

# Multimodal Saliency and Fusion for Movie Summarization based on Aural, Visual, and Textual Attention

Georgios Evangelopoulos, *Member, IEEE*, Athanasia Zlatintsi, *Student Member, IEEE*, Alexandros Potamianos, *Senior Member, IEEE*, Petros Maragos, *Fellow, IEEE*, Konstantinos Rapantzikos, *Member, IEEE*, Georgios Skoumas, *Student Member, IEEE*, and Yannis Avrithis, *Member, IEEE*

**Abstract**—Multimodal streams of sensory information are naturally parsed and integrated by humans using signal-level feature extraction and higher-level cognitive processes. Detection of attention-invoking audiovisual segments is formulated in this work on the basis of saliency models for the audio, visual and textual information conveyed in a video stream. Aural or auditory 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 orientation. Textual or linguistic saliency is extracted from part-of-speech tagging on the subtitles information available with most movie distributions. The individual saliency streams, obtained from modality-dependent cues, are integrated in a multimodal saliency curve, modeling the time-varying perceptual importance of the composite video stream and signifying prevailing sensory events. The multimodal saliency representation forms the basis of a generic, bottom-up video summarization algorithm. Different fusion schemes are evaluated on a movie database of multimodal saliency annotations with comparative results provided across modalities. The produced summaries, based on low-level features and content-independent fusion and selection, are of subjectively high aesthetic and informative quality.

**Index Terms**—Attention, audio saliency, fusion, movie summarization, multimodal saliency, multistream processing, text saliency, video summarization, visual saliency.

Manuscript received November 16, 2011; revised July 25, 2012, October 22, 2012; accepted January 07, 2013. Date of publication nulldate; date of current version nulldate. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Christophe De Vleeschouwer. Copyright © 2012 IEEE. Personal use of this material is permitted. However, permission to use this material for any other purposes must be obtained from the IEEE by sending a request to [pubpermissions@ieee.org](mailto:pubpermissions@ieee.org).

G. Evangelopoulos\*, A. Zlatintsi, P. Maragos, K. Rapantzikos, G. Skoumas and Y. Avrithis are with the School of Electrical and Computer Engineering, National Technical University of Athens, Athens GR-15773, Greece (e-mail: [gvegav, nzlat, maragos]@cs.ntua.gr, [rap,iavr]@image.ntua.gr, gskoumas@dblab.ece.ntua.gr). A. Potamianos is with the Department of Electronics and Computer Engineering, Technical University of Crete, Chania GR-73100, Greece (e-mail: potam@telecom.tuc.gr).

This research was partially supported by: (1) the project “COGNIMUSE” which is implemented under the “ARISTELA” Action of the Operational Program “Education and Lifelong Learning” and is co-funded by the European Social Fund (ESF) and National Resources; (2) the European Union (ESF) and Greek national funds through the Operational Program “Education and Lifelong Learning” of the National Strategic Reference Framework - Research Funding Program: Heracleitus II; (3) the EU project DIRHA with grant FP7-ICT-2011-7-288121.

Color versions of one or more of the figures in this paper are available online at <http://ieeexplore.ieee.org>.

Digital Object Identifier 10.1109/TMM.2013.XXXXXXXX

## I. INTRODUCTION

ATTENTIONAL selection and information abstraction are cognitive mechanisms employed by humans and animals for parsing, structuring and organizing perceptual stimuli. These mechanisms are grounded in most of the normal conscious or non-conscious activities such as guided search, communication and interaction, awareness and alert states, sensory grouping, memory, decision making, action taking, visual and auditory scene analysis etc. [1]–[3]. They are functionally correlated to modulations of neuronal activity and the product of a combination of bottom-up (sensory) and top-down (cognitive) processing. Attention is the process of focusing cognitive resources on prevailing properties, cues, temporal segments or individual streams of sensory information. Abstraction refers to the reduction of information representations through simplification and selection. Both processes have been the common ground and subject of neurophysiological, cognitive, behavioral and computational studies. In this work, we propose computational models for multimodal stream abstraction and attentional selection, based on the saliency of individual features for aural, visual and linguistic representations.

Attention may be of two modes, a top-down, task-driven and a bottom-up, stimulus-driven, that control the gating of the processed information (input filtering) and the selective access to neural mechanisms (capacity limitation), for example working memory [2]–[5]. Bottom-up attention or saliency is based on the sensory cues of a stimulus captured by its signal-level properties, like spatial, temporal and spectral contrast, complexity, scale etc. [6]–[8]. Similar to competitive selection, saliency can be attributed on the feature level, the stream level or the modality level. For example, a frequency tone may be acoustically salient, a voice can be perceivable among environmental sounds, and an audiovisual scene can be biased towards any of the two signals. Feature saliency is the property of a feature to dominate the signal representation while preserving information about the stimulus. Stream saliency is the property of a temporal segment to stand-out or ‘pop-out’ with respect to its surroundings in a time-evolving scene. Modality saliency is the importance of individual sensory or data modalities (aural, visual, linguistic, etc.) across time (intramodality) or perceptual scene (cross- or intermodality). Salient feature selection is done based either on their representational strength or their

appropriateness for a given application. Here, we develop a hierarchical, multimodal saliency representation where audio, visual and text cues compete at feature, stream and modality levels for the formation of salient multimodal events.

Events in video streams are temporal segments of traceable activity or change in scene properties, for example actions, periodic motions, highlights, or facial expressions. Neurophysiological studies with narrative media like movies suggest that event segmentation is automatic during active perception and occurs both at feature-level and at higher-level concept changes [9]. For a video stream with audiovisual content, perceptually salient events can be detected by tracking intramodal changes on the individual sources of auditory and visual stimuli, or crossmodal changes for coupled events, i.e., changes that span multiple modalities. Attention towards such salient events is triggered by changes or contrast in object appearance (texture and shape), motion activity and scene properties (*visual* events), changes in audio sources, textures or tempo (*aural* events), and the relevant –when available– transcribed dialogues or spoken narrations (*textual* events).

Computational models of single- and multimodal saliency [10]–[13] have been applied to emerging multimedia applications such as automatic video abstraction, summarization, indexing and browsing [14], [15]. Summarization refers to producing a shorter, in duration, version of a video that contains essential information for content understanding, without sacrificing much of the original’s informative, functional or aesthetical purpose. A summary can function as a video preview or an overview, thus aiding in quickly accessing whether the content is important, interesting or enjoyable. Automatic summarization can be broadly classified into two types [14]–[16]: *key-frame selection*, yielding a static, small set of important video frames, and *video skimming* (loosely referred to as *video summarization* here), giving a dynamic short clip that contains sub-segments of the original stream. Since content coverage in the final summaries is important, most summarization techniques employ prior information on the structure of the underlying source data. In the case of scripted, structured video, for example films, the hierarchical segregation in shots and scenes is explicitly utilized [11], [17].

Movies provide the substrate of video content that poses challenging research problems, and at the same time introduce the potential for a range of commercial and interdisciplinary applications. Movie data are multimodal, conveying audio, visual and text information in the form of screenplay and subtitles, scripted and structured and generated through professional and artistic filming and editing. Besides the sensory-level, movies are rich in semantics, either in the form of conceptual units (themes, concepts, stories) or in the form of structured content (frames, shots, scenes). Additionally, direction and montage effects are introduced on purpose in order to induce emotional or attentional responses to the viewer. In this work, we aim to elicit higher-level semantic or affective content from sensory-level saliency representations.

**Contributions and Overview:** We propose multimodal saliency representations of audiovisual streams, in which signal (audio and visual) and semantic (linguistic/textual) cues are integrated hierarchically. Each modality is independently ana-

lyzed in individual saliency representations: spectro-temporal for the audio channel (Sec. III), spatio-temporal for the visual channel (Sec. IV), and syntactic for the transcribed subtitle text (Sec. V). A multimodal saliency score per video frame is obtained by combining features within each modality and saliencies across modalities using linear and nonlinear fusion schemes (Sec. VI), and weighted integration where the weights can be: a) constant across time, b) stream-variance depended, associated to feature/modality uncertainty, and c) dynamically adaptive, in local, structure-depended windows (e.g., movie scene and shot boundaries). Based on the conjecture that temporal variation of saliency is correlated to the time-varying attentional capacity of the underlying streams, a video summarization algorithm is formulated on the basis of salient segments (Sec. VII) and applied on structured, multimodal movie data. The algorithm is content-independent and scalable, ranging from short movie clips to entire movies, and can be generalized to other types of audiovisual data. Summarization precision results, as a function of skim duration and fusion scheme, are presented on a new database of Academy-Awarded films, annotated with respect to salient, semantic and aesthetic content (Sec. VIII). The quality of the produced summaries is additionally evaluated using subjective user ratings related to content aesthetics (enjoyability) and coverage (informativeness) (Sec. IX).

## II. BACKGROUND/RELATED WORK

**Video Summarization:** Summaries of video data may be static or dynamic; personalized, domain-dependent or generic; interactive or unsupervised; based on objects, events or perceptual features, such as user attention. An extensive survey on methods and taxonomies from the vast literature on video abstraction and summarization can be found in [18] and detailed field overviews in [15], [16]. Early works on automatic skimming were primarily based on extracting low-level, visual features, such as color or motion [19], often complemented by mid-level cues. Representative key-frames were selected using supervised and unsupervised classification [20], [21], singular value decomposition [22] or probabilistic inference [23]. Summarization has been also approached in a semi-automatic manner, following video-editing principles, through semantics mining, manual editing and abstraction effects [24].

Attempts to incorporate multimodal and/or perceptual features have led to the design and implementation of various systems that take into account more than the visual stream of a video [25]. IBM’s CueVideo system [26] automatically extracts a number of low- and mid-level visual and audio features and clusters the visually similar shots. The Informedia project [27] and its offsprings combined speech and image processing with natural language understanding to automatically index video for intelligent search and retrieval [28]. Gaining insight from viewer behavior, user attention models were developed to guide the search for salient video segments [12]. Besides visual saliency, additional cues (motion, face, camera and audio attention) have been sought in order to capture salient information, detect important video segments and compose a summary [10]. Attention values were also coupled with scene

and shot detection, for example through partitions on graphs modeling video structure [11].

**Visual attention and image saliency:** Visual saliency and its computational representations, i.e., image saliency maps, have been the subject of efforts to model computationally the neural basis of biological attention [6], [7]. Bottom-up models, in which image regions are selected based on their distinctive physical feature properties, or saliency, have arguably been the most popular. This is supported by observations that saliency is related to a region’s visual conspicuity and can predict gaze and fixation allocation [29]. In most cases, the objective is the notion of a centralized spatial saliency map, where each image location is assigned a stimulus conspicuity or saliency value [7]. In an evaluation study of image saliency methods [30], a soft distinction is drawn in biologically-inspired, purely computational and statistical/information-theoretical approaches.

Cues for visual saliency have been sought in low-level (intensity, color, texture, motion orientation, size distributions) and mid-level features (edges, shapes and contours) and in some approaches high-level object or scene parsing [7], [31], [32]. An in-depth review on the elementary visual cues that can infuse saliency and guide the visual search is given in [33]. Besides spatial contrast, a number of methods rely on frequency or phase-selective tuning of the saliency map [34], [35]. In addition, several information-theoretic measures of saliency have been based on the distributions of features within and across local image patches and neighborhoods. Such measures include entropy [36], self-information [37], mutual-information [38] and spatial Bayesian surprise [39].

The equivalent of a saliency map for image sequences is a spatiotemporal map [35]. To capture the dynamics of visual scenes, a video sequence is represented as a solid in the 3D space. Saliency volumes are then computed through feature competition at voxel level and optimization with inter- and intra-feature constraints [40], [41].

**Aural attention and audio saliency:** The equivalent bottom-up component of auditory attention is due to temporal and spectral cues of the acoustical stimuli [3], [42], [43]. These are related to primitive sound features such as loudness, frequency, direction and their temporal or spatial contrast in an acoustic scene [5], [43], which are also involved in higher-level processing, e.g. parsing and recognition, of the acoustical stream. For example, auditory speech separation employs timbre, pitch and spatial location [43], while speaker identification relies also on accent and intonation [3].

Following the distinction in [3], attention to acoustical streams may be spatial, towards different sources that need to be localized [42], [44] or featural, non-spatial, towards distinctive acoustical features within each stream [10], [45], depending on the demands of the auditory task. The former implies the construction of spatial saliency maps, the latter the formation of temporal saliency streams. In addition, featural salient segments signify the temporal boundaries of audio events [46]. Models of auditory attention have been previously used to simulate psychophysical tasks and applied to audio analysis systems, such as grouping and stream segregation [42], sound source localization, auditory scene analysis [8], soundscape design [44], prominent syllable and word detection

[45], change or event detection and video summarization.

Building on the analogies of early visual and auditory processing, bottom-up, auditory saliency maps of an acoustic scene were developed inspired by the visual paradigm [8]. The auditory spectral representation is processed as an image, by extracting multiscale features (intensity, frequency and temporal contrast, orientation). Pyramidal decompositions in isotropic and oriented bands, give rise to feature maps that are subsequently combined across-scales to an integrated saliency map [44], [45]. Saliency maps can be reduced to a saliency stream by across-frequency integration or maximization. In addition, top-down, biased selection has been included in auditory attention models in the form of task-dependent control mechanisms [45] or attentional switching [44].

For the case of speech signals, salient features have been sought to micro-modulations in the envelope and phase variations of fundamental, nonstationary AM-FM components. These variations were employed for extracting various modulation-inspired representations like formant tracks and bandwidth [47], dominant components [48] and coefficients of energy-frequency distributions [49].

**Text saliency:** The saliency of language and text has been studied extensively for a wide range of applications most notably text summarization. The various features proposed for estimating saliency include: word frequency, term frequency-inverse document frequency (tf-idf) [50], part-of-speech tags [51], discourse structure [52], ontological relationships [53], lexical centrality in semantic graphs [54]. In recent, closed-loop approaches, models of saliency are trained from annotated text corpora using machine learning algorithms [55]–[57].

### III. AUDIO ANALYSIS

We approach saliency computation in an audio stream as a problem of assigning a measure of interest to audio frames, based on spectro-temporal cues. Applying the Amplitude Modulation - Frequency Modulation (AM-FM) speech model [58] to generic audio signals, where multiple sources are added linearly (temporal overlap) or concatenated (auditory streaming) [59], audio features are extracted through signal instantaneous amplitude and frequency. The importance of amplitude and frequency changes for aural saliency and auditory scene analysis has motivated a variety of studies where subject responses are measured with respect to tones of modulated frequency or loudness [3], [5], [8]. Amplitude and frequency modulations are also related to temporal acoustic micro-properties of sounds that are useful for auditory grouping [59] and recognition of audio sources and events. In the proposed model, saliency is quantified through the parameters of elementary AM-FM components, separated in time and frequency. An abstracted representation is obtained by tracking the components with maximal energy contribution across frequencies and time.

The input audio is processed as a sequence of signal frames, with the window length defining the scale of the representation, e.g. instantaneous (sample-wise) or set to match the associated visual modality rate (frame-wise). Frames are decomposed to a set of frequency bands; each band is modeled

by an AM-FM signal and further decomposed to instantaneous energy, amplitude and frequency signals. We are modeling salient structure in the signal level, as opposed to the context level in previous approaches [10], thus approaching saliency in a more generic, context-independent way.

### A. Audio Modeling

In this work, an audio source is either the physical signal-producing medium or a specific semantic part of the audio stream (speech, music, sound effects, natural and artificial sounds, background/environmental noise). A signal from such a source is modeled by a linear mixture of  $K$  nonstationary sinusoids modulated in frequency and amplitude, i.e. a *multi-component AM-FM* signal

$$s(t) = \sum_{k=1}^K s_k(t) = \sum_{k=1}^K a_k(t) \cos(\phi_k(t)) \quad (1)$$

where  $a_k$  and  $\phi_k$  are the instantaneous amplitude and phase signals of component  $k \in [1, K]$ . In the most general case we assume a fixed tessellation in  $K$  sub-components. Each signal  $s_k(t)$  is a real-valued amplitude and frequency modulated sinusoid of the form  $x(t) = a(t) \cos(\phi(t))$  with time-varying envelope  $a(t)$  and instantaneous frequency  $\omega(t) = d\phi(t)/dt$ . Amplitude accounts for subtle envelope variations in  $x(t)$ , while the frequency  $\omega(t) = d\phi(t)/dt$  accounts for small-scale instantaneous deviations from a carrier frequency  $\omega_c$ .

For a complex-valued sinusoid  $x(t) = a(t) \exp(j\phi(t))$ , demodulation in amplitude and frequency can be approximated via  $|a(t)| \approx |x(t)|$  and  $\phi(t) \approx \arctan(\text{Im}\{x(t)\}/\text{Re}\{x(t)\})$  [47]. For real signals, this involves construction of the analytic signal through the Hilbert transform, which requires a longer time window and results in errors related to approximating the quadrature signal. An approach of comparable modeling error, but reduced complexity and improved temporal resolution is based on the Teager-Kaiser energy operator

$$\Psi[x(t)] \equiv [\dot{x}(t)]^2 - x(t)\ddot{x}(t), \quad \dot{x}(t) = dx(t)/dt, \quad (2)$$

and the energy separation algorithm (ESA) [47], [58], [60]. Applied to  $x(t) = a(t) \cos(\phi(t))$ ,  $\Psi$  gives with negligible approximation error under realistic constraints [58],  $\Psi[x(t)] \approx a^2(t)\omega^2(t)$ , i.e., the instantaneous energy of a source of oscillations of amplitude  $a(t)$  and frequency  $\omega(t)$ . This energy is separated to its amplitude and frequency components by the energy separation algorithm using  $\Psi$  on the signal and its first derivative. Signal  $x(t)$  is thus described by the set of amplitude, frequency and energy signals  $\{a(t), \omega(t), \Psi[x(t)]\}$ .

For a multicomponent AM-FM of the form (1), the model requires separation of  $s(t)$  in a set of  $K$  narrowband signals  $s_k(t)$  for the energy separation constraints to hold. A global and a-priori separation is achieved by bandpass filtering through a linearly-spaced set of frequency-tuned filters. Gabor filters have optimum time-frequency localization and their complex responses come in quadrature pairs

$$g_k(t) = (2\pi\sigma_k^2)^{-1} \exp(-t^2/2\sigma_k^2) \exp(j\omega_k t), \quad k \in [1, K] \quad (3)$$

where  $\sigma_k$  determines the temporal support and filter bandwidth and  $\omega_k$  the central frequency. Assuming that a component instantaneously dominates a filter's response ( $g_k * s(t) \approx s_k(t)$ ),

its model parameters can be estimated by demodulating the output directly [60]. This results in a  $(3 \times K)$ -dimensional time-varying representation of the audio signal

$$s(t) \leftrightarrow \{a_k(t), \omega_k(t), \Psi[g_k * s](t)\}, \quad k \in [1, K]. \quad (4)$$

If we additionally require that a single component dominates locally the signal spectrum, we obtain a representation in terms of the *dominant modulation component*  $s_d(t) = a_d(t) \exp\{j\phi_d(t)\}$ , by maximizing an energy criterion  $E_k(t)$  over the  $K$ -dimensional component space [48]:

$$s_d(t) = s_{i(t)}(t), \quad i(t) = \arg \max_{k \in [1, K]} \{E_k(t)\}, \quad (5)$$

A criterion related to component saliency is the short-term energy operator response of the bandpassed signals, estimated over local windows:

$$E_k(t) = \Psi[g_k * s](t) * G_\sigma(t), \quad (6)$$

where  $G_\sigma(t)$  is a Gaussian window of time scale  $\sigma$ . The dominant component's local energy, amplitude and frequency yield a reduced, 3-dimensional, time-varying signal representation:

$$s(t) \approx s_d(t) \leftrightarrow [a_d(t) * G_\sigma(t), \omega_d(t) * G_\sigma(t), E_d(t)]. \quad (7)$$

### B. Audio Features

A discrete-time audio signal  $s[n] = s(nT)$  is modeled using  $K$  discrete AM-FM subcomponents 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 outputs of  $K$  bandpass filters, using convolution with real Gabor filters, a discrete-time energy operator  $\Psi_d(x[n]) \equiv (x[n])^2 - x[n-1]x[n+1]$  and the associated discrete ESA, at an almost instantaneous time resolution [58].

Representation in terms of the dominant modulation components is obtained by maximizing per analysis frame the discrete operator  $\Psi_d$ , in the  $K$ -dimensional energy space. For each frame  $m$  of length  $N$ , the dominant modulation component is the maximum energy response, averaged over the frame  $\{n : (m-1)N < n \leq mN\}$ :

$$E_d[m] = \max_k \{W[n-mN] * \Psi_d[g_k * s][n]\}, \quad (8)$$

where  $W[n]$  is a moving average filter and  $g_k$  the filter impulse response. The dominant energy filter  $j[m] = \arg \max_k \{W * \Psi_d[g_k * s]\}$  is submitted to demodulation via ESA and the instantaneous signals are averaged over frame duration to derive the dominant amplitude and dominant frequency features

$$A_d[m] = W[n] * |A_j[n]|, \quad \Omega_d[m] = W[n] * \Omega_j[n]. \quad (9)$$

Overall, each analysis frame yields average measurements for the source energy, instant amplitude and frequency from the filter that captures the prominent modulation components. The resulting feature vector

$$\mathbf{F}_a[m] = [F_{a1}, F_{a2}, F_{a3}] [m] = [A_d, \Omega_d, E_d] [m] \quad (10)$$

is a low dimensional descriptor of signal properties related to level of excitation, rate-of-change, frequency content and source energy (Fig. 1).

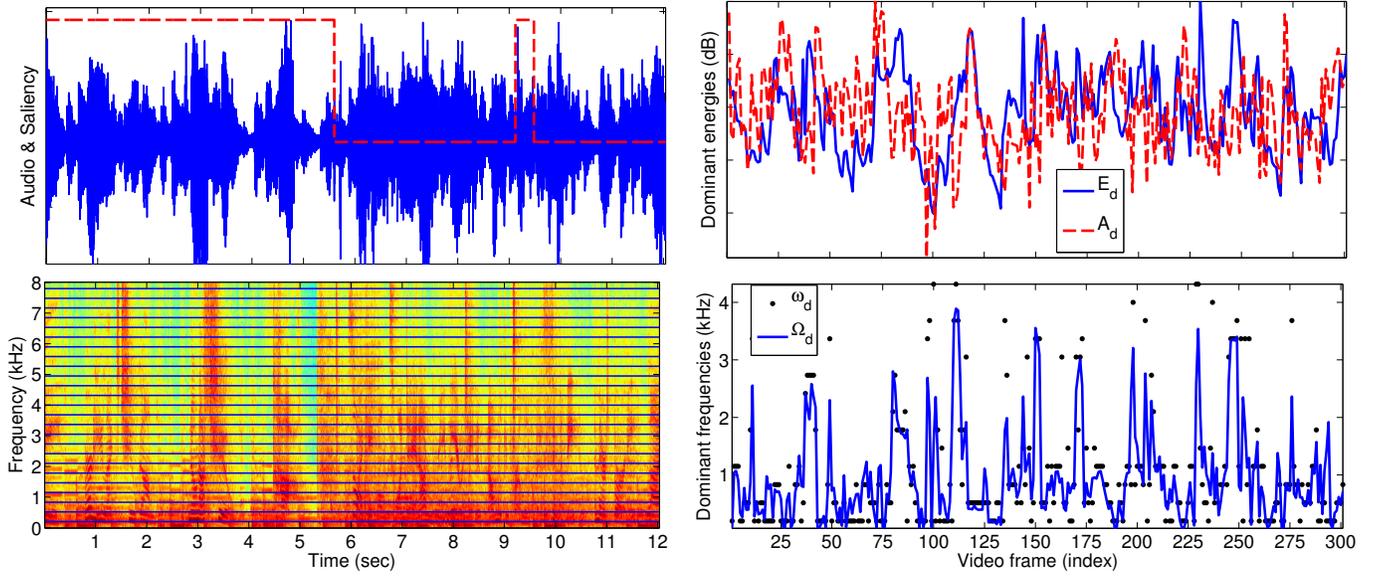


Fig. 1. *Audio stream*: waveform (left-top) with audio saliency annotation and spectrogram (left-bottom) using the employed audio analysis parameters (15 msec windows, 1/2 overlap). Horizontal lines denote the filterbank (25 filters, 400Hz bandwidth) central frequencies, i.e.  $\omega_k$  in Eq. (3). *Dominant modulation features*: energy (solid)  $E_d$  and amplitude (dashed)  $A_d$  (right-top); frequency  $\Omega_d$  (solid) and frequency of the dominant filter (black dots) (right-bottom). Audio data are 300 frames (12 sec) from film “Chicago”, containing music, singing and dialogue.

### C. Audio Saliency

A temporal index of auditory saliency is constructed by integrating the feature values of (10) in a single representation. The rationale is to build a data-driven, time-varying saliency function that resonates with the sensory-level attention invoked to a listener of the audio stream. The features extracted and selected through the audio model quantify spectral and temporal saliency from fundamental modulation structures. Our hypothesis for this modulation-based saliency is to be correlated to the bottom-up attention and listener fixation to parts of an audio signal.

In accordance with the overall bottom-up framework, an early integration scheme is applied. For low-level feature fusion, we define a saliency measure per analysis frame  $S_a[m]: \mathbb{R}^3 \rightarrow [0, 1]$  on the space spanned by the audio feature vectors. An intuitive choice of a weighted linear mapping

$$S_a[m] = w_1 F_{a1}[m] + w_2 F_{a2}[m] + w_3 F_{a3}[m], \quad (11)$$

where  $w_i, i \in \{1, 2, 3\}$  are positive scalars that sum to one, provides a mid-level representation over  $m$  that will depend on feature normalization and the weighting scheme.

Features are normalized with respect to their value range in order to theoretically ensure a mapping to  $[0, 1]$  and compensate for the difference in their dynamic range. Normalization is performed by least squares fitting of independent feature values to  $[0, 1]$  over a long-term window  $L_n$ . The choice of  $L_n$  can be associated with scale, auditory memory and temporal integration of the attentional selection process, in the sense that features retain their relative variation across a finite time extend. For large-durations of inputs, for example the audio channel from a full-length film, the normalizing window can be defined by logical, structural or thematic units. In this context, global normalization can be sub-optimal if a

listener’s attentional thresholds are progressively adapted, e.g., by comparing new representations to short-term memory.

The weighting scheme controls the type and relative contribution of each feature. It be fixed or time-adaptive; incorporate priors on feature significance; or obtained through supervised or semisupervised learning. A baseline of constant equal weights provides the baseline, uniform average of normalized features. Alternatively, assuming independent normal distributions for each, we account for feature uncertainty by setting the weights inversely proportional to feature variance, i.e.,  $w_i = 1/\sigma_i^2$ . This is a theoretically semi-optimal scheme under a weak probabilistic fusion framework [61] and provides the means for adaptivity across time with the variance estimated in local windows  $L_a$  of fixed or varying duration.

The developed audio saliency representation is a continuous-valued function of time, constrained in  $[0, 1]$  by the design of the fusion norm and formed through an unsupervised, bottom-up approach. It constitutes a 1D temporal saliency map, conceptually similar to spatial saliency for images (2D maps) [30] and spatiotemporal saliency maps (3D volumes) for videos [40]. Schemes for fusion, normalization, weighting and adaptation will be further discussed in Sec. VI.

## IV. VISUAL ANALYSIS

We define saliency computation in image sequences as a problem of assigning a measure of interest to each visual unit. This means that a saliency measure is produced by taking into account the actual spatiotemporal evolution of the input. Inspired by theories of grouping and perceptual organization we propose a model based on a volumetric representation of the visual input where features are grouped together according to several criteria related to Gestalt laws.

The input is a sequence of frames represented in our model as a volume in space-time. This volume is decomposed into a set of conspicuity features, each decomposed into multiple scales. Voxel interactions are allowed in three different ways: (a) intra-feature (proximity), between voxels of the same feature and same scale, (b) inter-scale (scale), between voxels of the same feature but different scale and (c) inter-feature (similarity), between voxels of different features. We implement this kind of interactions through global minimization of an energy, which is strongly related to Gestalt's figure/ground separation, since the background is continuously suppressed after each iteration. The stable solution of the energy minimization leads to the final saliency volume. It is important to notice that this formulation allows for selective enhancement of features rather than naive smoothing of conspicuous features. Broadly speaking, the constraints enhance coherency of similar and neighboring voxels according to each of the criteria.

Let  $V$  be a volume representing a set of consequent input frames, defined on a set of points  $Q$ , with  $q = (x, y, t)$  an individual space-time point. Points  $q \in Q$  form a grid in the discrete Euclidean 3D space defined by their coordinates. Under this representation, point  $q$  becomes the equivalent to a voxel in this volume and  $V(q)$  is the value of the volume at  $q$ .  $V$  is decomposed into a set of conspicuity volumes  $C_i$  with  $i = 1, \dots, M$  corresponding to three different features, namely intensity, color and orientation. Each conspicuity volume is further decomposed into multiple scales  $\ell$  and a set  $\mathbf{C} = \{C_{i,\ell}\}$  is created with  $i = 1, \dots, M$  and  $\ell = 0, 1, \dots, L$  representing a Gaussian volume pyramid. The final saliency distribution is obtained by minimizing an energy function  $E$  composed of a data term  $E_d$  and a smoothness term  $E_s$ :

$$E(\mathbf{C}) = \lambda_d \cdot E_d(\mathbf{C}) + \lambda_s \cdot E_s(\mathbf{C}). \quad (12)$$

The data term models the interaction between the observation and the current solution, while the smoothness term is composed of the three constraints.

#### A. Visual Features

In order to establish a common encoding and allow interaction between different features, each of the volumes participating in the energy minimization is initialized by conspicuity and not by pure feature value. Such encoding establishes a common conspicuity range among all features that makes them comparable. This means, for example, that the most conspicuous voxel in the intensity volume must have the same value as the one in the color volume.

*Intensity conspicuity*  $C_1$  is obtained by applying to the intensity, given by  $F_1 = (r + g + b)/3$ , where  $r, g, b$  are the color components of volume  $V$ , a local contrast operator that marks a voxel as more conspicuous when its value differs from the average value in the surrounding region:

$$C_1(q) = \left| F_1(q) - \frac{1}{|N_q|} \sum_{u \in N_q} F_1(u) \right|, \quad (13)$$

where  $q \in Q$  and  $N_q$  is the set of the 26-neighbors of  $q$ . The 26-neighborhood is the direct extension in 3D of the 8-neighborhood in the 2D image space.

*Color conspicuity* is based on the color opponent theory that suggests the control of color perception by two opponent systems: a blue-yellow and a red-green mechanism. Such spatial and chromatic opponency exists for the red/green, green/red, blue/yellow, and yellow/blue color pairs in human primary visual cortex [62] [63]:

$$C_2(q) = (RG + BY)(q) \quad (14)$$

$$RG = |R - G|, \quad BY = |B - Y|,$$

with  $R = r - (g + b)/2$ ,  $G = g - (r + b)/2$ ,  $B = b - (r + g)/2$ , and  $Y = (r + g)/2 - |r - g|/2 - b$ .

Orientation is computed using spatiotemporal steerable filters tuned to respond to moving stimuli. The responses  $E^\theta$  are obtained by convolving the intensity volume  $F_1$  with the second derivatives  $G_2$  of a 3D Gaussian filter and their Hilbert transforms  $H_2$ . The quadrature response is taken to eliminate phase variation. More details are given in [64]. Energies are computed at orientations  $\theta$  defined by the angles related to the three different spatiotemporal axis. In order to get a purer measure, the response of each filter is normalized by the sum of the consort and *orientation conspicuity* is computed by

$$C_3(q) = \frac{\sum_{\theta} E_{\theta}(q)}{\sum_u \sum_{\theta} E_{\theta}(u)}. \quad (15)$$

#### B. Energy Formulation

Each of the conspicuity volumes encodes the saliency of the contained voxels according to the corresponding feature only. These volumes should interact in order to produce a single saliency measure for each voxel. The proposed model achieves this through a regularization framework, whereby conspicuity volumes compete along a number of directions, namely interaction among voxels at the intra-feature, inter-scale and inter-feature level. As discussed above, the different interactions are implemented as a competition modeled by energies inspired by the Gestalt laws. Specifically, proximity and closure laws give rise to the intra-feature constraint, according to which voxels that are located near each other tend to be part of a group and small gaps are closed due to the induced forces. The similarity law is related to all energies, since voxels similar in terms of intra-feature, inter-feature and inter-scale value tend to group. Finally, the common fate law is related to the entire minimization approach which produces space-time regions that can be perceived as coherent and homogenous. Hence, we expect that voxels conspicuous enough to pop out in all dimensions will become ever salient during the minimization process.

The *data term*,  $E_d$ , preserves a relation between the observed and initial estimate in order to avoid excessive smoothness of the result, since the energies involved in  $E_s$  tend to smooth the visual input according to different criteria. The constraint is formulated as an energy relating the observed to the initial voxel values. For a set of conspicuity volumes  $\mathbf{C}$  the data term is defined as

$$E_d(\mathbf{C}) = \sum_i \sum_{\ell} \sum_q (C_{i,\ell}(q) - C_{i,\ell}^0(q))^2, \quad (16)$$

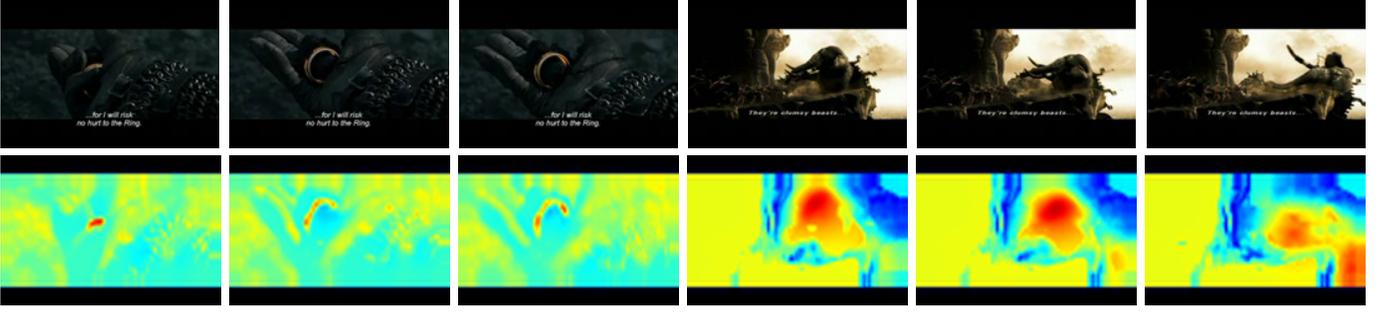


Fig. 2. Sample video frames (top) and corresponding color-encoded saliency maps (bottom) from movies “Lord of the Rings I” (left) and “300” (right).

where  $C_{i,\ell}^0(q)$  is the initial estimate,  $i = 1, \dots, M$ ,  $\ell = 1, \dots, L$  and  $q \in Q$ . The sum limits are omitted for simplicity.

The *smoothness term*,  $E_s$ , is formulated as

$$E_s(\mathbf{C}) = E_1(\mathbf{C}) + E_2(\mathbf{C}) + E_3(\mathbf{C}), \quad (17)$$

where  $E_1, E_2, E_3$  denote the intra-feature, inter-feature and inter-scale constraints respectively.  $E_1$  models *intra-feature* coherency, i.e., defines the interaction among neighboring voxels of the same feature, at the same scale and enhances voxels that are noncoherent with their neighborhood:

$$E_1(\mathbf{C}) = \sum_i \sum_\ell \sum_q \left( C_{i,\ell}(q) - \frac{1}{|N_q|} \sum_{r \in N_q} C_{i,\ell}(r) \right)^2 \quad (18)$$

$E_1$  produces small spatiotemporal blobs of similar valued voxels.  $E_2$  models *inter-feature* coherency, i.e., it enables interaction among different features so that voxels being conspicuous across all feature volumes are grouped together and form coherent regions. It involves competition between a voxel in one feature volume and the corresponding voxels in all other feature volumes:

$$E_2(\mathbf{C}) = \sum_i \sum_\ell \sum_q \left( C_{i,\ell}(q) - \frac{1}{M-1} \sum_{j \neq i} C_{j,\ell}(q) \right)^2 \quad (19)$$

$E_3$  models *inter-scale* coherency among ever coarser resolutions of the input, i.e., aims to enhance voxels that are conspicuous across different pyramid scales:

$$E_3(\mathbf{C}) = \sum_i \sum_\ell \sum_q \left( C_{i,\ell}(q) - \frac{1}{L-1} \sum_{n \neq \ell} C_{i,n}(q) \right)^2 \quad (20)$$

Voxels that retain high values along all scales are more salient. This effect is in conformance also to the scale saliency definition proposed by Kadir and Brady [36].

### C. Energy Minimization

To minimize (12) we adopt a steepest gradient descent algorithm where the value of each feature voxel is updated along a search direction, driving the value in the direction of the estimated energy minimum

$$C_{i,\ell}^\tau(q) = C_{i,\ell}^{\tau-1}(q) + \Delta C_{i,\ell}^{\tau-1}(q), \quad (21)$$

$$\Delta C_{i,\ell}^{\tau-1}(q) = -\gamma \cdot \frac{\partial E(\mathbf{C}^{\tau-1})}{\partial C_{i,\ell}^{\tau-1}(q)} + \mu \cdot \Delta C_{i,\ell}^{\tau-2}(q), \quad (22)$$

where  $\tau$  is the iteration number,  $\gamma$  is the learning rate and  $\mu$  a momentum term that controls the algorithm’s stability. The two parameters are important both for stability and speed of convergence. Practically, few iterations are enough for the estimate to reach a near optimal solution.

Equation (22) requires the computation of the energy partial derivative

$$\frac{\partial E(\mathbf{C})}{\partial C_{k,m}(s)} = \lambda_d \cdot \frac{\partial E_d(\mathbf{C})}{\partial C_{k,m}(s)} + \lambda_s \cdot \sum_{c=1}^3 \frac{\partial E_c(\mathbf{C})}{\partial C_{k,m}(s)} \quad (23)$$

where  $k = 1, \dots, M$ ,  $m = 1, \dots, L$ ,  $s \in Q$  and  $E_c$  with  $c = 1, \dots, 3$  the three energy constraints of the smoothness term. The detailed analytic derivation of the partial derivatives of  $E_d$  and  $E_c$  can be found in [41].

### D. Visual Saliency

The convergence criterion for the minimization process is defined by  $\max_q |\Delta C_{i,\ell}^{\tau-1}(q)| < \epsilon$ , where  $\epsilon$  is a small constant. The output is a set of modified conspicuity multiscale volumes  $\hat{\mathbf{C}} = \{\hat{C}_{i,\ell}\}$  and saliency is computed as the average of all volumes across features and scales:

$$S = \frac{1}{ML} \sum_{i=1}^M \sum_{\ell=1}^L \hat{C}_{i,\ell}. \quad (24)$$

A more detailed description of the method for different applications can be found in [40], [41]. Figure 2 depicts the computed saliency on three frames of movies “Lord of the Rings I” and “300”, where higher values correspond to more salient regions (e.g., the shining ring or the falling elephant).

In order to create a single saliency value per frame, we use the same features involved in the saliency volume computation, namely, intensity, color and spatiotemporal orientation. 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. Each frame is assigned a single saliency value through the the weighted average:

$$S_v = \sum_{k=1}^3 \sum_q S(q) \cdot F_{k,1}(q), \quad (25)$$

where the second sum is taken over voxels  $q$  of saliency volume  $S(q)$  and  $F_{k,1}(q)$  stands for the  $k$ -th feature volume at the first pyramid level.

## V. TEXT ANALYSIS

Text saliency scores are assigned on the basis of empirical word-level relative importance values. As a preprocessing step, spoken language information in the audio stream has to be automatically recognized or manually annotated. In addition, the produced transcripts have to be time-aligned with the audio stream, in order to establish a temporal word-frame correspondence. In this work, we utilize the annotation available in the subtitles of movies and commercial video streams, although the proposed approach can be also applied to the output of an automatic speech recognizer.

### A. Audio Segmentation using Forced Alignment

Although subtitles provided with commercially released video material are roughly time aligned with the audio stream, the synchronization is not perfect. To correct time-stamp bias and achieve accurate word-level alignment, we perform forced segmentation on 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 in order 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 time-stamps.

In this work, we use the Sonic ASR toolkit [65] and general-purpose acoustic models, i.e., 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 portions of the audio stream where speech overlapped with loud music or noises. Audio alignment results can be further improved by employing acoustic model adaptation techniques [66] or acoustic modeling of various noise types [67], [68].

### B. Syntactic Text Tagging

The time-aligned transcripts are analyzed using a shallow syntactic parser that (mainly) performs part-of-speech (POS) tagging. We employ a decision-tree-based probabilistic tagger [69], although in principle any POS tagger can be used. Text saliency scores are assigned to each word based on the POS tag of that word. The motivation behind this approach is the well-known fact that (on-average) some POS convey more information than others. The most salient POS tags are proper nouns, followed by nouns, noun phrases and adjectives [70]. Verbs can specify semantic restrictions on their pre- and post-arguments, which usually belong to the aforementioned classes. Finally, there is a list of words (often referred as stop-words) that have very little semantic content.

POS taggers contain anywhere from 30 to 100 different tags. We have grouped those into six POS classes to simplify the text saliency computation process. The first (and most salient) class contains the proper nouns, e.g., names, 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 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. The somewhat arbitrary assignment of POS tag classes to saliency scores was chosen based on observations of linguistic experts [70], however the weights can be learned from saliency annotations or scores assigned to movie dialogues.

All in all, each word is assigned a saliency score based on the POS category assigned to it by the tagger. For example, the POS label and assigned weights for two sentences from “Lord of the Rings I” are:

|        |     |          |      |         |      |     |         |
|--------|-----|----------|------|---------|------|-----|---------|
| “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” |      |     |         |
| NN     | VBZ | VVG      | IN   | PN      |      |     |         |
| 0.7    | 0.5 | 0.5      | 0.2  | 1.0     |      |     |         |

Note how proper nouns (PN), e.g., “Sauron”, “Mordor”, are very salient and are assigned a score of 1, common nouns (NN) a score of 0.7, noun phrases (NP) and verbs (VBZ, VVG) a score of 0.5, while “stop-words” (IN) are assigned a score of 0.2. The noun phrases (NP) tags produced by the parser are consistent with the (phrase-level) Penn-Tree bank tags. Since NPs contain a common noun and typically one or two words of lower saliency (e.g., determiner, pronoun, adjective) they are assigned a mid-range score of 0.5.

### C. Text Saliency

Based on the assignment of frames to words from the forced segmentation procedure and the word saliency scores assigned by the POS tagger, a text saliency temporal curve  $S_t$  is computed as follows:

$$S_t[m] = \sum_p w_p \chi_p[m], \quad \chi_p[m] \in \{0, 1\},$$

$$p \in \{1, \dots, 6\}, \quad w_p \in \{0.2, 0.5, 0.7, 1\}, \quad (26)$$

where  $m$  is the frame index,  $p$  is the POS class index,  $w_p$  is the saliency score for class  $p$ , and  $\chi_p[m]$  is an indicator function that is 1 if frame  $m$  is aligned with a (portion of a) word in POS class  $p$ , else 0.

## VI. MULTIMODAL FUSION

Fusion of different modalities and their representations can be performed at three levels: a) *low-level* fusion (feature vectors), b) *middle-level* fusion (saliency curves) or c) *high-level* fusion (curve features, salient segments, events). The process of combining feature or saliency curves can in general be linear or nonlinear [10], [71], have memory, or vary with time. In addition, representations from heterogeneous modalities may require preprocessing, normalization and alignment prior to

integration. We consider and evaluate experimentally different fusion schemes, within and across modalities. Specifically, two problems are examined: **Intramodal fusion:** Features for each modality are normalized and combined to produce modality-specific saliency curve (Sec. III, IV, V). **Intermodal fusion:** Saliency curves from different modalities are combined in a composite, multimodal saliency. The discussed methods for saliency fusion, normalization and adaptation have been also applied for the intramodal, feature integration problem.

### A. Saliency Integration

The individual saliency cues are combined in a multimodal, audio-visual-text saliency using frame-level fusion:

$$S_{\text{avt}}[m] = \text{fusion}(S_a, S_v, S_t, m). \quad (27)$$

Variants of fusion norm frameworks are considered and validated experimentally, namely: a) weighted linear combinations with fixed, equal or unequal, weights; b) variance-based weights, inversely proportional to each modality's uncertainty  $\sigma_i^2$ ,  $i \in \{a, v, t\}$ ; c) nonlinear norms, e.g., max, min; d) time-adaptive, dynamic weights, using syntactic video structure (e.g., scene and shot changes). For intramodal (audio and visual) fusion, we used the experimentally optimum of the same variants of baselines (11) and (25).

**Linear Fusion:** The straightforward and most intuitive scheme, sets the information gain of the multimodal curve equal to the sum of the unimodal gains, through a memoryless, weighted average of audio, visual and text saliency values:

$$S_{\text{lin}} = w_a S_a + w_v S_v + w_t S_t. \quad (28)$$

In general, the weights can be unequal, time-varying, adaptive, depending on priors such as the uncertainty of the feature streams etc. Assuming that the individual saliency features are normalized in  $[0, 1]$  and the weights form a convex combination, linear fusion gives a multimodal saliency in  $[0, 1]$ . Our baseline system (LE-F) is based on linear fusion with all three saliency curves equally weighted.

**Variance-based Fusion:** Each saliency stream is weighted inversely proportional to its variance:

$$S_{\text{var}} = \frac{1}{\text{var}(S_a)} S_a + \frac{1}{\text{var}(S_v)} S_v + \frac{1}{\text{var}(S_t)} S_t. \quad (29)$$

The linear scheme (28) is optimum (to a second-order approximation) under the maximum a posteriori (MAP) criterion if the monomodal MAP estimates are close and the weights equal the negatives of the 2nd-order derivatives of the monomodal posteriors at their maxima [61]. If the underlying distributions are Gaussian, then it is also exact and the weights become inversely proportional to the variances. The same variance-based scheme (VA-F) can also be applied to feature combination within each modality for intramodal fusion.

**Nonlinear Fusion:** Two nonlinear fusion schemes are considered for intermodal (resp. intramodal) fusion, namely the minimum (MI-F) and maximum (MA-F) rules applied on saliency (resp. feature) values at each frame:

$$S_{\text{min}} = \min\{S_a, S_v, S_t\}, \quad S_{\text{max}} = \max\{S_a, S_v, S_t\}. \quad (30)$$

**$\alpha$ -Divergence:** All the above linear and nonlinear fusion schemes can be considered as special cases of the so-called  **$\alpha$ -mean** proposed in [72] for integration of probability distributions. Let  $S_k$  be nonnegative quantities, i.e., saliency values; then, for  $\alpha \in [-\infty, +\infty]$ , their  $\alpha$ -mean is defined by

$$f_\alpha^{-1} \left[ \sum_k w_k f_\alpha(S_k) \right], \quad f_\alpha(x) \triangleq \begin{cases} x^{\frac{1-\alpha}{2}}, & \text{for } \alpha \neq 1, \\ \log x, & \text{for } \alpha = 1, \end{cases} \quad (31)$$

where  $w_k$  are given weights and form a convex combination. For  $\alpha = -\infty, -1, 1, 3, +\infty$ , we obtain respectively the maximum, weighted arithmetic mean, geometric mean, harmonic mean and minimum of the values  $S_k$ . In [72] it is shown that the  $\alpha$ -mean is optimum in minimizing the  $\alpha$ -divergence.

### B. Normalization Scale and Weight Adaptation

Movies are structured hierarchically in progressively larger units, of increasing duration (frames, shots, scenes, settings, thematics). A global normalization of feature and saliency values for fusion assumes a mode of prolonged, attentional viewer fixation around the global maxima of the resulting curves. To introduce locality in feature scaling, three linear schemes are considered based on semantically-chosen normalization windows: (a) global normalization (GL-N), (b) scene-based normalization (SC-N) and (iii) shot-based normalization (SH-N), with values scaled independently across movies, scenes and shots respectively. For SC-N (resp. SH-N) normalization, we impose the same peak-to-peak variation for all scenes (shots) in a movie clip. Other schemes such as nonlinear (e.g., log) scaling or root mean square normalization across shots or scenes may be plausible options but are not considered here. In accordance, the same locality is considered for dynamic adaptation of the fusion rule, through weight updating across global or local windows. In the case of the inverse-variance weighting scheme of Eq. (29), the variance of each stream can be computed at a global (VA-GL-F), shot (VA-SH-F) or scene (VA-SC-F) level.

## VII. VIDEO SUMMARIZATION

We present a dynamic summarization algorithm that selects the most salient audio and video sub-clips in order to produce a coherent and informative summary. Clips are selected based on their attentional capacity through the computed multimodal, audio-visual-text (AVT) saliency. One approach for creating summaries is to select, based on a user- or application-defined skimming index, portions of video around key frames and align the corresponding ‘‘audio sentences’’ [10]. Here, summaries are created using a predefined *skimming percentage*.

First, a smoother, coarse attention curve is created using median filtering on the initial AVT saliency, since information from key-frames or saliency boundaries is not necessary. A saliency threshold  $T_c$  is selected so that a required *percent of summarization*  $c$  is achieved. Frames  $m$  with saliency value  $S_{\text{av}}[m] > T_c$  are selected to be included in the summary. For example, for 20% summarization,  $c = 0.2$ , the threshold  $T_c$  is selected so that the cardinality of the set of selected frames  $D = \{m : S_{\text{av}}[m] > T_c\}$  is 20% of the total number of frames.

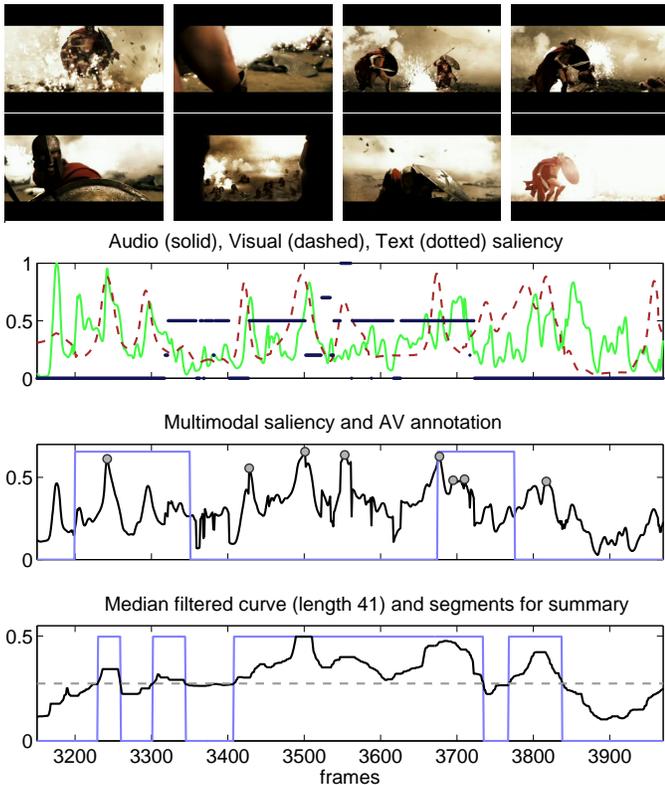


Fig. 3. Saliency curves, multimodal fusion and manual vs. automatic segment selection for movie summarization (800 frames, scene from the movie “300”). Keyframes (top) correspond to the indicated saliency peaks.

The result from this leveling step is a video frame indicator function  $I_c$  for the desired level of summarization  $c$  that equals 1,  $I_c[m] = 1$ , if frame  $m$  is selected for the summary and 0 otherwise. The resulting indicator function  $I_c$  is further processed to form contiguous blocks of video segments. This processing involves eliminating isolated segments of small duration and merging neighboring blocks in one segment. Finally, the selected clips are tailored together using overlap-add (fade-in fade-out) for both the audio and visual streams. More details are provided in [13].

Results presented in this paper use the following parameters, for videos at a rate of 25 frames per second: a 41-frame median filter is used for saliency curve smoothing; selected clips that are shorter than 20 frames long are ignored; selected clips that are at most 10 frames apart are joined together; and fade-in fade-out is applied over 10 frames. In Fig. 3 an example of individual saliency curves and their multimodal fusion (linear with equal weights) is shown for a short clip from movie “300”. Video frames associated with high saliency values (marked with circles on the multimodal curve) are shown on top and segments selected from the median-filtered curve for the summary ( $c = 0.5$ ) are shown at the bottom.

## VIII. DATABASE

The Movie Summarization (MovSum) Database consists of half-hour continuous segments from seven movies (three and a half hours in total), namely: “A Beautiful Mind” (BMI), “Chicago” (CHI), “Crash” (CRA), “The Departed”

(DEP), “Gladiator” (GLA), “Lord of the Rings - the Return of the King” (LOR) and the animation movie “Finding Nemo” (FNE). Oscar-winning movies from various film genres (drama, musical, action, epic, fantasy, animation) were selected to form a *systematic, genre-independent database* of acclaimed, high production quality videos. In this paper we present results for seven movies, however the expansion and annotation of the database is an ongoing task.

Movie clips were first manually segmented into shots and scenes. A shot is defined as the interval between editing transitions (e.g., cut, fade) while a scene, is defined as a complete, continuous chain of actions (shots) that occur at the same place and time. The average shot and scene duration was 2.5 sec and 3.5 min, respectively. Next, labeling of perceptual, semantic, and affective content was performed, as follows: (a) *Sensory information*: monomodal (audio, visual) and multimodal (AV) saliency of the sensory content, i.e., segments that are, respectively, acoustically, visually or audio-visually interesting. (b) *Cognitive information*: the combination of sensory/perceptual events and semantics/pragmatic events. It includes the *semantic information* layer, i.e., segments that are conceptually important as stand-alone sensory/semantic events, henceforth referred to as audio-visual-semantic events (AVS). (c) *Affective information*: both intended emotions and experienced emotions have been annotated. More details on the affective annotation and the associated emotion tracking task are provided in [73].

Annotation was performed by three expert viewers using ANVIL video annotation tool [74]. Movie segments that were considered salient at the audio sensory (A), visual sensory (V), audio-visual sensory (AV), and audio-visual sensory/semantic (AVS) level were labeled in separate annotation runs. The output of each run was a binary saliency indicator function. Table I shows the (average) percentage of frames labeled as A, V, AV, AVS, for each of the seven movies. To achieve a high degree of annotation uniformity for this highly subjective task, the annotators followed guidelines from a labeler’s manual on how to perform the labeling of all individual layers. Table II shows the average (pairwise) correlation agreement, overall satisfactory, for each annotation layer and movie. Note that the agreement is higher for the sensory (A, V, AV) layers

TABLE I  
AVERAGE PERCENTAGE OF FRAMES LABELED SALIENT.

| Layer | BMI  | CHI  | CRA  | DEP  | GLA  | LOR  | FNE  | Mean |
|-------|------|------|------|------|------|------|------|------|
| A     | 25.4 | 56.3 | 55.0 | 33.4 | 60.9 | 58.3 | 54.6 | 49.1 |
| V     | 30.1 | 46.3 | 37.9 | 32.4 | 39.2 | 43.3 | 36.9 | 38.0 |
| AV    | 27.4 | 47.7 | 43.1 | 37.8 | 49.6 | 50.7 | 39.7 | 42.3 |
| AVS   | 63.2 | 76.6 | 64.8 | 71.8 | 68.5 | 72.7 | 67.6 | 69.3 |

TABLE II  
AVERAGE (PAIR-WISE) CORRELATION BETWEEN LABELERS.

| Layer | BMI  | CHI  | CRA  | DEP  | GLA  | LOR  | FNE  | Mean |
|-------|------|------|------|------|------|------|------|------|
| A     | 0.54 | 0.48 | 0.46 | 0.49 | 0.51 | 0.52 | 0.42 | 0.49 |
| V     | 0.31 | 0.33 | 0.32 | 0.45 | 0.38 | 0.43 | 0.38 | 0.37 |
| AV    | 0.45 | 0.45 | 0.41 | 0.54 | 0.44 | 0.50 | 0.44 | 0.46 |
| AVS   | 0.29 | 0.24 | 0.27 | 0.29 | 0.31 | 0.33 | 0.23 | 0.28 |

compared to the sensory-semantic AVS layer. However, the ground-truth saliency indicator functions, used for evaluation purposes in Sec. IX, consist of frames that have been labeled salient by at least two labelers. Thus, despite the lower agreement between annotators observed for certain movies, the final saliency ground-truth was formed on the basis of consistently-labeled salient frames only.

## IX. EVALUATION AND DISCUSSION

In this section, we present objective (quantitative) and subjective (qualitative) results for the proposed saliency representations and video summarization algorithms. The different fusion and normalization schemes for multimodal saliency are systematically evaluated, with respect to frame-level precision, against the annotated ground-truth and compared to learning-based classification. The produced summaries are also qualitatively evaluated in terms of informativeness and enjoyability.

### A. Objective Evaluation of Fusion Schemes

We evaluate three normalization schemes, i.e., global (GL-N), scene-level (SC-N), shot-level (SH-N), and six fusion schemes, i.e., linear (LE-F), min (MI-F), max (MA-F), inverse variance (VA-GL-F, VA-SC-F, VA-SH-F), using the annotations of the movie database as ground-truth. Normalization and fusion schemes are evaluated on three different tasks: a) intramodal fusion of audio features into an audio saliency curve (audio feature fusion), evaluated on audio saliency ground-truth (A-A); b) intramodal fusion of visual features into a visual saliency curve (visual feature fusion), evaluated on visual saliency ground-truth (V-V); c) intermodal fusion of audio, visual and text curves (AVT) into multimodal saliency, evaluated on ground-truth of audio-visual (AVT-AV) and audio-visual-semantic (AVT-AVS) annotations.

Results are presented in terms of frame-level precision scores for all tasks. Specifically, the automatically-computed binary indicator functions on the output of the summarization algorithm (Sec. VII) are compared to the annotated database ground-truth. Precision, i.e., the percentage of correctly detected salient frames given the ground-truth labeling to salient/non-salient frames (precision = correctly detected / all detected), best characterizes the frame-level performance on these salient event detection tasks. Note that for the intramodal experiments (audio feature and visual feature fusion) the saliency indicator and associated summaries are produced using only the corresponding modality features (audio and visual respectively). Various percents of summarization are considered, corresponding to different levels of decision thresholds on the computed saliency curves.

Frame precision scores for intra- and intermodal fusion are presented in Table III for summaries that include 20%, 33% and 50% of the original number of frames, and for the tasks of audio feature fusion, visual feature fusion and audio-visual-text saliency fusion. All possible combinations among normalization and fusion schemes are evaluated. Best results in each task are shown in bold. There are two clear trends across tasks and evaluation settings: a) global normalization (GL-N) significantly outperforms shot-level and scene-level

normalization schemes, and b) nonlinear minimum fusion (MI-F) and inverse variance weighting (VA-GL-F, VA-SC-F, VA-SH-F) outperform uniform linear combination (LE-F) and max fusion (MA-F). For the remainder of this paper, we focus only on the best performing normalization (GL-N) and fusion (MI-F, VA-GL-F, VA-SC-F, VA-SH-F) schemes. Linear fusion (LE-F) results are also reported as baseline.

In Fig. 4, frame precision results are shown as a function of summarization percentage (ranging from 5% to 70%). Results are shown for global normalization and the best performing schemes for the inter- and intramodal fusion tasks (same as in Table III). For audio feature fusion, in Fig. 4(a), MI-F performs best, while LE-F performs significantly worse compared to the rest. All inverse variance schemes perform well, with VA-SH-F being the best (performing close to MI-F). For visual feature fusion, in Fig. 4(b), MI-F significantly outperforms the similarly performing inverse variance and linear fusion schemes. Note the low absolute precision scores achieved for the visual task (compared to audio detection precision). Results for multimodal saliency fusion are shown in Fig. 4(c),(d) on AV and AVS annotation, respectively. Relative scheme performance is consistent in both: MI-F performs best, followed closely by VA-SH-F, while linear fusion performs very poorly. Note that (absolute) precision scores are higher in Fig. 4(d) compared to (c), i.e., including content semantics in the evaluation improves saliency precision. Although detection relies on low-level features, semantic information might be slipping through from text saliency. Overall, the VA-SH-F and MI-F schemes work very well for both intra- and intermodal fusion. Nonlinear min (MI-F) works notably well for single modality fusion (e.g., visual), especially for shorter summaries.

Finally, we examine the relative performance of single (A, V) and multiple modalities (AV, AVT) on a common annotation layer, i.e., audio-visual-semantic (AVS), with frame-level precision results presented in Fig. 6. Clearly, the audio features provide the best stand-alone performance, significantly higher than the visual ones. Fusing the audio-visual-text curves improves average precision, especially for longer summaries, using six out of seven movies (shown here for all seven).

### B. Machine Learning

As a proof-of-concept, we consider a data-driven, machine learning approach to monomodal and multimodal frame selection, where classifiers are trained using the frame-wise audio (A) and visual (V) features, or pooled feature vectors for audiovisual (AV) and audio-visual-text (AVT). The purpose of this baseline is two-fold: a) to validate the efficiency of the proposed saliency features, irrespective of the fusion scheme, b) to serve as a supervised learning-based benchmark for summarization via frame selection.

Specifically, we employ the raw feature vectors of each modality along with their first and second time derivatives computed over three and five frames respectively. Nearest neighbor classifiers (NNR- $k$ )<sup>1</sup> are trained per individual or multiple modalities on the annotated saliency ground-truth (A,

<sup>1</sup>Similar results can be obtained using Gaussian mixture models or Support Vector Machine classifiers.

TABLE III  
 FRAME-LEVEL SUMMARIZATION PRECISION FOR FEATURE (AUDIO, VISUAL) AND SALIENCY (AUDIO-VISUAL-TEXT) FUSION, EVALUATED ON MONOMODAL (A, V), MULTIMODAL (AV) AND MULTIMODAL-SEMANTIC (AVS) GROUND-TRUTH ANNOTATIONS.

| Fusion Level     |         | Audio Feature                              |             |             | Visual Feature        |             |             | AVT Saliency          |             |             | AVT Saliency          |             |             |
|------------------|---------|--|-------------|-------------|-----------------------|-------------|-------------|-----------------------|-------------|-------------|-----------------------|-------------|-------------|
| Annotation Layer |         | Audio (A)                                  |             |             | Visual (V)            |             |             | AudioVisual (AV)      |             |             | AV-Semantic (AVS)     |             |             |
|                  |         | Summarization Percent                      |             |             | Summarization Percent |             |             | Summarization Percent |             |             | Summarization Percent |             |             |
| Algorithm        |         | 20%  | 33%         | 50%         | 20%                   | 33%         | 50%         | 20%                   | 33%         | 50%         | 20%                   | 33%         | 50%         |
| Norm             | Fusion  | Frame-Level Summarization Precision Scores |             |             |                       |             |             |                       |             |             |                       |             |             |
| GL-N             | LE-F    | 62.7                                       | 60.7        | 57.6        | 42.8                  | 39.7        | 37.0        | 40.5                  | 38.7        | 38.5        | 74.8                  | 72.5        | 72.1        |
| GL-N             | MI-F    | <b>89.2</b>                                | <b>79.2</b> | <b>69.2</b> | <b>49.6</b>           | <b>45.5</b> | <b>40.8</b> | <b>79.5</b>           | <b>69.9</b> | <b>58.9</b> | <b>91.2</b>           | <b>87.9</b> | <b>83.2</b> |
| GL-N             | MA-F    | 44.8                                       | 47.4        | 48.6        | 40.0                  | 37.9        | 35.5        | 37.4                  | 36.9        | 38.0        | 69.0                  | 69.2        | 69.2        |
| GL-N             | VA-GL-F | 85.8                                       | 75.6        | 65.7        | 42.7                  | 40.1        | 37.5        | 64.7                  | 56.2        | 49.9        | 83.0                  | 79.1        | 76.1        |
| GL-N             | VA-SC-F | 82.7                                       | 72.1        | 64.2        | 43.1                  | 39.9        | 37.6        | 66.0                  | 59.8        | 53.8        | 84.6                  | 81.4        | 78.0        |
| GL-N             | VA-SH-F | 87.6                                       | 79.0        | 68.3        | 41.4                  | 39.5        | 37.4        | 79.3                  | 68.6        | 58.0        | 90.2                  | 85.9        | 81.8        |
| SC-N             | LE-F    | 60.1                                       | 59.5        | 57.3        | 38.8                  | 38.2        | 36.2        | 41.0                  | 39.5        | 40.0        | 75.9                  | 73.4        | 72.7        |
| SC-N             | MI-F    | 74.6                                       | 69.6        | 64.3        | 43.0                  | 40.6        | 39.1        | 63.4                  | 58.2        | 53.6        | 85.2                  | 82.3        | 79.5        |
| SC-N             | VA-GL-F | 68.5                                       | 64.3        | 59.7        | 38.8                  | 38.0        | 36.2        | 51.5                  | 48.0        | 44.7        | 77.5                  | 76.2        | 74.3        |
| SC-N             | VA-SC-F | 68.7                                       | 62.0        | 58.1        | 38.3                  | 38.1        | 36.1        | 50.6                  | 48.9        | 46.2        | 76.9                  | 75.9        | 73.9        |
| SH-N             | LE-F    | 67.3                                       | 63.2        | 58.8        | 41.1                  | 39.8        | 38.2        | 40.4                  | 40.3        | 40.9        | 74.6                  | 74.3        | 73.2        |
| SH-N             | MI-F    | 64.0                                       | 61.9        | 59.2        | 43.1                  | 41.6        | 40.1        | 49.6                  | 49.2        | 47.6        | 75.6                  | 75.6        | 74.8        |
| SH-N             | VA-GL-F | 67.3                                       | 63.3        | 58.8        | 41.5                  | 40.2        | 38.3        | 43.5                  | 43.8        | 43.2        | 75.0                  | 74.8        | 74.1        |
| SH-N             | VA-SC-F | 67.4                                       | 63.5        | 59.3        | 42.1                  | 40.2        | 38.5        | 43.5                  | 43.1        | 42.7        | 74.8                  | 74.6        | 73.7        |

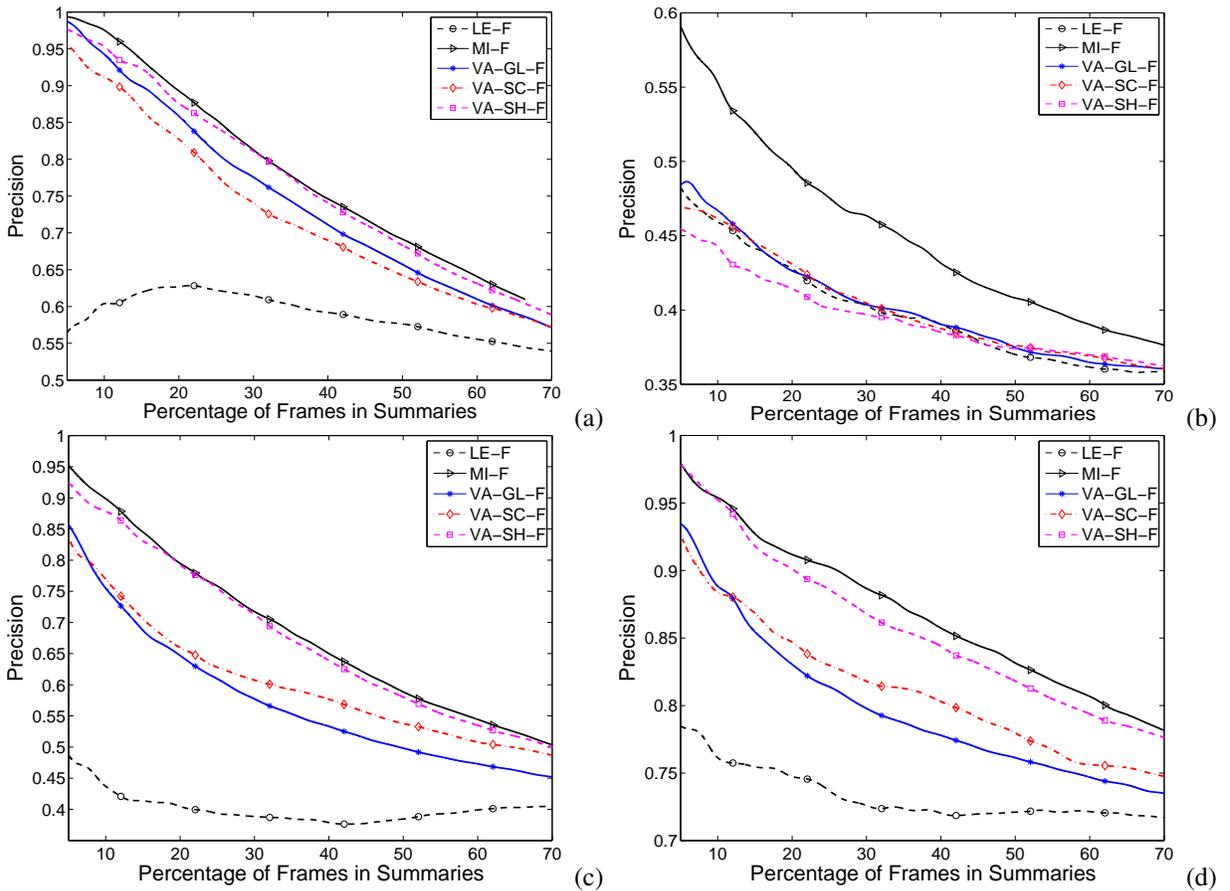


Fig. 4. Frame-level summarization precision scores for (a) audio feature fusion, (b) video feature fusion, (c),(d) multimodal (audio-visual-text) saliency fusion. Multimodal fusion is evaluated on (c) audio-visual human annotation and (d) audio-visual-semantic human annotation.

V, AV, AVS) for a two-class classification problem (salient-nonsalient frames). Each classifier output is thus an indicator function of salient frames, that can be used to form learning-based summaries. In order to select the frames that most likely correspond to event candidates for a summary, results are

reported using a median-filtered version on the raw classifier output with window of length  $2M + 1$ . To obtain results for variable compression rates, a confidence score is defined for each classification result, i.e., each frame. We set that to be the fraction of the  $k$  nearest neighbors with labels in class

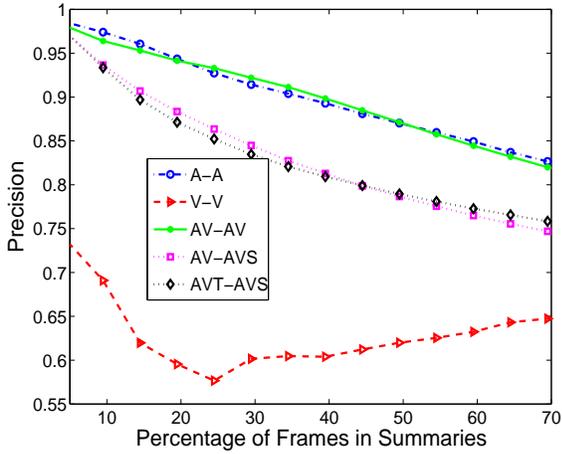


Fig. 5. Frame-level NNR- $k$  classification precision using different (features-annotation) salient class labeling for audio (A-A), visual (V-V), audiovisual (AV-AV, AV-AVS), audio-visual-text (AVT-AVS).

1 (salient events); this roughly corresponds to the posterior probability of event class for that frame; for details see [75].

Frame precision results are shown in Fig. 5, for saliency classification from audio on audio (A-A), visual on visual (V-V), audiovisual on audio-visual-semantic (AV-AVS) and audio-visual-text (AVT-AVS). A seven-fold cross-validation was used in a leave-one-movie-out manner, i.e., NNR- $k$  models are trained on six movies and tested on the seventh. The parameters were empirically set to  $k = 250$  neighbors for NNR and  $M = 2$  sec for median filtering, by optimizing for audio classification accuracy scores. In general, the precision scores achieved using the classifiers are better than those from the bottom-up saliency approach for the monomodal saliency (A) and (V) and the multimodal audiovisual (AV) schemes, on the expense of the need for training and parameter-validation ( $k$  and  $M$ ). However, results from the fused, multimodal AVT curve (Fig. 6) are better than classification on the pooled audio-visual-text feature vector.

### C. Subjective Evaluation of Summaries

Summaries obtained for  $c = 0.2$  (5 times faster than real time) were subjectively evaluated by 11 naive and 3 expert subjects in terms of informativeness and enjoyability on a 0 – 100% scale similarly to [10], [13] for six out of seven movies of the database. In total, five automatically produced summaries were used with global normalization (GL-N) and the following fusion schemes: LE-F (baseline method), inverse variance (VA-GL-F, VA-SC-F, VA-SC-F) and min fusion (MI-F). In addition, an “expert” summary, manually created by a movie production professional, was included in the evaluation. The 14 subjects that participated in the study viewed the original thirty-minute clip, for each of the movies, followed by the six summaries (six-minute each) in random order.

To better normalize the ratings the following scale was communicated to the subjects: poor between 0 – 40%, fair 40 – 60%, good 60 – 75%, very good 75 – 90% and excellent 90 – 100%. Average subjective informativeness and enjoyability ratings for the five fusion schemes and the expert

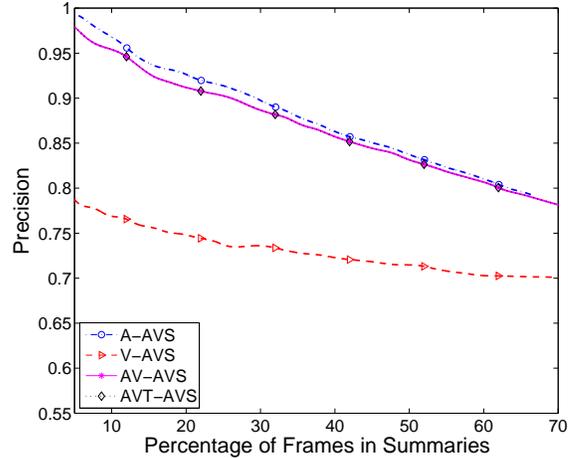


Fig. 6. Frame-level summarization precision scores for audio saliency (A-AVS), visual saliency (V-AVS) and audio-visual-text saliency curve (AVT-AVS) using GL-N MI-F, all evaluated on audio-visual-semantic annotation.

summary are shown in Table IV for each of the six movies, along with objective scores (frame-level accuracy evaluated on AVS annotation) shown for comparison.

Expert summaries achieved very high subjective ratings, up to 95% for both informativeness and enjoyability. Automatic summaries also received high scores, up to 80%, but the performance gap between manually and automatically created summaries is large. This is expected since professionally created summaries contain no artifacts such as abrupt changes in the audio and visual stream and use high-level information (semantics/pragmatics) to select the relevant sub-clips. High marks for expert summaries might also be attributed to subject bias: professional skims were clearly a class apart from the automatically created summaries and subjects might have over-emphasized this difference.

The best performing fusion scheme across all six movies for subjective informativeness ratings is MI-F, followed by VA-SH-F and LE-F. The results are consistent with objective evaluation results with the exception of LE-F that gets surprisingly high subjective ratings (this could be due to the good performance of LE-F for the visual stream that is perceptually important). Note that performance is movie dependent, i.e., VA-SC-F and VA-GL-F also score well for certain movies. The high (absolute) informativeness scores for fantasy/epic and animated films (LOR, GLA, FNE) may be due to sharp scene changes, crisp narration, high-intensity color motifs and audio effects. Low informativeness for CRA and CHI summaries could be due to the long dialogue segments and music scenes, respectively, that are hard to automatically summarize.

Subjective enjoyability scores are higher for VA-SH-F and VA-SC-F fusion. Among the global fusion schemes MI-F performs best followed by LE-F and VA-GL-F. Global, non-adaptive fusion schemes tend to select short segments, resulting in summaries that feel “choppy” or “fast-forward” like. VA-SC-F selects longer segments but might miss important plot elements, thus often forming enjoyable skims that are not necessarily very informative. An expert subject evaluated the seventh movie (BMI) in terms of both informativeness and

TABLE IV  
SUBJECTIVE EVALUATION OF AVT SUMMARIES AT ( $\times 5$ ) RATE AND  
FRAME PRECISION ON AVS ANNOTATION SHOWN FOR COMPARISON.

| Subjective Informativeness           |             |             |             |             |             |             |             |
|--------------------------------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|
| Movie                                | CHI         | CRA         | DEP         | GLA         | LOR         | FNE         | Mean        |
| Expert                               | 91.5        | 90.1        | 83.1        | 87.9        | 88.7        | 92.5        | 89.0        |
| LE-F                                 | 70.3        | 65.4        | 62.6        | 67.6        | 72.3        | 63.7        | 67.0        |
| VA-GL-F                              | 61.4        | <b>66.6</b> | 62.3        | 57.3        | 58.6        | 71.0        | 62.9        |
| VA-SC-F                              | <b>71.5</b> | 65.9        | 67.5        | 52.4        | 52.8        | 58.4        | 61.4        |
| VA-SH-F                              | 53.9        | 59.5        | 69.1        | <b>71.1</b> | <b>78.7</b> | 73.4        | 67.6        |
| MI-F                                 | 50.9        | 62.9        | <b>72.6</b> | 70.8        | 77.4        | <b>74.1</b> | <b>68.1</b> |
| Subjective Enjoyability              |             |             |             |             |             |             |             |
| Movie                                | CHI         | CRA         | DEP         | GLA         | LOR         | FNE         | Mean        |
| Expert                               | 89.0        | 93.2        | 90.1        | 92.1        | 92.3        | 95.6        | 92.1        |
| LE-F                                 | 68.3        | 62.0        | 62.0        | 65.5        | 62.7        | 66.2        | 64.4        |
| VA-GL-F                              | 61.5        | 65.1        | 54.7        | 58.9        | 57.8        | 69.7        | 61.3        |
| VA-SC-F                              | <b>74.3</b> | <b>75.1</b> | <b>80.8</b> | 68.8        | 70.7        | 69.3        | 73.2        |
| VA-SH-F                              | 66.5        | 65.1        | 74.3        | <b>76.8</b> | <b>80.9</b> | <b>81.0</b> | <b>74.1</b> |
| MI-F                                 | 55.3        | 69.9        | 71.0        | 66.7        | 74.4        | 78.3        | 69.3        |
| Objective Accuracy (Frame Precision) |             |             |             |             |             |             |             |
| Movie                                | CHI         | CRA         | DEP         | GLA         | LOR         | FNE         | Mean        |
| LE-F                                 | 56.6        | 58.5        | 59.5        | 46.7        | 46.7        | 55.5        | 53.9        |
| VA-GL-F                              | 62.4        | 68.1        | 68.2        | 59.1        | 56.9        | 65.2        | 63.3        |
| VA-SC-F                              | 63.9        | 69.0        | 64.8        | 60.1        | 59.1        | 63.9        | 63.5        |
| VA-SH-F                              | 67.0        | <b>70.7</b> | <b>74.3</b> | 68.3        | 64.5        | <b>70.0</b> | 69.1        |
| MI-F                                 | <b>67.4</b> | 70.1        | 73.6        | <b>69.4</b> | <b>65.8</b> | 69.8        | <b>69.4</b> |

enjoyability concluding that MI-F and VA-SH-F are the best performing fusion schemes on both subjective qualities.

Overall, minimum and inverse variance fusion schemes perform best with respect to both informativeness and enjoyability subjective scores. The performance in informativeness ratings of linear fusion is somewhat surprising considering its poor detection performance. Shot- and scene-based adaptive inverse variance fusion performed best in terms of skim enjoyability.

## X. CONCLUSION

A multimodal saliency curve integrating cues from the aural, visual and text streams of videos was proposed based on audio, image and language processing, and hierarchical, low-level fusion. Used as an indicator function for attention-invoking salient event detection, the developed representation formed the basis for dynamic movie summarization under a scalable, generic and content-independent algorithm. Summarization performance was quantitatively and qualitatively evaluated on a movie database with multilayer, multimodal saliency annotation. Subjective evaluations showed that the saliency-based video skims can have both functional and aesthetic value, i.e., being informative and pleasing to the viewer. Among the various explored feature normalization, adaptation and fusion schemes, global normalization (GL-N), shot-variance (VA-SH-F) and min-fusion (MI-F) schemes work very well for both intra- and intermodal fusion. Min fusion proved well-suited for cue integration within modalities, especially for visual features and shorter summaries. Minimum and inverse variance schemes performed best in terms of informativeness, while adaptive shot- and scene-based inverse variance in terms of enjoyability. Extensions of this work include: the development of mid-level fusion algorithms, both inside and across

modalities, such as learning schemes and nonlinear feature correlations; incorporation of higher-level features from movie transcript information; and top-down modeling through movie semantics syntax, and structure for bridging the semantic gap. Sample video skims and on-going evaluations can be found at <http://cvsp.cs.ntua.gr/research>.

## ACKNOWLEDGMENTS

The authors wish to thank: the students and staff of CVSP Lab, NTUA for participating in the subjective evaluation studies; N. Malandrakis and I. Rodomagoulakis for the additional movie annotations; T. Apostolidis for providing the expert movie summaries; the anonymous reviewers for their suggestions towards improving this paper.

## REFERENCES

- [1] M. I. Posner and S. E. Petersen, "The attention system of the human brain," *Annual Review of Neuroscience*, vol. 13, no. 1, pp. 25–42, 1990.
- [2] E. I. Knudsen, "Fundamental components of attention," *Annual Review of Neuroscience*, vol. 30, no. 57–58, Jun. 2007.
- [3] 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. 437–55, Aug. 2007.
- [4] R. Desimone and J. Duncan, "Neural mechanisms of selective visual attention," *Annual Review Neuroscience*, vol. 18, pp. 193–222, Jan. 1995.
- [5] 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.
- [6] 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.
- [7] L. Itti and C. Koch, "Computational modelling of visual attention," *Nature Reviews Neuroscience*, vol. 2, no. 3, pp. 194–203, 2001.
- [8] C. Kayser, C. Petkov, M. Lippert, and N. Logothetis, "Mechanisms for allocating auditory attention: an auditory saliency map," *Current Biology*, vol. 15, no. 21, pp. 1943–1947, 2005.
- [9] J. M. Zacks, N. K. Speer, K. M. Swallow, and C. J. Maley, "The brain's cutting-room floor: Segmentation of narrative cinema," *Frontiers in human neuroscience*, vol. 4, pp. 1–15, Oct. 2010.
- [10] Y. Ma, X. 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.
- [11] C.-W. Ngo, Y.-F. Ma, and H.-J. Zhang, "Video summarization and scene detection by graph modeling," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 15, no. 2, pp. 296–305, Feb. 2005.
- [12] J. You, G. Liu, L. Sun, and H. Li, "A multiple visual models based perceptive analysis framework for multilevel video summarization," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 17, no. 3, pp. 273–285, Mar. 2007.
- [13] 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. IEEE Int'l Conf. Acoustics, Speech and Signal Processing (ICASSP)*, Taipei, Taiwan, 19–24 Apr. 2009.
- [14] 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.
- [15] Y. Li, S.-H. Lee, C.-H. Yeh, and C.-C. Kuo, "Techniques for movie content analysis and skimming: tutorial and overview on video abstraction techniques," *IEEE Signal Process. Mag.*, vol. 23, no. 2, pp. 79–89, Mar. 2006.
- [16] B. T. Truong and S. Venkatesh, "Video abstraction: A systematic review and classification," *ACM Trans. Multimedia Comput. Commun. Appl.*, vol. 3, no. 1, Feb. 2007.
- [17] M. Ellouze, N. Boujemaa, and A. M. Alimi, "IM(S)2: Interactive movie summarization system," *J. Visual Communication and Image Representation*, vol. 21, no. 4, pp. 283–294, May 2010.
- [18] 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.

- [19] Y. Zhuang, Y. Rui, T. Huang, and S. Mehrotra, "Adaptive key frame extraction using unsupervised clustering," in *Proc. IEEE Int'l Conf. Image Processing (ICIP)*, Chicago, IL, 4-7 Oct. 1998, pp. 866-870.
- [20] S. Uchihashi, J. Foote, A. Girgensohn, and J. Boreczky, "Video Manga: generating semantically meaningful video summaries," in *Proc. 7th ACM Int'l Conf. on Multimedia*, Orlando, FL, 1999, pp. 383-392.
- [21] A. Hanjalic and H. J. Zhang, "An integrated scheme for automated video abstraction based on unsupervised cluster-validity analysis," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 9, no. 8, pp. 1280-1289, Dec. 1999.
- [22] Y. H. Gong and X. Liu, "Video summarization using singular value decomposition," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, Hilton Head, SC, 13-15 Jun. 2000, pp. 174-180.
- [23] X. Orriols and X. Binefa, "An EM algorithm for video summarization, generative model approach," in *Proc. IEEE Int'l Conf. Computer Vision (ICCV)*, vol. 2, Vancouver, BC, 07-14 Jul. 2001, pp. 335-342.
- [24] R. Lienhart, S. Pfeiffer, and W. Effelsberg, "Video abstracting," *Commun. ACM*, vol. 40, no. 12, pp. 54-62, Dec. 1997.
- [25] M. Smith and T. Kanade, "Video skimming and characterization through the combination of image and language understanding techniques," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, San Juan, Puerto Rico, 17-19 Jun. 1997, pp. 775-781.
- [26] D. Ponceleon, A. Amir, S. Srinivasan, T. Syeda-Mahmood, and D. Petkovic, "CueVideo: automated multimedia indexing and retrieval," in *Proc. 7th ACM Int'l Conf. on Multimedia*, Orlando, FL, 1999.
- [27] A. Hauptmann, "Lessons for the future from a decade of Informedia video analysis research," in *Proc. 4th Int'l Conf. Image and Video Retrieval (CIVR)*, Singapore, 20-22 Jul. 2005.
- [28] M. Rautiainen, T. Ojala, and T. Seppanen, "Cluster-temporal browsing of large news video databases," in *Proc. IEEE Int'l Conf. on Multimedia and Expo (ICME)*, vol. 2, 2004, pp. 751-754.
- [29] D. Parkhurst, K. Law, and E. Niebur, "Modeling the role of saliency in the allocation of overt visual attention," *Vision Research*, vol. 42, no. 1, pp. 107-123, Jan. 2002.
- [30] A. Toet, "Computational versus psychophysical image saliency: A comparative evaluation study," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 11, pp. 2131-2146, Nov. 2011.
- [31] A. Torralba, A. Oliva, M. S. Castelhana, and J. M. Henderson, "Contextual guidance of eye movements and attention in real-world scenes: The role of global features in object search," *Psychological Review*, vol. 113, no. 4, pp. 766-786, 2006.
- [32] H. Li and K. N. Ngan, "A co-saliency model of image pairs," *IEEE Trans. Image Process.*, vol. 20, no. 12, pp. 3365-75, Dec. 2011.
- [33] J. M. Wolfe and T. S. Horowitz, "What attributes guide the deployment of visual attention and how do they do it?" *Nature reviews. Neuroscience*, vol. 5, no. 6, pp. 495-501, Jun. 2004.
- [34] R. Achanta, S. Hemami, F. Estrada, and S. Susstrunk, "Frequency-tuned salient region detection," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, Miami, FL, 20-26 Jun. 2009, pp. 1597-1604.
- [35] C. Guo and L. Zhang, "A novel multiresolution spatiotemporal saliency detection model and its applications in image and video compression," *IEEE Trans. Image Process.*, vol. 19, no. 1, pp. 185-198, Jan. 2010.
- [36] T. Kadir and M. Brady, "Saliency, scale and image description," *Int'l J. Computer Vision*, vol. 45, no. 2, pp. 83-105, Nov. 2001.
- [37] N. Bruce and J. Tsotsos, "Saliency, attention, and visual search: an information theoretic approach," *Journal of Vision*, vol. 9, no. 3, 2009.
- [38] D. Gao, S. Han, and N. Vasconcelos, "Discriminant saliency, the detection of suspicious coincidences, and applications to visual recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 6, pp. 989-1005, Jun. 2009.
- [39] I. Gkioulekas, G. Evangelopoulos, and P. Maragos, "Spatial Bayesian surprise for image saliency and quality assessment," in *Proc. IEEE Int'l Conf. Image Processing (ICIP)*, Hong Kong, 26-29 Sep. 2010, pp. 1081-1084.
- [40] K. Rapantzikos, Y. Avrithis, and S. Kollias, "Dense saliency-based spatiotemporal feature points for action recognition," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 2009.
- [41] —, "Spatiotemporal features for action recognition and salient event detection," *Cognitive Computation*, vol. 3, no. 1, pp. 167-184, 2011.
- [42] S. N. Wrigley and G. J. Brown, "A computational model of auditory selective attention," *IEEE Trans. Neural Netw.*, vol. 15, no. 5, pp. 1151-63, Sep. 2004.
- [43] E. R. Hafter *et al.*, "Auditory Attention and Filters," in *Auditory Perception of Sound Sources*, 2007, vol. 29.
- [44] B. D. Coensel and D. Botteldooren, "A model of saliency-based auditory attention to environmental sound," in *20th Int'l Congress on Acoustics (ICA)*, Sydney, Australia, 2010.
- [45] O. Kalinli and S. Narayanan, "Prominence detection using auditory attention cues and task-dependent high level information," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 17, no. 5, pp. 1009-1024, Jul. 2009.
- [46] X. Zhuang, X. Zhou, M. A. Hasegawa-Johnson, and T. S. Huang, "Real-world acoustic event detection," *Pattern Recognition Letters*, vol. 31, no. 12, pp. 1543-1551, Sep. 2010.
- [47] A. Potamianos and P. Maragos, "Speech formant frequency and bandwidth tracking using multiband energy demodulation," *Journal of the Acoustical Society of America*, vol. 99, no. 6, pp. 3795-3806, Jun. 1996.
- [48] G. Evangelopoulos and P. Maragos, "Multiband modulation energy tracking for noisy speech detection," *IEEE Trans. Audio, Speech, Language Process.*, vol. 14, no. 6, pp. 2024-2038, Nov. 2006.
- [49] D. Dimitriadis, P. Maragos, and A. Potamianos, "Robust AM-FM features for speech recognition," *IEEE Signal Process. Lett.*, vol. 12, no. 9, pp. 621-624, Sep. 2005.
- [50] G. Salton and C. Buckley, "Term-weighting approaches in automatic text retrieval," in *Information Processing and Management*, vol. 24, no. 5, 1988, pp. 513-523.
- [51] C. Lioma and R. Blanco, "Part of speech based term weighting for information retrieval," in *Proc. 31th European Conf. on Information Retrieval Research*, Toulouse, France, 6-9 Apr. 2009, pp. 412-423.
- [52] D. Cristea and O. Postolache, "Summarisation through discourse structure," in *Proc. 6th Int'l Conf. Computational Linguistics and Intelligent Text Processing (CICLing)*, Mexico City, Mexico, 13-19 Feb. 2005.
- [53] L. Hennig, W. Umbrath, and R. Wetzker, "An ontology-based approach to text summarization," in *Proc. IEEE/WIC/ACM Int'l Conf. on Web Intelligence and Intelligent Agent Technology*, 2008, pp. 291-294.
- [54] G. Erkan and D. R. Radev, "LexRank: graph-based lexical centrality as saliency in text summarization," *Journal of Artificial Intelligence Research*, vol. 22, pp. 457-479, Dec. 2004.
- [55] W. T. Chuang and J. Yang, "Extracting sentence segments for text summarization: a machine learning approach," in *Proc. 23rd Int'l ACM SIGIR Conf. Information Retrieval*, Athens, Greece, 2000, pp. 152-159.
- [56] J. Kupiec, J. Pedersen, and F. Chen, "A trainable document summarizer," in *Proc. 18th Int'l ACM SIGIR Conf. Information Retrieval*, Seattle, WA, 1995, pp. 68-73.
- [57] J. L. Neto, A. A. Freitas, and C. A. A. Kaestner, "Automatic text summarization using a machine learning approach," in *Proc. 16th Brazilian Symp. Artificial Intelligence*, 11-14 Nov. 2002, pp. 205-215.
- [58] P. Maragos, J. Kaiser, and T. Quatieri, "Energy separation in signal modulations with application to speech analysis," *IEEE Trans. Signal Process.*, vol. 41, no. 10, pp. 3024-3051, Oct. 1993.
- [59] R. P. Carlyon, "How the brain separates sounds," *Trends in Cognitive Sciences*, vol. 8, no. 10, pp. 465-71, Oct. 2004.
- [60] A. Bovik, P. Maragos, and T. Quatieri, "AM-FM energy detection and separation in noise using multiband energy operators," *IEEE Trans. Signal Process.*, vol. 41, no. 12, pp. 3245-3265, Dec. 1993.
- [61] A. L. Yuille and H. H. Bülthoff, "Perception as Bayesian inference," in *Bayesian Decision Theory and Psychophysics*. Cambridge University Press, 1996, pp. 123-161.
- [62] E. Hering, *Outlines of a Theory of the Light Sense*. Harvard Univ. Press, 1964.
- [63] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Trans. on Pattern Anal. Mach. Intell.*, vol. 20, no. 11, pp. 1254-1259, 1998.
- [64] K. G. Derpanis and J. M. Gryn, "Three-dimensional n-th derivative of Gaussian separable steerable filters," in *Proc. IEEE Int'l Conf. Image Processing (ICIP)*, Genoa, Italy, 11-14 Sep. 2005, pp. III-553-6.
- [65] B. Pellom and K. Hacioglu, "SONIC: The University of Colorado continuous speech recognizer," University of Colorado, Boulder, Tech. Rep. TR-CSLR-2001-01, Mar. 2001.
- [66] M. Gales, "Maximum likelihood linear transformations for HMM-based speech recognition," *Comp. Speech & Language*, vol. 12, no. 2, pp. 75-98, Apr. 1998.
- [67] L. Deng, A. Acero, M. Plumpe, and X. Huang, "Large-vocabulary speech recognition under adverse acoustic environments," in *Proc. Int'l Conf. Spoken Language Processing (ICSLP)*, Denver, CO, 16-20 Sep. 2000, pp. 806-809.
- [68] D. Kim and M. Gales, "Noisy constrained maximum-likelihood linear regression for noise-robust speech recognition," *IEEE Trans. Audio, Speech, Language Process.*, vol. 19, no. 2, pp. 315-325, Feb. 2011.
- [69] H. Schmid, "Probabilistic part-of-speech tagging using decision trees," in *Proc. Int'l Conf. New Methods in Language Processing*. Manchester, UK, 1994, pp. 44-49.
- [70] D. Jurafsky and J. Martin, *Speech and Language Processing*, 2nd ed. Prentice Hall, 2008.

- [71] P. K. Atrey, M. A. Hossain, A. El Saddik, and M. S. Kankanhalli, "Multimodal fusion for multimedia analysis: a survey," *Multimedia Systems*, vol. 16, no. 6, pp. 345–379, Apr. 2010.
- [72] S. Amari, "Integration of stochastic models by minimizing  $\alpha$  - divergence," *Neural Computation*, vol. 19, no. 10, pp. 2780–2796, Oct. 2007.
- [73] N. Malandrakis, A. Potamianos, G. Evangelopoulos, and A. Zlatintsi, "A supervised approach to movie emotion tracking," in *Proc. IEEE Int'l Conf. Acoustics, Speech and Signal Processing (ICASSP)*, Prague, 22-27 May 2011, pp. 2376–2379.
- [74] M. Kipp, "Multimedia annotation, querying, and analysis in ANVIL," in *Multimedia Information Extraction*, 2012, pp. 351–367.
- [75] A. Zlatintsi, P. Maragos, A. Potamianos, and G. Evangelopoulos, "A saliency-based approach to audio event detection and summarization," in *Proc. 20th European Signal Processing Conference (EUSIPCO)*, Bucharest, Romania, 27-31 Aug. 2012.



**Georgios Evangelopoulos** (S '02, M '07) received the Diploma (2001) and Ph.D. (2007) degrees in Electrical & Computer Engineering (ECE) from National Technical University of Athens (NTUA), Greece. He was previously appointed a postdoctoral research associate at NTUA, adjunct lecturer in School ECE, NTUA and postdoctoral fellow at Department of Computer Science, University of Houston. He was involved in National, European Union and United States research programs on speech, multimedia and biometrics. His research interests lie

in the areas of Image Analysis and Computer Vision, Face Recognition, Multimedia/Multimodal Abstraction, Computational Attention and Audio/Speech Analysis. He is a member of IEEE and IEEE Signal Processing Society and serves as a regular reviewer and conference reviewer/technical committee member for IEEE.



**Athanasia Zlatintsi** (S17) received the Master of Science in media technology from the Royal Institute of Technology (KTH), Stockholm, Sweden, in 2006. Since 2007 she has been a research assistant at the Computer Vision, Speech Communication, and Signal Processing Group, NTUA, participating in research projects while she is currently working towards her Ph.D. degree. Her research interests lie in the areas of music and audio signal processing and include analysis and recognition.



**Alexandros Potamianos** received the Diploma in Electrical and Computer Engineering from the National Technical University of Athens, Greece in 1990. He received the M.S and Ph.D. degrees in Engineering Sciences from Harvard University, Cambridge, MA, USA in 1991 and 1995, respectively. From 1995 to 1999 he was a Senior Technical Staff Member at the Speech and Image Processing Lab, AT&T Shannon Labs, Florham Park, NJ. From 1999 to 2002 he was a Technical Staff Member and Technical Supervisor at the Multimedia Communications

Lab at Bell Labs, Lucent Technologies, Murray Hill, NJ. In the spring of 2003, he joined the Dept. of ECE at the Technical University of Crete, Chania, Greece as an associate professor. His current research interests include speech processing, analysis, synthesis and recognition, dialog and multimodal systems, nonlinear signal processing, natural language understanding, artificial intelligence and multimodal child-computer interaction. Prof. Potamianos has authored or co-authored over a hundred papers in professional journals and conferences. He is the co-author of the paper "Creating conversational interfaces for children" that received a 2005 IEEE Signal Processing Society Best Paper Award. He holds four patents. He is a member of IEEE Speech & Language TC and Multimedia TC.



**Petros Maragos** (F17) received the EE Diploma from NTUA in 1980 and the M.Sc. and Ph.D. from Georgia Tech in 1982 and 1985. He has worked as ECE professor at Harvard University (1985-1993), at Georgia Tech (1993-1998), and at NTUA (1998-present). His research interests include image processing and computer vision, audio, speech and language processing, multimedia, cognitive systems, and robotics. He has served as an associate editor for several IEEE Transactions and other journals; as co-organizer of several conferences and workshops; and as a member of three IEEE SPS committees. He is the recipient or co-recipient of several awards, including a 1987 NSF PYIA, the 1988 IEEE SPS Young Author Paper Award, 1994 IEEE SPS Senior Best Paper Award, 1995 IEEE W.R.G. Baker Prize Award, 1996 Pattern Recognition Honorable Mention Award, 2011 CVPR Gesture Recognition Workshop best paper award, and the 2007 EURASIP Technical Achievement Award. He is a fellow of IEEE and of EURASIP.

**Konstantinos Rapantzikos** received the diploma the MS degree in Electronic and Computer Engineering from the Technical University of Crete (Greece) in 2000 and 2002 respectively, and the Ph.D. degree from the National Technical University of Athens (NTUA) in 2008. His interests include computational modeling of human vision, visual action recognition, biomedical imaging and motion estimation. He has published 13 articles in international journals and books and 25 in proceedings of international conferences.



**Georgios Skoumas** received the Diploma in electronic and computer engineering from the Technical University of Crete in 2008, and the MPhil degree in computer science from the University of Cambridge in 2010. He has been a research assistant at Idiap research institute from January 2011 to August 2011. He is currently pursuing the Ph.D. degree in the Knowledge and Database Systems Laboratory (KDBSL), School of Electrical and Computer Engineering, National Technical University of Athens, under the supervision of Professor Timos Sellis. He received a Marie Curie Fellowship in 2011. His research interests lie in the areas of data mining and machine learning.



**Yiannis Avrithis** is a senior researcher at the Image, Video and Multimedia Systems Laboratory (IVML) of the National Technical University of Athens (NTUA), carrying out research on image and video analysis, computer vision and machine learning, and lecturing in NTUA. His research interests include visual feature detection, representation of visual appearance and geometry, image matching and registration, content-based indexing and retrieval, object detection and recognition, image classification, spatiotemporal image / video segmentation and tracking, and video summarization. He has been involved in 15 European and 9 National research projects, and has published 3 edited volumes, 20 articles in journals, 92 in conferences and workshops, 8 book chapters and 7 technical reports in the above topics. He has contributed to the organization of 19 conferences and workshops, and is a reviewer in 15 scientific journals and 13 conferences.