VIDEO EVENT DETECTION AND SUMMARIZATION USING AUDIO, VISUAL AND TEXT SALIENCY

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

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

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

ABSTRACT

Detection of perceptually important video events is formulated here on the basis of saliency models for the audio, visual and textual information conveyed in a video stream. Audio saliency is assessed by cues that quantify multifrequency waveform modulations, extracted through nonlinear operators and energy tracking. Visual saliency is measured through a spatiotemporal attention model driven by intensity, color and motion. Text saliency is extracted from part-of-speech tagging on the subtitles information available with most movie distributions. The various modality curves are integrated in a single attention curve, where the presence of an event may be signified in one or multiple domains. This multimodal saliency curve is the basis of a bottom-up video summarization algorithm, that refines results from unimodal or audiovisual-based skimming. The algorithm performs favorably for video summarization in terms of informativeness and enjoyability.

Index Terms— multimodal saliency, audio, video, text processing, video abstraction, movie summarization

1. INTRODUCTION

Video streams involve multiple information modalities that convey cues related to the nature and properties of the underlying events. For example, *visual* events may include objects, motions and scene changes, *aural* events can be changes in audio sources, while dialogues, subjects and key-words may be deemed *textual* events. Perceptual attention is triggered by changes in the involved events like scene transitions, progressions or newly introduced themes. Computational models of attention have been previously developed using multimodal analysis, i.e., the concurrent analysis of multiple information modalities [1, 2, 3, 4, 5]. Automatic video content access, analysis and abstraction have thus emerged as potential applications.

Video 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 [6]. Earlier works on video skimming were primarily based on processing the visual input and low-level features like color or motion [7]. Other skimming schemes like hierarchical frame clustering [8] or fuzzy classification [9] have also produced encouraging results. In an attempt to incorporate multimodal and/or perceptual features various systems have been designed and implemented. The Informedia project [10, 11] and its offsprings [12] combined speech and image processing with natural language understanding to automatically index video for intelligent search and retrieval.

Gaining insight from viewer behavior, user attention models were developed for detecting salient video parts. Motion, face, camera and audio attention models were cues to capture salient information and identify the segments to compose a summary [2]. In our previous work, saliency was modeled independently in each modality, using meaningful temporal modulations in multiple frequencies for the audio and spatiotemporal features (color, motion, intensity) for the visual stream [4, 5]. An integrated audiovisual saliency curve formed the basis of a bottom-up, content-independent, summarization technique.

In this work, we extend the audiovisual saliency-based video summarization algorithm in [5] to include text saliency automatically extracted from the subtitles information available with each movie distribution. For the computation of the frame-based text saliency metric the steps followed are: (i) extract the movie transcript from the subtitle file and perform shallow syntactic analysis including part-of-speech tagging, (ii) segment the audio stream using speech recognition technology to find the beginning and ending frame for each word in the transcript, and (iii) assign a text saliency value to each frame based on the parser tag assigned to the corresponding word. The audio, visual and text saliency scores are linearly combined to derive a multimodal saliency score for each frame. Similarly to audio and visual saliency, the text-based saliency metric uses only local information. As a result the movie summarization algorithm is easily scalable to different video content ranging from short movie clips to whole movies.

2. AUDIO ANALYSIS

The analysis and saliency-modeling of the audio stream is based on strong modulation structures of the signal waveform, using the AM-FM model for audio signals (speech, music, environmental sounds), i.e., $s(t) = \sum_{k=1}^{K} a_k(t)\cos\left(\phi_k(t)\right)$. The instantaneous amplitude $a_k(t)$ and frequency $\omega_k(t) = d\phi_k(t)/dt$ can be estimated from a set of Gabor filters $h_k(t)$ by applying the nonlinear energy operator Ψ and the energy separation algorithm. A compact, low-dimensional representation emerges by tracking the energy-dominant modulation component along multiple frequency bands [13], i.e., the component $j = j[m] \in \{1, 2, \ldots, K\}$ maximizing the average energy

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

Audio is described by means of this energy, along with the corresponding mean amplitude and frequency

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

This work was supported in part by the Greek research programs ΠΕΝΕΔ-2003-ΕΔ 865, 866, 554 [cofinanced by E.U.-European Social Fund (75%) and the Greek Ministry of Development-GSRT (25%)] and by the European Union FP6-IST Network of Excellence 'MUSCLE'.

where m the analysis frame index of N samples duration, n is the sample index in m, h_k the impulse response of the kth filter and $s[n] = s(nT)$, $A_k[n] = a_k(nT)$ and $\Omega_k[n] = T\omega_k(nT)$ are the discrete-time signals. More details and algorithmic implementations can be sought in [13, 4].

This 3D feature vector $\vec{F}_{\rm a}[m] = [{\rm MTE}, {\rm MIA}, {\rm MIF}]\,[m]$ conveys information on audio properties such as level of excitation, rate-of-change, frequency content and source energy, related to the presence and progression of audio events. Audio saliency is defined by a weighted linear integration of the normalized features

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

derived here with equal weights $w_i = 1/3, i = \{1, 2, 3\}$ and feature normalization by least squares fit of their individual values to $[0, 1]$.

3. VISUAL ANALYSIS

Computation of visual saliency is based on the notion of a centralized saliency map [1] initiated by a feature competition scheme. 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 while spatiotemporal orientations are computed using steerable filters and measuring their strength along particular directions. The different orientations are then fused to produce a single orientation volume. More details can be found in [3]. Volumes for each feature, are decomposed into multiple scales. The pyramidal decomposition allows the model to represent smaller and larger events in separate subdivisions of the channels.

A spatiotemporal saliency volume is computed with the incorporation of feature competition by defining cliques at the voxel level and using an optimization procedure with both inter- and intra- feature constraints. This is implemented through an energy-based measure that involves voxel operations between coarse and finer scales of the pyramid: 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 $\vec{F}_{\rm v} = [F_{\rm v_1}, F_{\rm v_2}, F_{\rm v_3}]$ with $F^0_{\rm v_k}$ corresponding to the original volume of each, level ℓ of the pyramid is obtained by convolution with an isotropic 3D Gaussian G and dyadic down-sampling $F^\ell_{\rm v_k} = \left(G * F^{\ell-1}_{\rm v_k}\right) \downarrow_2, \quad \ell = 1, 2, ..., p$, where $\downarrow_2$ denotes decimation by 2 in each dimension. For each voxel q of volume F the energy is defined as

$$E_v(F^c_{\rm v_k}(q)) = \lambda_1 \cdot E_1(F^c_{\rm v_k}(q)) + \lambda_2 \cdot E_2(F^c_{\rm v_k}(q)), \qquad (4)$$

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

$$E_1(F^c_{\rm v_k}(q)) = F^c_{\rm v_k}(q) \cdot |F^c_{\rm v_k}(q) - F^h_{\rm v_k}(q)| \qquad (5)$$

and acts as a center-surround operator that promotes areas that differ from their spatiotemporal surroundings, thus attracting attention. The second, *smoothness*, term is defined as

$$E_2(F^c_{\rm v_k}(q)) = F^c_{\rm v_k}(q) \cdot \frac{1}{|N(q)|} \cdot \sum_{r \in N(q)} \left(F^c_{\rm v_k}(r) + V(r)\right), \quad (6)$$

where V is the spatiotemporal orientation volume that indicates motion activity in the scene and $N(q)$ is the 26- neighborhood of voxel q. This 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 volumes $F^1_{\rm v_k}$ at the first pyramid level, i.e., $S(q) = \frac{1}{3} \cdot \sum_{k=1}^{3} F^1_{\rm v_k}(q)$.

Visual saliency values for each frame are obtained by first normalizing the feature volumes in $[0, 1]$ and then pointwise multiplying them by the saliency volume $S(q)$ in order to suppress low saliency voxels. The weighted average is taken to produce a value per frame

$$S_{\rm v} = \sum_{k=1}^{3} w_k \sum_q S(q) \cdot F^1_{\rm v_k}(q), \qquad (7)$$

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

4. TEXT ANALYSIS AND PROCESSING

Subtitles include transcripts of the audio track of videos, as well as, time stamps roughly corresponding to the location of the transcript in the audio stream. They are available in all commercially released video material like movies and TV series. Subtitles are provided in standardized format consisting of movie transcripts followed by time stamps, thus making their processing simple.

4.1. Syntactic Text Tagging

The first step of transcript processing consists of shallow syntactic tagging that includes part-of-speech (POS) tagging. For this purpose, we use a decision-tree-based probabilistic part-of-speech tagger described in [14]. After each word is classified into the corresponding part-of-speech, text saliency weights are assigned to each word based on the POS tag assigned to that word. As far as semantic information is concerned, some POS are more important than others. For example, the most salient POS tags are proper nouns, followed by nouns, noun phrases and adjectives [15]. Verbs can specify semantic restrictions on their pre-arguments and post-arguments which usually belong to the aforementioned classes. Finally, there is a list of words that have very little semantic content; such words are often referred to as "stop-words" and are filtered out in natural language processing and web applications. Next, we assign a saliency measure to groups of POS tags based on the framework outlined above.

POS taggers contain anywhere from 30 to 100 different tags. We have created six POS classes in order to simplify the text saliency computation. The first (and most salient) class contains the proper nouns, e.g., names of movie heroes, cities. The second contains common nouns, the third contains noun phrases, the fourth adjectives, the fifth verbs and the sixth class the remaining parts of speech, e.g., pronouns, prepositions, conjunctions, adverbs. The following weights are assigned[1] to each of the six classes: 1.0, 0.7, 0.5, 0.5, 0.5, 0.2. Note that scores are normalized between 0.2 and 1, i.e., even "stop-words" are assigned a small weight. All in all, each word is assigned a saliency score based on the POS category assigned to it by the tagger.

Next we show the output of the POS tagger and the assigned weights for two sentences from the movie "Lord of the Rings I". Note how proper nouns (PN), e.g. names, are very salient and are assigned a score of 1, noun phrases (NP) and verbs (VBZ, VVG) a score of 0.5, while "stop-words" (IN) are assigned a score of 0.2.

[1] One could actually train these saliency scores based on hand-labeled saliency scores assigned to movie dialogues by users. Here we have chosen a somewhat arbitrary assignment of POS tag classes to saliency scores based on observations of linguistic experts in [15].

Taken	*by*	*Isildur*	*from*	*the*	*hand*	*of*	*Sauron*
NP	NP	PN	IN	NP	NP	IN	PN
0.5	0.5	1.0	0.2	0.5	0.5	0.2	1.0

Evil	*is*	*stirring*	*in*	*Mordor*
NP	VBZ	VVG	IN	PN
0.5	0.5	0.5	0.2	1.0

4.2. Audio Segmentation using Forced Alignment

Although it is generally accepted that the movie subtitles provided by the production company are well synchronized with the audio stream, it is obvious that sometimes there is a delay in the time that subtitle appears. To correct such bias, we perform forced segmentation of the audio stream using the speech transcript and phone-based acoustic models, i.e., an automatic speech recognition (ASR) system. The original timestamps in the subtitles are used to find the approximate location of the text in the audio stream, i.e., to initialize the forced segmentation procedure. We avoid losing relevant speech segments in the audio stream by adding a small fixed amount of time before the start time and after the end time of the subtitle timestamps.

Forced segmentation is achieved using the SONIC ASR toolkit [16]. The acoustic models used were content-dependent tri-phone hidden Markov models trained on clean speech. The grammar used is based on the phonetic transcription of the corresponding text in the subtitles with garbage models in the beginning and end of each sentence. Informal evaluation of the forced segmentation results showed good performance on approximately 85% of the sentences analyzed. Errors occurred for part of the audio stream where speech overlapped with loud music or noises.

4.3. Text Saliency Curve

Based on the assignment of frames to words from the forced segmentation procedure and the word saliency scores assigned by the POS tagger a frame-based text saliency curve is computed as

$$S_{\mathrm{t}}[m] = w_p \chi_p[m],$$
$$p \in \{1, \ldots, 6\},\ w_p \in \{0.2, 0.5, 0.7, 1\},\ \chi_p[m] \in \{0, 1\}. \quad (8)$$

5. MULTIMODAL SALIENCY: AUDIO, VISUAL, TEXT

In a video stream with aural, visual and text information available, attention is modeled by constructing a composite, temporal index of saliency. In this final step the various cues are combined in a multimodal saliency curve (AVT) by low-level fusion $S_{\mathrm{avt}}[m] = \mathrm{fusion}(S_{\mathrm{a}}, S_{\mathrm{v}}, S_{\mathrm{t}}, m)$. A linear or nonlinear fusion scheme can be employed. In this paper, a weighted linear combination of the audio, visual and text saliency is used

$$S_{\mathrm{avt}} = w_{\mathrm{a}} S_{\mathrm{a}} + w_{\mathrm{v}} S_{\mathrm{v}} + w_{\mathrm{t}} S_{\mathrm{t}}, \quad (9)$$

where the three curves are weighted equally.This coupled curve serves as a continuous-valued indicator function of salient events, in one or more of the aural, visual or textual domains.

6. VIDEO SUMMARIZATION

The segment selection and skim rendering algorithm [5], based on the multimodal saliency curve follows the steps:

1. AVT is filtered with a median filter of length $2M + 1$ frames.

2. A saliency threshold S_c is selected so that the required *percent of summarization* c is achieved. Frames n with AVT value $S_{\mathrm{avt}}(n) > S_c$ are selected. 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 is a video frame indicator function I_c for the desired level of summarization c.

3. The selected frames are joined into segments. Selected segments that are shorter than N frames 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. 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 audio samples. 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.

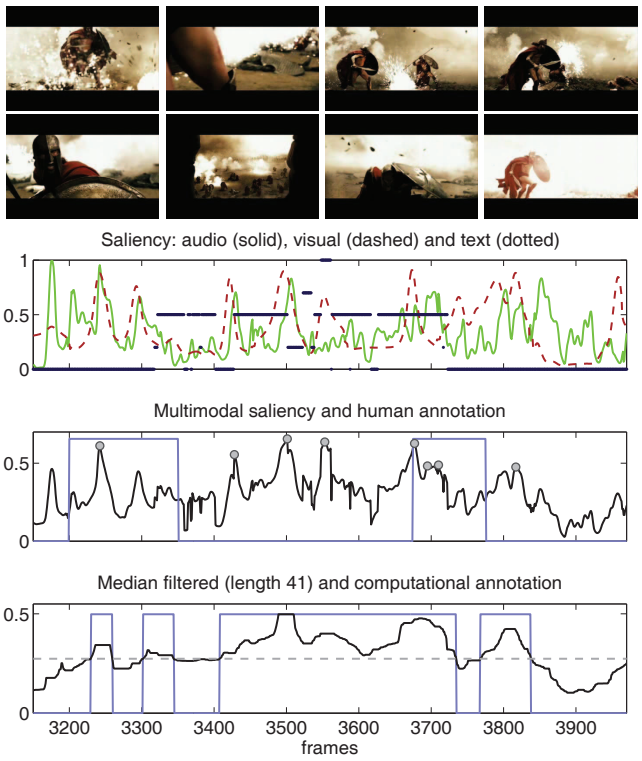

Fig. 1. Saliency curves, annotation and keyframe selection. Human annotation performed by inspection (middle) and automatic by the summarization algorithm (bottom). Keyframes displayed are noted by circles on the AVT curve ("300" movie).

7. EVALUATION & DISCUSSION

The developed multimodal saliency-based method was applied for video summarization on three segments (ca. 5-7 min) from the movies "Lord of the Rings I" (LOTR1), "300" and "Cold Mountain" (CM) [5]. In Fig. 1, we give an example of the saliency curves and their fusion, comparisons of manual vs. automatic saliency annotation and video abstraction (keyframes and salient segments) for 800

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

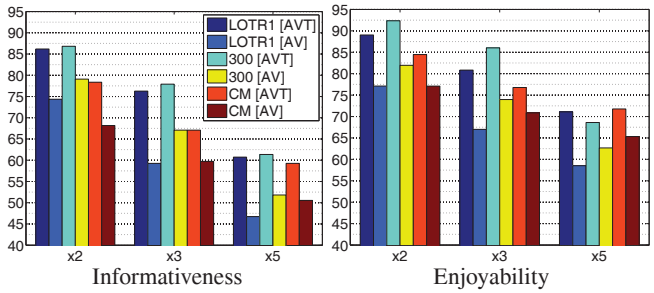

Fig. 2. Subjective evaluation scores of video skims in different rates ($\times 2, 3, 5$), using audiovisual (AV) and multimodal (AVT) saliency.

frames of a scene. Note the agreement of strong peaks in the audio and visual curves, a result of intentionally added sound and visual effects in the scene to draw viewer attention. The salient regions were estimated computationally as in the summarization algorithm, by thresholding the median-filtered AVT, using a 41-length filter. Overall correct salient frame classification exceeded an average of 80% in the three videos for that filter length.

Summaries obtained for $c = 0.5, 0.33, 0.2$, i.e., skimming 2, 3 and 5 times faster than real time, were subjectively evaluated by eleven naive subjects in terms of informativeness and enjoyability, employed in [2, 5], in a 0-100 scale. Each viewer rated both the multimodal and the audiovisual-only (AV) skims and each score was normalized by the score given to the original clip. These preliminary though indicative results, shown in Fig. 2 and Table 1, demonstrate the capability to generate meaningful skims from generic movie content. The high values of "300" may be attributed to the fact that it is an action film with sharp scene changes, soundtrack and crisp narration. The lower performance of "CM" can be explained by the limited use of audiovisual effects, the absence of clear dialog and the transcribed dialect and expressions not recognized by the text model. An interesting observation is that all movies in $\times 5$ have very similar ratings, i.e., around 60% informativeness and 70% enjoyability.

Compared to evaluation of the audiovisual-only framework [5], informativeness improved relatively by 18.2% on the average and enjoyability by 13.8%. We present movie- and skim- depended comparisons in Fig. 2. Changes in skim quality, aesthetics and captured video events that viewers found more essential were most obvious on the $\times 2$ skims. The quality of the resulting audio and video streams, i.e, smoothness, flow, perceptibility, transition etc., rated from very satisfying to acceptable as the skim rate increased.

8. CONCLUSIONS

A multimodal saliency curve integrating the aural, visual and text streams of videos was formed based on efficient audio, image and language processing and employed as a metric for video event detection and abstraction. The proposed video summarization algorithm is based on the fusion of the three streams and the detection of salient video segments. The algorithm is generic and independent of the video semantics, syntax, structure or genre. Subjective evaluation showed that informative and pleasing video skims can be obtained using such multimodal 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 summarizer. Extensions of this work will include more sophisticated fusion algorithms, both inside and among the various modalities, e.g., learning schemes, non-linear feature correlations and variance-adaptive stream weights, the incorporation of extra higher-level features to movie transcript information, and a systematic framework for more thorough evaluations of the summarization algorithm. Sample video skims and on-going evaluations can be found at `http://cvsp.cs.ntua.gr/research`.

Video	AVT scores and relative improvement to AV		
	x2	x3	x5
	Informativeness		
LOTR1	86.2 (+15.9%)	76.3 (+28.7%)	60.7 (+30.0%)
300	86.8 (+ 9.8%)	77.9 (+16.1%)	61.4 (+18.4%)
CM	78.4 (+14.9%)	67.1 (+12.3%)	59.3 (+17.3%)
	Enjoyability		
LOTR1	89.0 (+15.5%)	80.8 (+20.7%)	71.1 (+21.5%)
300	92.4 (+12.7%)	86.0 (+16.3%)	68.6 (+ 9.5%)
CM	84.5 (+ 9.6%)	76.8 (+08.4%)	71.8 (+ 9.9%)

Table 1. Subjective evaluation scores (%) on 0-100 scale.

9. REFERENCES

[1] 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.

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

[3] K. Rapantzikos, N. Tsapatsoulis, Y. Avrithis, and S. Kollias, "Bottom-up spatiotemporal visual attention model for video analysis," *IET Image Processing*, vol. 1, no. 2, pp. 237–248, 2007.

[4] 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.

[5] G. Evangelopoulos, K. Rapantzikos, A. Potamianos, P. Maragos, A. Zlatintsi, and Y. Avrithis, "Movie summarization based on audiovisual saliency detection," in *Proc. IEEE Int'l Conf. Image Processing (ICIP)*, San Diego, CA, 2008.

[6] L. Ying, S.-H. Lee, C.-H. Yeh, and C.-C.J. Kuo, "Techniques for movie content analysis and skimming," *IEEE Signal Processing Mag.*, vol. 23, no. 2, pp. 79–89, Mar 2006.

[7] 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.

[8] 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.

[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] M.A. Smith and T. Kanade, "Video skimming and characterization through the combination of image and language understanding techniques," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 1997, p. 775.

[11] 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.

[12] E.M. Voorhees, "Overview of TREC 2002," in *Proc. of the Eleventh Text REtrieval Conf. (TREC-11)*, E.M. Voorhees and D. Harman, Eds. Springer, 2003.

[13] 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.

[14] H. Schmid, "Probabilistic part of speech tagging using decision trees," in *Proc. Intl. Conf. New Methods in Language Processing*, 1994.

[15] D. Jurafsky and J.H. Martin, *Speech and Language Processing*, Prentice Hall, 2nd edition, 2008.

[16] B. Pellom and K. Hacioglu, "Sonic: The University of Colorado continuous speech recognizer," Tech. Rep. TR-CSLR-2001-01, University of Colorado, Boulder, 2001.