The ground-truth framewise saliency, used for evaluation purposes, consists of frames that have been labeled salient by at least two labelers.

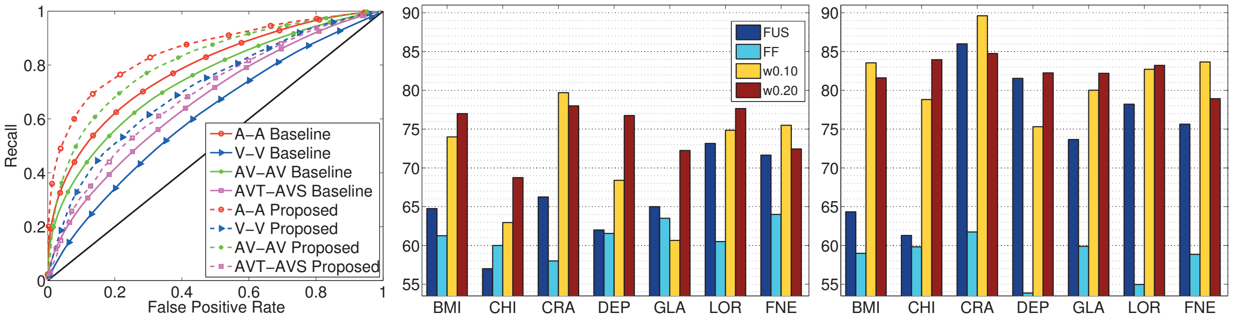
Machine Learning Approach: For the multimodal salient event detection we follow a simple classification approach, instead of experimenting with various fusion schemes as in [14]. The resulting audio-visual feature vectors (27 audio and 4 visual features) along with its first and second time derivatives (computed over 3 and 5 frames respectively) and the 4 text features, comprise the feature set for the classification process, where we employ a K-Nearest Neighbor Classifier (KNN); following the same framework as in [42, 14]. Specifically, we consider framewise saliency as a two-class classification problem, and a seven fold cross-validation is adopted by using the labeled frames from six movies and tested on the seventh. We also define a confidence score for each classification result, thus each frame, in order to obtain results for various compression rates.

5.1. Movie Summarization Algorithm

The new algorithm extends our baseline movie summarization algorithm [14] and it includes new features intending to make the summaries smoother, regarding audio and video transitions, while also adding better comprehension concerning the semantics.

For the creation of the summaries we use the classifier's output, which consists of the frames classified as salient. Thus, we use the segments/frames (chosen based on high confidence scores), as an indicator function curve that marks the most salient audio-visual-text events. The preprocessing steps that are followed are: 1) Median filtering of the audiovisual confidence scores C_{AV}, so as to obtain a smoother and coarse AV attention curve, followed by scene-based normalization (the boundaries of the scenes are extracted from the manual segmentation of the database). 2) The text confidence scores C_T that were only trained on speech segments are used; while frames without speech are set to zero. 3) Late fusion of the Audio-Visual and Text modalities is performed, where a weight w for the text stream is chosen: $C_{AVT} = C_{AV} + w \cdot C_T$. In this paper, we experimentally set the text weight to be $w = 0.10$ or $w = 0.20$.

In order to create summaries that do not include only salient events but "meaningful" as well, we perform boundary correction of the extracted events. This is achieved using ideas from mathematical morphology and specifically, the reconstruction opening: $\rho^-(M|X) \triangleq$ connected components of X intersecting M [43]. In such a way, we can extract large-scale components by detecting only smaller markers inside them. First, we use the boundaries of the manually segmented shots and then the single-word level boundaries of the automatically aligned text. More specifically, we use as marker M the salient events that are selected to be included in the final summary and as reference X the shots and the single words. The



(a) ROC Curves (b) Informativeness (c) Enjoyability

Fig. 2: Objective and subjective evaluation results by 20 humans. a) Saliency classification ROC curves for the different modalities. b) Informativeness and c) enjoyability results of AVT summaries at (×5) rate, were FUS denotes the best summary obtained using fusion methods as presented in [14], FF denotes "fast-forward", while the two newly produced summaries are differentiated by the weight used in the text modality.

reconstruction is regarded extremely significant for the performance of the summaries, as shown on the results of the human evaluation, especially for the comprehension of the semantics and the creation of smoother transitions [42].

The steps of the summarization algorithm are the following: a) sorting of the confidence scores so as to define the frames/segments that will be included in the summary, according to the number of frames needed for a five times (×5) faster summary than real time. b) Shot reconstruction, and c) "speech reconstruction" in order to assure that no words will be "clipped". d) The final step of the algorithm includes a process based on [3] for the combination of frames into segments. Thus, segments that are shorter than N frames are deleted from the summary, while neighboring segments selected for the summary are merged if they are less than K frames apart, where $N = 7$ and $K = 20$ (experimentally tuned). e) The final rendering of the selected segments into a summary is performed by using simple overlap-add on video frames to tailor together neighboring segments.

5.2. Results and Discussion

Objective Evaluation: Figure 2a shows ROC curves for saliency classification, while changing the percentage of frames in summary (100% percentage corresponds to perfect recall score), for audio on audio (A-A), visual on visual (V-V), audiovisual on audiovisual (AV-AV) and audiovisual-text on audio-visual-semantics (AVT-AVS) annotation. The results for the proposed method (AV-AV and AVT-AVS) are produced using the new summarization algorithm, while for the A-A and V-V results we use the sorted median filtered confidence scores. For the baseline method the results are shown for the sorted RAW confidence scores as presented in [14].

We note that the proposed system outperforms the baseline movie summarization system both when evaluating each modality individually as well as when two (AV) or three (AVT) modalities are combined. However, greater improvement can be seen for the monomodal salient event detection than the multimodal one. Regarding the proposed system best performance is accomplished for the audio modality (A-A evaluation). Moreover, we mark that the audiovisual modality (AV-AV) manages to yield a quite as high score as well. We have to highlight the fact that the classification approach used is a framewise detection task, while the human annotators marked salient events as segments and not as single frames.

Subjective Evaluation of Movie Summaries: Summaries obtained five times faster than real time were subjectively evaluated by 20 users in terms of informativeness and enjoyability on a 0–100% scale similarly to [14]. In total, four summaries were evaluated, namely: two summaries based on the proposed method using different weights for the text modality $w = 0.1$ or 0.2, our best performing summary produced using the fusion methods presented in [14] (the summaries were chosen based on enjoyability), and a fourth fast-forward like summary, which was created by subsampling 2 seconds every 10 seconds of the original clip. The subjects participating in the evaluation first viewed the original half-hour clip, for each of the movies, followed by the four summaries (ca. 6 min each) in randomized order.

As shown in Fig. 2b and 2c the proposed method performs much better in terms of both metrics compared to the best performing summaries based on fusion and the fast-forward like summaries; specifically up to 80% for informativeness and 90% for enjoyability. Regarding the proposed method we marked that the assignment of different weights in the text modality is important and it relates to the movie genre, usually a smaller weight is needed for a dialogue based movie than an action movie. Additionally, the reconstruction of shots and speech segments contributed a lot to the enjoyability, since it resulted to smoother transitions. Note that the subjects that participated in this evaluation were different than the ones in [14], that is why the results of the fusion based summaries may vary. Regarding the FF summaries none of the subjects realized that they were intentionally added for evaluation (as a naive approach indicating a lower bound for our metrics); and in this way we managed to show that a uniform sampling of movie frames is not adequate in order to create an acceptable summary.

6. CONCLUSIONS

A multimodal approach was adopted for perceptually salient event detection with application in movie summarization. We showed that our newly proposed spatio-temporal frontend for visual saliency estimation can be further improved when incorporating the two additional modalities of audio and text. Our experimental evaluation using a simple classifier confirms the adequacy of the proposed algorithms. The combined framework shows to be promising as it outperforms the baseline system over the saliency annotated MovSum database. The subjective evaluation of the automatic created movie summaries quantitatively verifies the appropriateness of both methods and our newly proposed movie summarization algorithm. For future work, we intend to further refine our methods and the movie summarization algorithm automating the weight selection for the text modality as well as the segmentation of shots and scenes.

Acknowledgments: The authors would like to thank E. Palogiannidi for the useful discussions on affective text analysis and the students of NTUA, CVSP Lab for participating in the subjective evaluation.

7. REFERENCES

[1] Y. Wang, Z. Liu, and J.-C. Huang, “Multimedia content analysis using both audio and visual clues,” *IEEE Signal Process. Mag.*, vol. 17, no. 6, pp. 12–36, Nov. 2000.

[2] 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. on Multimedia*, vol. 7, no. 5, pp. 907–919, Oct 2005.

[3] G. Evangelopoulos, A. Zlatintsi, G. Skoumas, K. Rapantzikos, A. Potamianos, P. Maragos, and Y. Avrithis, “Video event detection and summarization using audio, visual and text saliency,” in *Proc. Int’l Conf. Acoustics, Speech and Signal Processing*, 2009.

[4] Y. Zhuang, Y. Rui, T.S. Huang, and S. Mehrotra, “Adaptive key frame extraction using unsupervised clustering,” in *Proc. Int’l Conf. Image Processing*, 1998.

[5] A. Hanjalic and H. J. Zhang, “An integrated scheme for automated video abstraction based on unsupervised cluster-validity analysis,” *IEEE Trans. Circ. Syst. Video Technol.*, vol. 9, no. 8, 1999.

[6] Y. H. Gong and X. Liu, “Video summarization using singular value decomposition,” in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2000.

[7] X. Orriols and X. Binefa, “An EM algorithm for video summarization, generative model approach,” in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2001.

[8] Z. Li, P. Ishwar, and J. Konrad, “Video condensation by ribbon carving,” *IEEE Trans. on Image Processing*, vol. 18(11), pp. 2572–2583, 2009.

[9] Y. Pritch, A. Rav-Acha, and S. Peleg, “Nonchronological video synopsis and indexing,” *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 30, no. 11, pp. 1971–1984, 2008.

[10] Y.-F. Ma, L. Lu, H.-J. Zhang, and M. Li, “A user attention model for video summarization,” in *Proc. ACM Multimedia*, 2003.

[11] F. Wang and C.-W. Ngo, “Summarizing rushes videos by motion, object and event understanding,” *IEEE Trans. on Multimedia*, 2010.

[12] A. Coutrot and N. Cuyader, “How saliency, faces, and sound influence gaze in dynamic social scenes,” *Jour. of Vision*, vol. 14, no. 8, pp. 1–17, 2014.

[13] A. Money and H. Agius, “Video summarization: A conceptual framework and survey of the state of the art,” *J. Visual Communication and Image Representation*, vol. 19, no. 2, pp. 121–143, Feb. 2008.

[14] G. Evangelopoulos, A. Zlatintsi, A. Potamianos, P. Maragos, K. Rapantzikos, G. Skoumas, and Y. Avrithis, “Multimodal saliency and fusion for movie summarization based on aural, visual, textual attention,” *IEEE Trans. on Multimedia*, vol. 15, no. 7, pp. 1553–1568, Nov. 2013.

[15] P. Nakov, S. Rosenthal, Z. Kozareva, V. Stoyanov, A. Ritter, and T. Wilson, “Semeval 2013 task 2: Sentiment analysis in twitter,” in *Proc. of 2nd Joint Conf. on Lexical and Computational Semantics (*SEM), 7th Int’l. Workshop on Semantic Evaluation*, 2013, pp. 312–320.

[16] C. M. Lee and S. S. Narayanan, “Toward detecting emotions in spoken dialogs,” *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 2, pp. 293–303, 2005.

[17] P. Turney and M. L. Littman, “Unsupervised learning of semantic orientation from a hundred-billion-word corpus,” Tech. Rep. ERC-1094, National Research Council of Canada, 2002.

[18] N. Malandrakis, A. Potamianos, E. Iosif, and S. Narayanan, “Distributional semantic models for affective text analysis,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 11, pp. 2379–2392, 2013.

[19] 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.

[20] L. Itti, C. Koch, and E. Niebur, “A model of saliency-based visual attention for rapid scene analysis,” *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 20, no. 11, pp. 1254–1259, 1998.

[21] G. Wyszecki and W. S. Stiles, *Color Science*, J. Wiley & Sons, NY, 2nd edition, 1982.

[22] S. Gao, K. Yang, C. Li, and Y. Li, “A color constancy model with double-opponency mechanisms,” 2013.

[23] K. Maninis, P. Koutras, and P. Maragos, “Advances on action recognition in videos using and interest point detector based on multiband spatio-temporal energies,” in *Proc. Int’l Conf. Image Processing*, 2014.

[24] D. Gabor, “Theory of Communication,” *IEE Journal (London)*, vol. 93, pp. 429–457, 1946.

[25] J. G. Daugman, “Uncertainty Relation for Resolution in Space, Spatial Frequency and Orientation Optimized by Two-Dimensional Visual Cortical Filters,” *J. Opt. Soc. Amer. A*, vol. 2(7), pp. 1160–1169, 1985.

[26] J. G. Daugman, “Two-dimensional spectral analysis of cortical receptive field profiles,” *Vision Research*, vol. 20, no. 10, pp. 847–856, 1980.

[27] D. J. Heeger, “Model for the extraction of image flow,” *J. Opt. Soc. Amer.*, vol. 4, no. 8, pp. 1455–1471, 1987.

[28] J. P. Havlicek, D. S. Harding, and A. C. Bovik, “Multidimensional quasi-eigenfunction approximations and multicomponent am-fm models,” *IEEE Trans. Image Processing*, vol. 9, no. 2, pp. 227–242, 2000.

[29] A. C. Bovik, N. Gopal, T. Emmoth, and A. Restrepo, “Localized Measurement of Emergent Image Frequencies by Gabor Wavelets,” *IEEE Trans. Information Theory*, vol. 38, pp. 691–712, 1992.

[30] J. B. Fritz, M. Elhilali, S.V. David, and S.A. Shamma, “Auditory attention–focusing the searchlight on sound,” *Current opinion in neurobiology*, vol. 17, no. 4, pp. 437–455, Aug. 2007.

[31] 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.

[32] M. Elhilali, J. Xiang, S. A. Shamma, and J. Z. Simon, “Interaction between attention and bottom-up saliency mediates the representation of foreground and background in an auditory scene,” *PLoS biology*, vol. 7, no. 6, Jun. 2009.

[33] H. M. Teager and S. M. Teager, “Evidence for nonlinear sound production mechanisms in the vocal tract,” in *Speech Production and Speech Modelling*, W.J. Hardcastle and A. Marchal, Eds., vol. 15. NATO Advanced Study Institute, Series D, Boston, MA: Kluwer, Jult 1989.

[34] J. F. Kaiser, “On a simple algorithm to calculate the ’energy’ of a signal,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 1990.

[35] D. Dimitriadis, P. Maragos, and A. Potamianos, “On the effects of filterbank design and energy computation on robust speech recognition,” *IEEE Trans. on Audio, Speech and Language Process.*, vol. 9, no. 6, pp. 1504–1516, Aug. 2011.

[36] R. Plomp and W.J.M. Levelt, “Tonal consonance and critical bandwidth,” *Journal of the Acoustical Society of America (JASA)*, vol. 38, pp. 548–560, 1965.

[37] E. Zwicker and H. Fastl, *Psychoacoustics, Facts and Models*, Springer, 2nd edition, 1999.

[38] W. A. Sethares, *Tuning, Timbre, Spectrum, Scale,*, Springer-Verlag, 1998.

[39] P. N. Vassilakis, *Perceptual and Physical Properties of Amplitude Fluctuation and their Musical Significance*, Ph.D. thesis, Los Angeles: University of California, 2001.

[40] Z. Harris, “Distributional structure,” *Word*, vol. 10, no. 23, pp. 146–162, 1954.

[41] M. Bradley and P. Lang, “Affective norms for English words (ANEW): Stimuli, instruction manual and affective ratings. Tech. report C-1.,” The Center for Research in Psychophysiology, Univ. of Florida, 1999.

[42] A. Zlatintsi, P. Maragos, A. Potamianos, and G. Evangelopoulos, “A saliency-based approach to audio event detection and summarization,” in *Proc. European Signal Process. Conf.*, 2012.

[43] P. Maragos, *The Image and Video Processing Handbook*, chapter Morphological Filtering for Image Enhancement and Feature Detection, pp. 135–156, Elsevier Acad. Press, 2nd edition, 2005.