1

Audio-Visual Attention Modeling and Salient Event Detection

Georgios Evangelopoulos[1], Konstantinos Rapantzikos[1], Petros Maragos[1],
Yannis Avrithis[1], and Alexandros Potamianos[2]

[1] National Technical University of Athens
[2] Technical University of Crete

Although human perception appears to be automatic and unconscious, complex sensory mechanisms exist that form the preattentive component of understanding and lead to awareness. Considerable research has been carried out into these preattentive mechanisms and computational models have been developed for similar problems in the fields of computer vision and speech analysis. The focus here is to explore aural and visual information in video streams for modeling attention and detecting salient events. The separate aural and visual modules may convey explicit, complementary or mutually exclusive information around the detected audiovisual events. Based on recent studies on perceptual and computational attention modeling, we formulate measures of attention using features of saliency for the audiovisual stream. Audio saliency is captured by signal modulations and related multifrequency band features, extracted through nonlinear operators and energy tracking. Visual saliency is measured by means of a spatiotemporal attention model driven by various feature cues (intensity, color, motion). Features from both modules mapped to one-dimensional, time-varying saliency curves, from which statistics of salient segments can be extracted and important audio or visual events can be detected through adaptive, threshold-based mechanisms. Audio and video curves are integrated in a single attention curve, where events may be enhanced, suppressed or vanished. Salient events from the audiovisual curve are detected through geometrical features such as local extrema, sharp transitions and level sets. The potential of inter-module fusion and audiovisual event detection is demonstrated in applications such as video key-frame selection, video skimming and video annotation.

1.1 Approaches and Applications

Attention in perception is formally modeled either by stimulus-driven, bottom-up processes, or by goal-driven, top-down mechanisms that require prior knowledge of the depicted scene or the important events [21]. The former, bottom-up approach is based on signal-level analysis with no prior information reacquired or learning incorporated.

In analyzing the visual and aural information of video streams the main issues that arise are: i) choosing appropriate features that capture important signal properties, ii) combining the information corresponding to the different modalities to allow for interaction and iii) defining efficient salient event detection schemes. In this chapter, the potential of using and integrating aural and visual features is explored, to create a model of audiovisual attention, with application to saliency-based summarization and automatic annotation of videos. The two modalities are processed independently with the saliency of each described by features that correspond to physical changes in the depicted scene. Their integration is performed by constructing temporal indexes of saliency that reveal dynamically evolving audiovisual events.

Multimodal video analysis (i.e., analysis of various information modalities) has gained in popularity with automatic summarization being one of the main targets of research. Summaries provide the user with a short version of the video that ideally contains all important information for understanding the content. Hence, the user may quickly access and evaluate if the video is important, interesting or enjoyable. The tutorial in [38] classifies video abstraction into two main types: *key-frame selection* which yields a static small set of important video frames and *video skimming* (loosely referred to in this chapter as *video summarization*) which results a dynamic short subclip of the original video containing important aural and visual spatiotemporal information.

Earlier works were mainly based on processing only the visual input. Zhuang et al. [40] extracted salient frames based on color clustering and global motion, while Ju et al. [13] used gesture analysis in addition to the latter low-level features. Furthermore Avrithis et al. [2] represent the video content by a high-dimensional feature curve and detect key-frames at the curvature points. Another group of methods is based on frame clustering to select representative frames [30, 37]. Features extracted from each frame of the sequence form a feature vector and are used in a clustering scheme. Frames closer to the centroids are then selected as key-frames. Other schemes based on sophisticated temporal sampling [33], hierarchical frame clustering [30, 10], where the video frames are hierarchically clustered by visual similarity, and fuzzy classification [6] have also proposed summarization schemes with encouraging results.

In an attempt to incorporate multimodal or/and perceptual features in the analysis and processing of the visual input, various systems have been designed and implemented within a variety of projects. The Informedia project and its offsprings combined speech, image, natural language understanding and image processing to automatically index video for intelligent search and retrieval [32, 12, 11]. This approach generated interesting results. In the Video Browsing and Retrieval system (VIRE) [31] a number of low-level visual and audio features are extracted and stored using MPEG-7, while MediaMill [26] provides a tool for automatic shot and scene segmentation for general content. IBMs CueVideo system [1] automatically extracts a number of low- and midlevel visual and audio features. The visually similar shots are clustered using color correlograms. Going one step further towards human perception, Ma et al. [20, 21] proposed a method for detecting the salient parts of a video that is based on user attention models. They used motion, face and camera attention along with audio attention models (audio saliency and speech/music) as cues to capture salient information and identify the audio and video segments to compose a summary.

We present a saliency-based method to detect important audiovisual segments and focus more on the potential benefits of feature-based attention modeling and multi-sensory signal integration. As content importance in a video stream is quite subjective, it is not easy to evaluate methods in the field. Hence, in an attempt to assess the proposed method both quantitatively and qualitatively, we present video summarization results on commercial videos and samples from the MUSCLE movie database[3], annotated with respect to saliency of the scene evaluated by human observers. The reference videos are clips from the movies “300” and “Lord of The Rings I”. Automatic and manual annotations are studied and compared on the selected movie clips with respect to audiovisual saliency of the depicted scenes.

The remaining of the chapter is organized as follows: Section 1.2 and Section 1.3 describe the audio saliency and the visual saliency modules, respectively. Schemes for detecting salient events are proposed in Section 1.4 and experimental evaluation and applications are given in Section 1.5. Conclusions are drawn and open issues for future work are discussed in Section 1.6.

1.2 Audio Saliency

Streams of audio information may be composed from a variety of sounds, like speech, music, environmental sounds (nature, machines, noises), a result of multiple sources that correspond to natural, artificial, man-made, on purpose or randomly occurring phenomena. An audio event is a bounded region in the time continuum, in terms of a beginning and end, that is characterized by a variation or transitional state to one or more sound-producing sources. Events are “sound objects” that change dynamically with time, while retaining a set of characteristic properties that identify a single entity. Perceptually, event boundaries correspond to points of maximum quantitative or qualitative change of physical features [39].

Aural attention is triggered perceptually by changes in the involved events of an audio stream. These may be changes of the nature/source of events, newly introduced sounds, or transitions and abnormalities in the course of a specific event, in real-life or synthetic recordings. Such transitions correspond to changes of salient audio properties, e.g invariants, whose selection is crucial for efficient audio representations for event detection and recognition.

[3] http://poseidon.csd.auth.gr/EN/MUSCLE_moviedb/index.htm

Biological observations indicate that one of the segregations performed by the auditory system in complex channels is in terms of temporal modulations, while according to psychophysical experiments, modulated carriers seem more salient perceptually to human observers compared to stationary signals [17, 36]. Moreover, following Gestalt theories, the salient audio signal structures constitute meaningful audio Gestalts which in turn define manifestations of audio events [24]. Thus, we formulate a curve modeling audio attention based on saliency measures of meaningful temporal modulations in multiple frequencies.

1.2.1 Audio Processing and Salient Features

Processing the audio stream of multimodal systems, involves confronting a number of subproblems that compose what may be thought of as audio understanding. In that direction, the notions of audio events and salient audio segments are the backbone of audio detection, segmentation, recognition and identification. Starting from lower and going toward higher level, i.e., more complicated problems, the subproblems of audio analysis can be roughly categorized as: a) detection, where the presence of auditory information is verified and separated from silence or background noise conditions [7]; b) attention modeling and audio saliency, where the perceptual importance is valued [20, 21]; c) source separation, where the auditory signal is decomposed to different generating sources and sound categories (e.g speech, music, natural or synthetic sounds); d) segmentation and event labeling, where the aural activity is assigned boundaries and dynamic events are sought after [19]; and e) recognition of sources and events, where the sources and events are matched to stored lexicon representations.

Descriptive signal representations are essential for all the above subproblem categories and much work has been devoted in robust audio feature extraction for applications [27, 15, 19, 21]. Psychophysical experiments indicate the nature of features responsible for audio perception [23, 36]. These are representations both in the temporal and spectral domain, that incorporate properties and notions such as scale, structure, dimension and perceptual invariance. Well-established features for audio analysis, classification and recognition include time-frequency representations (e.g., spectrograms), temporal measurements (e.g., energy, zero-crossings rate, pitch, periodicity), spectral measurements (e.g., component or resonance position and variation, bandwidth, spectral flux) and cepstral measurements like the Mel-Frequency Cepstral Coefficients (MFCCs).

Recent advances in the field of nonlinear speech modeling relate salient features of speech signals to their inherent non-stationarity and the presence of micro-modulations in the amplitude and frequency variation of their constructing components. Experimental and theoretical indications about modulations in various scales during speech production led to proposing an AM-FM modulation model for speech in [22]. The model was then employed for extracting various "modulation-based" features like formant tracks and bandwidth, mean amplitude and frequency of the components [25] as well as the coefficients of their energy-frequency distributions (TECCs) [5].

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

Audio AM-FM Modeling and Multiband Demodulation

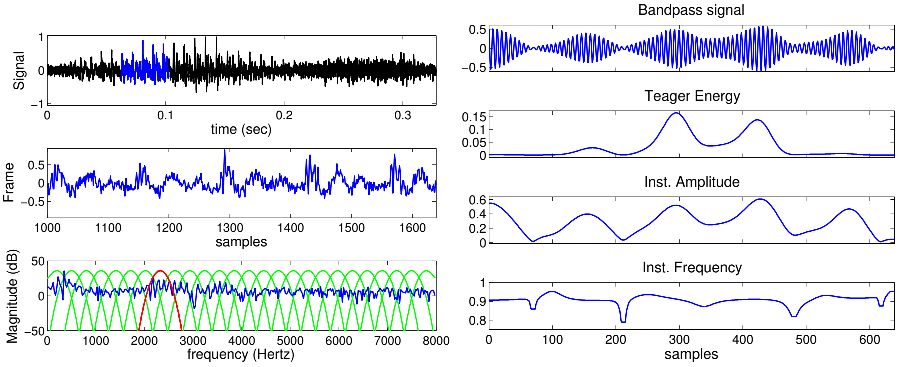
Assume that a single audio component is modeled by a real-valued AM-FM signal of the form $x(t) = a(t)\cos\left(\int_0^t \omega(\tau)d\tau\right)$, with time-varying amplitude envelope $a(t)$ and instantaneous frequency $\omega(t)$ signals. Demodulation of $x(t)$ can be approached via the use of the Teager-Kaiser nonlinear differential energy operator $\Psi[x(t)] \equiv [\dot{x}(t)]^2 - x(t)\ddot{x}(t)$, where $\dot{x}(t) = dx(t)/dt$ [34, 14]. Applied to an AM-FM signal $x(t)$, Ψ yields the instantaneous energy of the source producing the oscillation, i.e., $\Psi[x(t)] \approx a^2(t)\omega^2(t)$, with negligible approximation error under realistic constraints [22]. The instantaneous energy is separated to its amplitude and frequency components by the *energy separation algorithm* (ESA) [22] using Ψ as its main ingredient.

In order to apply ESA for demodulating a wideband audio signal, modeled by a sum of AM-FM components, it is necessary to isolate narrowband components in advance. Bandpass filtering decomposes the signal in frequency bands, each assumed to be dominated by a single AM-FM component in that frequency range. In the *multiband demodulation analysis* (MDA) scheme, components are isolated globally using a set of frequency-selective filters [3, 25, 7]. Here MDA is applied through a filterbank of linearly-spaced Gabor filters $h(t) = \exp(-\alpha^2 t^2)\cos(\omega_c t)$, with ω_c the central filter frequency and α its rms bandwidth. Gabor filters are chosen for being compact and smooth while attaining a minimum joint time-frequency uncertainty [9, 22, 3].

Demodulation via ESA of a single frequency band, obtained by one Gabor filter, can be seen in Fig. 1.1(b). The choice of the specific band corresponds to an energy-based dominant component selection criterion that will be further employed in the following for audio feature extraction. Postprocessing by median filtering may be used to alleviate singularities in the resulting demodulation measurements.

Audio Features

The AM-FM modulation superposition model for speech [22], motivated by the presence of multi-scale modulations during speech production [34], is applied here to generic audio signals. Thus an audio signal is modeled by a



(a) signal, spectrum & filterbank (b) demodulation

Fig. 1.1. Short-time audio processing and dominant modulation extraction. (a) a vowel frame (20ms) from a speech waveform is analyzed in multiple bands (bottom) and (b) the dominant, w.r.t average source energy, band is demodulated in instantaneous amplitude and frequency (smoothed by 13-pt median) signals.

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

To model a discrete-time audio signal $s[n] = s(nT)$, we use K discrete AM-FM components whose instantaneous amplitude and frequency signals are $A_k[n] = a_k(nT)$ and $\Omega_k[n] = T\omega_k(nT)$, respectively. The model parameters are estimated from the K filtered components using a discrete-time energy operator $\Psi_d(x[n]) \equiv (x[n])^2 - x[n-1]x[n+1]$ and a related discrete ESA, which is a computationally simple and efficient algorithm with an excellent, almost instantaneous, time resolution [22]. Thus, at each sample instance n the audio signal is represented by three parameters (energy, amplitude and frequency) for each of the K components, leading to $3 \times K$ feature vector.

A representation in terms of a single component per analysis frame emerges by maximizing an energy criterion in the multi-dimensional filter response space [3, 7]. For each frame m of N samples duration, the dominant modulation component is the one with *maximum average Teager energy* (MTE):

$$\mathrm{MTE}[m] = \max_{1 \le k \le K} \frac{1}{N} \sum_{n} \Psi_d((s * h_k)[n]), \quad (m-1)N + 1 \le n \le mN \tag{1.1}$$

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

$$j = \arg\max_{1\leq k\leq K}(\overline{\mathit{\Psi}_d[(s*h_k)(n)]}),\ \mathrm{MTE}[m] = (\overline{\mathit{\Psi}_d[(s*h_j)(n)]}) \qquad (1.2)$$

$$\mathrm{MIA}[m] = (\overline{|A_j[n]|})\ ,\ \mathrm{MIF}[m] = (\overline{\mathit{\Omega}_j[n]}). \qquad (1.3)$$

Thus, each frame yields average measurements for the source energy, instant amplitude and frequency from the filter that captures the “strongest” modulation signal component. In this context strength refers to the amount of energy required for producing component oscillations. The dominant component is the most salient signal modulation structure and energy MTE may be thought of as the salient *modulation energy*, jointly capturing essential amplitude-frequency content information.

The resulting three-dimensional feature vector of the mean dominant modulation parameters

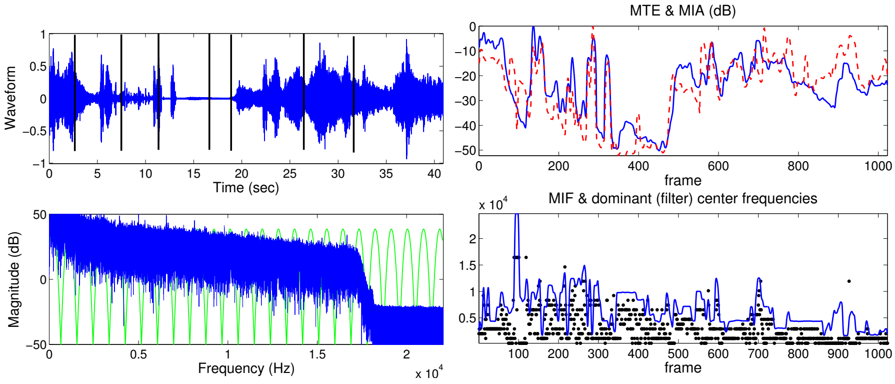
$$\mathbf{F}_{\mathrm{a}}[m] = [F_{\mathrm{a}1}, F_{\mathrm{a}2}, F_{\mathrm{a}3}]\,[m] = [\mathrm{MTE}, \mathrm{MIA}, \mathrm{MIF}]\,[m] \qquad (1.4)$$

is a low dimensional descriptor, compared to the potential $3\times K$ vector from all outputs, of the “average instantaneous” modulation structure of the audio signal involving properties such as level of excitation, rate-of-change, frequency content and source energy.

In discrete implementation, audio analysis frames usually vary between 10-25 ms. For speech signals, such a choice of window length covers all pitch duration diversities between different speakers. Sequentially, the discrete energy operator is applied to the set of filter outputs and an averaging operation is performed. Central frequency steps of the filter design varying between 200-400 Hz, yield filterbanks consisting of 20-40 filters.

An example of the short-time features extracted from a movie audio stream (1024 frames from “300”) can be seen in Fig. 1.2. The chosen segment was manually annotated by a human observer, with respect to the various sources present and their boundaries. These are indicated by the vertical lines in the signal waveform. The different sources include speech (2 different speakers), music, noise, sound effects and a general “mix-sound” category. The wideband spectrum is decomposed using 25 filters, of 400 Hz bandwidth, and the dominant modulation features are shown in (b), after median (7-point) and Hanning (5-point) post-smoothing. Features are mapped from audio-to-video temporal index by keeping maximum intraframe values. Note how a) the envelope features complement the frequency measure (i.e high-frequency sounds of low energy and the opposite), b) manual labeling matches sharp transitions to one or more features and c) frequency is characterized by longer, piece-wise constant “sustain periods.”

This representation in terms of the salient modulation properties of sounds, is additionally supported by cognitive theories of event perception [23]. For example, rapid amplitude and frequency modulations are related to temporal



(a) audio & wideband spectrum (b) modulation features

Fig. 1.2. Feature extraction from multi-source audio stream. (a) Waveform with manual labeling of the various sources/events (vertical lines) and wideband spectrum with filterbank, (b) top: MTE (solid) and MIA (dashed), bottom: MIF with dominant carrier frequencies superimposed (1024 frames from "300" video).

acoustic micro-properties of sounds that appear to be useful for recognition of sources and events. A simplistic approach for the structure of audio events involves three parts: an onset, a relatively constant duration and an offset portion. Event onset and decay are captured by the envelope variations of the amplitude and energy measurements. On the other hand, spectral variations, retrieved perceptually from the sustain period, and variations in the main signal component are captured by the dominant frequency feature.

1.2.2 Audio Attention Curve

The attention curve for the audio signal is constructed by the saliency values, provided by the set of audio features (1.4). Conceptually, salient information is modeled through source excitation and average rate of spectral and temporal change.

The simplest scenario of an *audio saliency* curve is a weighted linear combination of the normalized audio features

$$S_{\mathrm{a}}[m] = w_1 F_{\mathrm{a1}}[m] + w_2 F_{\mathrm{a2}}[m] + w_3 F_{\mathrm{a3}}[m], \qquad (1.5)$$

where $[w_1, w_2, w_3]$ is a weighting vector. Normalization is performed by least squares fit of their individual value ranges to $[0, 1]$. For this chapter we use equal weights $w_1 = w_2 = w_3 = 1/3$, which amounts to uniform linear averaging and viewing the normalized features $F_{\mathrm{a}i}$ as equally important for the level of saliency and the attention provoked by the audio signal.

A different, perceptually motivated approach is a non-linear feature fusion, based on time-varying "energy weights." According to the structure and rep-

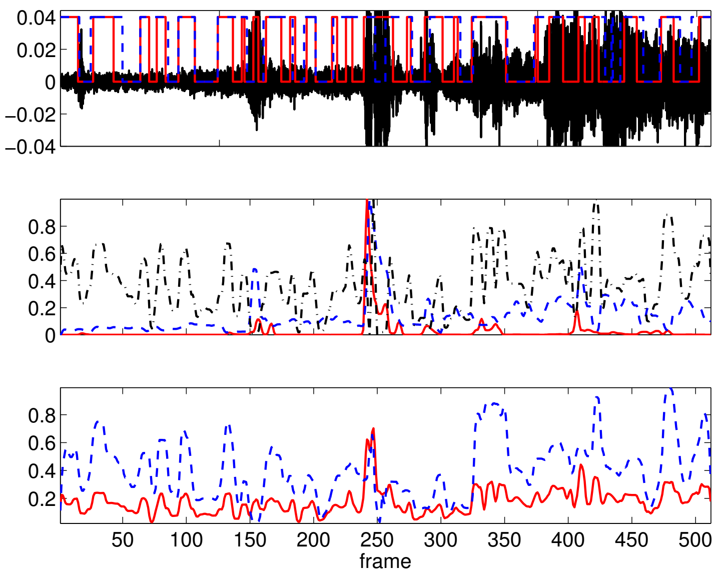


Fig. 1.3. Audio saliency curves. Top: audio waveform and threshold-based saliency indicator functions, Middle: normalized audio features, MTE (solid), MIA (dashed), MIF (dash dotted). Bottom: saliency curve (linear in solid, nonlinear in dashed). Indicator functions correspond to the two audio fusion schemes, (512 frames from the "Lord of the Rings I" stream).

resentation by the auditory system of audio events [23], temporal variation information is extracted by the onset and offset portions, while spectral change, from the intermediate sustain periods. As the energy measurement has been previously used for detecting speech event boundaries [7], we incorporate it as an index of event transitional points. Using the average source energy gradient as a weighting factor, we acquire the following *nonlinear audio-to-audio* integration scheme

$$S_{\mathrm{a}}[m] = w_{\mathrm{e}}[m]F_{\mathrm{a2}}[m] + (1 - w_{\mathrm{e}}[m])F_{\mathrm{a3}}[m], \quad w_{\mathrm{e}} = \left|\frac{dF_{\mathrm{a1}}}{dm}\right| \tag{1.6}$$

The effect of this gradient energy weighting process is that, in sharp event transitions (modeling beginning, ending or change of activity) the amplitude feature is employed more (hence, the temporal variation is more salient). The frequency is weighted more at relatively constant activity periods where the spectral variation is perceptually more important.

An example of the feature integration for saliency curve construction is presented in Fig. 1.3. Audio features, normalized and mapped to the video frame index, are combined linearly by (1.5) or nonlinearly by (1.6) to yield the corresponding saliency curves. A saliency indicator function is then obtained by applying on the resulting curves an adaptive threshold-based detection scheme.

1.3 Visual Saliency

The visual saliency computation module is based on the notion of a centralized saliency map [18] computed through a feature competition scheme. The motivation behind this scheme is the experimental evidence of a biological counterpart in the Human Visual System (interaction/competition among the different visual pathways related to motion/depth (M pathway) and gestalt/depth/color (P pathway) respectively) [16]. An overview of the visual saliency detection architecture is given in Fig. 1.4. In this framework, a video sequence is represented as a solid in the 3D Euclidean space, with time being the third dimension. Hence, the equivalent of a spatial saliency map is a spatiotemporal volume where each voxel has a certain value of saliency. This saliency volume is computed with the incorporation of feature competition by defining cliques at the voxel level and use an optimization procedure with both inter- and intra- feature constraints.

1.3.1 Visual Features

The video volume is initially decomposed into a set of feature volumes, namely intensity, color and spatiotemporal orientations. For the intensity and color features, we adopt the opponent process color theory that suggests the control of color perception by two opponent systems: a blue-yellow and a red-green mechanism. The extent to which these opponent channels attract attention of humans has been previously investigated in detail, both for biological [35] and computational models of attention [20]. According to the *opponent color* scheme, if r, g, b are the red, green and blue volumes respectively, the luminance and color volumes are obtained by

$$I = (r + g + b)/3, \quad RG = R - G, \quad BY = B - Y, \tag{1.7}$$

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

Spatiotemporal orientations are computed using steerable filters [8]. A steerable filter may be of arbitrary orientation and is synthesized as a linear combination of rotated versions of itself. Orientations are obtained by measuring the filter strength along particular directions θ (the angle formed by the plane passing through the t axis and the $x - t$ plane) and ϕ (defined on the $x - y$ plane). The desired filtering can be implemented using the three dimensional filters $G_2^{\theta,\phi}$ (e.g second derivative of a 3D Gaussian) and their Hilbert transforms $H_2^{\theta,\phi}$, by taking the filters in quadrature to eliminate the

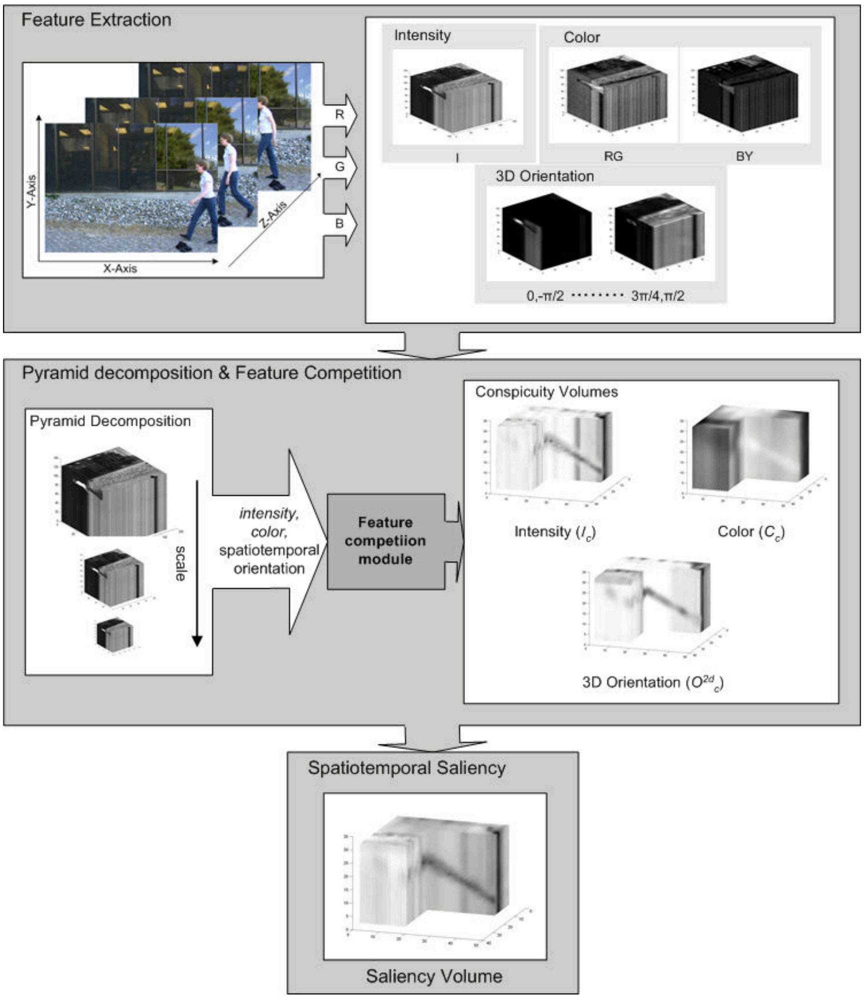


Fig. 1.4. Visual saliency module

phase sensitivity present in the output of each filter. This is called the oriented energy:

$$E(\theta, \phi) = [G_2^{\theta,\phi} * I]^2 + [H_2^{\theta,\phi} * I]^2, \tag{1.8}$$

where

$$\theta \in \{0, \frac{\pi}{4}, \frac{\pi}{2}, \frac{3\pi}{4}\}, \quad \phi \in \{-\frac{\pi}{2}, -\frac{\pi}{4}, 0, \frac{\pi}{4}, \frac{\pi}{2}\}. \tag{1.9}$$

By selecting θ and ϕ as in (1.9), 20 volumes of different spatiotemporal orientations are produced, which must be fused together to produce a single orientation volume that will be further enhanced and compete with the rest of the feature volumes. We use an operator based on Principal Component

Analysis (PCA) and generate a single spatiotemporal orientation conspicuity volume V. More details can be found in [28].

1.3.2 Visual Attention Curve

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

Feature competition is implemented in the model using an energy-based measure. In a regularization framework the first term of this energy measure may be regarded as the data term E_1 and the second as the smoothness one E_2, since it regularizes the current estimate by restricting the class of admissible solutions [29]. The energy involves voxel operations between coarse and finer scales of the volume pyramid, which means that if the center is a voxel at level $c \in \{2, ..., p-d\}$, where p is the maximum pyramid level and d is the desired depth of the center-surround scheme, then the surround is the corresponding voxel at level $h = c+\delta$ with $\delta \in \{1, 2, ..., d\}$. Hence, if we consider the intensity and two opponent color features as elements of the vector $\mathbf{F}_{\mathrm{v}} = F_{\mathrm{v}_1}, F_{\mathrm{v}_2}, F_{\mathrm{v}_3}$ and if $F^0_{\mathrm{v}_k}$ corresponds to the original volume of each of the features, each level ℓ of the pyramid is obtained by convolution with an isotropic 3D Gaussian G and dyadic down-sampling:

$$F^{\ell}_{\mathrm{v}_k} = \left(G * F^{\ell-1}_{\mathrm{v}_k}\right) \downarrow_2, \quad \ell = 1, 2, ..., p. \tag{1.10}$$

where $\downarrow_2$ denotes decimation by 2 in each dimension. For each voxel q of a feature volume F the energy is defined as

$$E_v(F^c_{\mathrm{v}_k}(q)) = \lambda_1 \cdot E_1(F^c_{\mathrm{v}_k}(q)) + \lambda_2 \cdot E_2(F^c_{\mathrm{v}_k}(q)), \tag{1.11}$$

where λ_1, λ_2 are the importance weighting factors for each of the involved terms. The first term of (1.11) is defined as

$$E_1(F^c_{\mathrm{v}_k}(q)) = F^c_{\mathrm{v}_k}(q) \cdot |F^c_{\mathrm{v}_k}(q) - F^h_{\mathrm{v}_k}(q)| \tag{1.12}$$

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

$$E_2(F^c_{\mathrm{v}_k}(q)) = F^c_{\mathrm{v}_k}(q) \cdot \frac{1}{|N(q)|} \cdot \sum_{r \in N(q)} \left(F^c_{\mathrm{v}_k}(r) + V(r)\right), \tag{1.13}$$

where V is the spatiotemporal orientation volume that may be regarded as an indication of motion activity in the scene and $N(q)$ is the 26- neighborhood of voxel q. The second energy term involves competition among voxel neighborhoods of the same volume and allows a voxel to increase its saliency value only if the activity of its surroundings is low enough. The energy is then minimized using an iterative steepest descent scheme and a *saliency volume* S is created by averaging the conspicuity feature volumes $F^1_{\mathrm{v}_k}$ at the first pyramid level:

$$S(q) = \frac{1}{3} \cdot \sum_{k=1}^{3} F^1_{\mathrm{v}_k}(q). \tag{1.14}$$

Overall, the core of the visual saliency detection module is an iterative minimization scheme that acts on 3D local regions and is based on center-surround inhibition regularized by inter- and intra- local feature constraints. A detailed description of the method can be found in [28]. Figure 1.5 depicts the computed saliency for three frames of "Lord of the Rings I" and "300" sequences. High values correspond to high salient areas (notice the shining ring and 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 motion. Each of the feature volumes is first normalized to lie in the range $[0, 1]$ and then point-to-point multiplied by the saliency one in order to suppress low saliency voxels. The weighted average is taken to produce a single *visual saliency* value for each frame:

$$S_{\mathrm{v}} = \sum_{k=1}^{3} \sum_{q} S(q) \cdot F^1_{\mathrm{v}_k}(q), \tag{1.15}$$

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

1.4 Audio-Visual Saliency

Integrating the information extracted from audio and video channels is not a trivial task, as they correspond to different sensor modalities (aural and visual). Audiovisual fusion for modeling multimodal attention can be performed at three levels: i) *low-level* fusion (at the extracted saliency curves), ii) *middle-level* fusion (at the corresponding feature vectors), iii) *high-level* fusion (at the detected salient segments and features of the curves).

In a video stream with both aural and visual information present, audiovisual attention is modeled by constructing a temporal sequence of audiovisual saliency values. In this saliency curve, each value corresponds to a measure of importance of the multi-sensory stream at each time instance. In

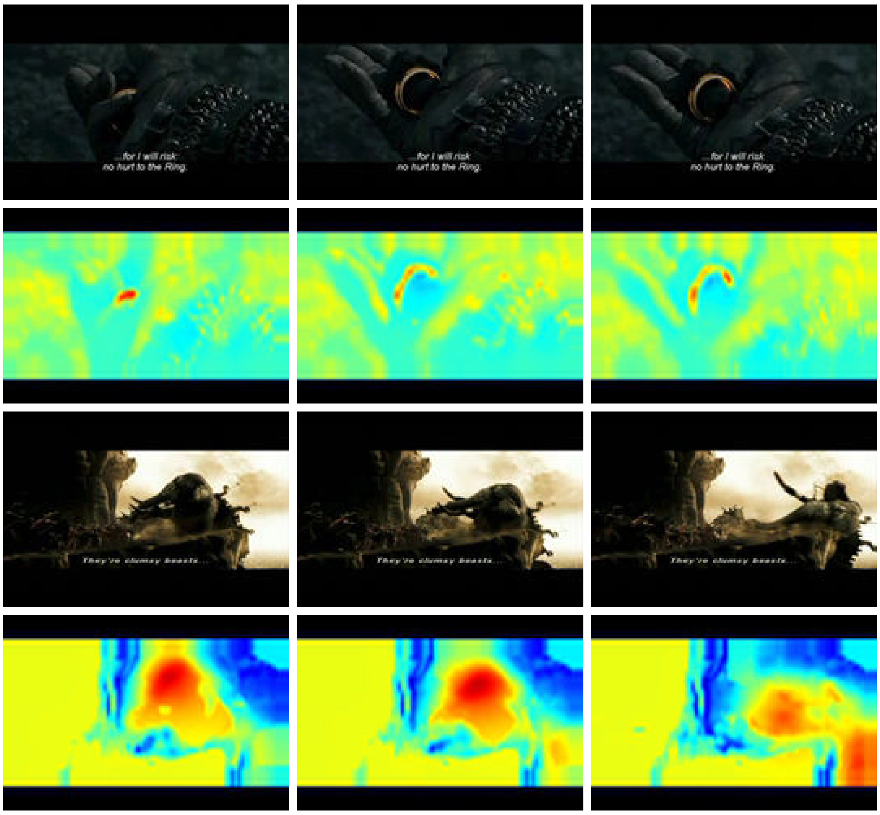


Fig. 1.5. Original frames from the movies "Lord of the Rings I" (top) and "300" (bottom) and the corresponding saliency maps (better viewed in color).

both modalities, features are mapped to saliency (aural and visual) curve values $(S_\mathrm{a}[m], S_\mathrm{v}[m])$, and the two curves are integrated to yield an audiovisual saliency curve

$$S_\mathrm{av}[m] = \mathrm{fusion}(S_\mathrm{a}, S_\mathrm{v}, m), \tag{1.16}$$

where m the frame index and $\mathrm{fusion}(\cdot)$ is the process of combining or fusing the two modalities. This is a low-level fusion scheme. In general, this process of combining the outputs of the two saliency detection modules may be nonlinear, have memory or vary with time. For the purposes of this chapter, however, we use the following straightforward linear memoryless scheme

$$S_\mathrm{av}[m] = w_\mathrm{a} \cdot S_\mathrm{a}[m] + w_\mathrm{v} \cdot S_\mathrm{v}[m]. \tag{1.17}$$

Assuming that the individual audio and visual saliency curves are normalized in the range $[0, 1]$ and the weights form a convex combination, this coupled audiovisual curve serves as a continuous-valued indicator function of salient events, in the audio, the video or a common audiovisual domain. The weights can be equal, constant or adaptive depending for example on the uncertainty of the audio or video features. Actually, the above weighted linear scheme corresponds to what is called in [4] *"weak fusion"* of modalities and is optimum under the maximum a posteriori criterion, if the individual distributions are Gaussian and the weights are inversely proportional to the individual variances, as explained in Section **??** of Chapter **??**.

The coupled audiovisual saliency curve provides the basis for subsequent detection of salient events. Audiovisual events are defined as bounded time-regions of aural and visual activity. In the proposed method, events correspond to attention-triggering signal portions or points of interest extracted from the saliency curves. The boundaries of events and the activity locus points, correspond to a maximum change in the audio and video saliency curves and the underlying features. Thus, transition and reference points in the audiovisual event stream can be tracked by analyzing the geometric features of the curve. Such geometric characteristics include:

- **Extrema points**: these are the local maxima or minima of the curve and can be detected by a 'peak-peaking' method.
- **Peaks & Valleys**: the region of support around maxima and minima, respectively. These can be extracted automatically (e.g., by a percentage to maximum) or via a user-defined scenario depending on the application (e.g., a skimming index).
- **Edges**: One-dimensional edges correspond to sharp transition points in the curve. A common approach is to detect the zero-crossings of a Derivative-of-Gaussian operator applied to the signal.
- **Level Sets**: points where the values of the curve exceed a learned or heuristic level-threshold. These sets can define indicator functions of salient activity.

Saliency-based events can be tracked at the individual saliency curves or at the integrated one. In the former case, the resulting geometric feature-events can be subjected to higher-level fusion (e.g., by logical OR, AND operators). As a result, events in one of the modalities may suppress or enhance the events present in the other. A set of audio, visual and audiovisual events can be seen in the example-application of Figs. 1.6 and 1.7. The associated movie-trailer clip contained a variety of events in both streams (soundtracks, dialogues, effects, shot-changes, motion), aimed to attract the viewer's attention. Peaks detected in the audiovisual curve revealed in many cases an agreement between peaks (events) tracked in the individual saliency curves.

1.5 Applications and Experiments

The developed audiovisual saliency curve has been applied to saliency-based video summarization and annotation. Summarization is performed in two directions: key-frame selection for static video storyboards via local maxima

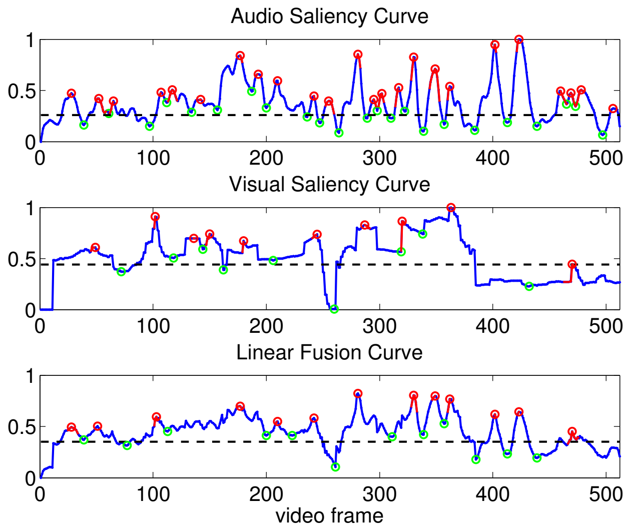


Fig. 1.6. Saliency curves and detected features (maxima, minima, lobes and levels) for audio (top), video (middle) and audiovisual streams (bottom) of the movie trailer "First Descend".

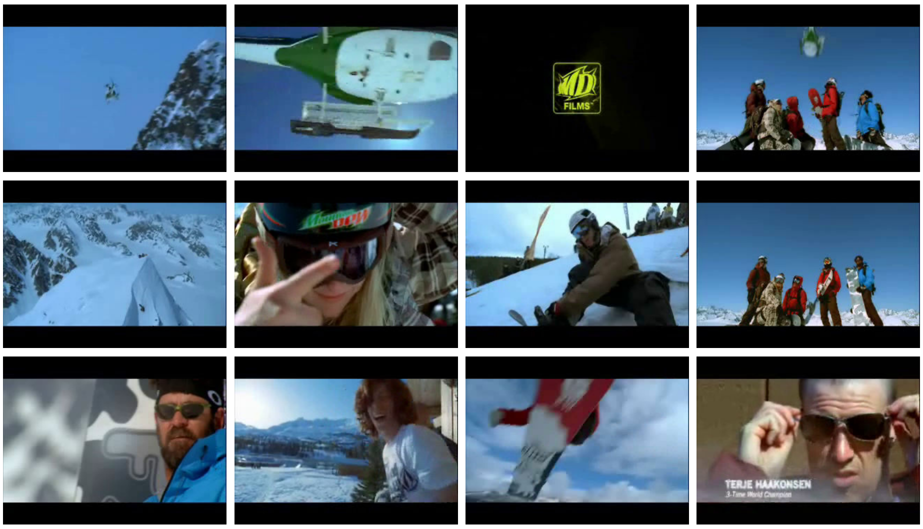


Fig. 1.7. Key-frames selection using local maxima (peaks) of corresponding audio-visual saliency curve. Selected frames correspond to the peaks in the bottom curve of Fig. 1.6 (12 out of 13 frames, peak 4 is not shown).

detection and dynamic video skimming based on a user-defined skimming percentage. Annotation refers to labeling various video parts with respect to their attentional strength, based on sensory information solely. In order to provide statistically robust and as far as possible objective results, the results are compared to human annotation.

1.5.1 Experimental setup

The proposed method has been applied both to videos of arbitrary content and to a human annotated movie database, that consists of 42 scenes extracted from 6 movies of different genres. For demonstration purposes we selected two clips ($\simeq$10 min each) from the movies "Lord of the Rings I" and "300" and present a series of applications and experiments that highlight different aspects of the proposed method.

The clips were viewed and annotated according to the audio, visual and audiovisual saliency of their content. This means that parts of the clip were labeled as salient or non-salient, depending on the importance and the attention attracted by their content. The viewers were asked to assign a saliency factor to any part according to loose guidelines, since strict rules cannot be applied due to the high subjectivity of the procedure. The guidelines were related to the audio-only, visual-only and audiovisual changes and events, but not to semantic interpretation of the content. The output of this procedure is a saliency indicator function, corresponding to the video segments that were assigned a non-zero saliency factor. For example, Fig. 1.8 depicts the saliency curves and detected geometric features, while Fig. 1.9 the indicator functions obtained manually and automatically on a frame sequence from one movie clip.

1.5.2 Key-frame Detection

Key-frame selection to construct a static abstract of a video, was based on the local maxima, through peak detection on the proposed saliency curves. The process and the resulting key-frames are presented in Figs. 1.6 and 1.7 respectively for a film trailer ("First Descend")[4] rich in audio (music, narration, sound effects, machine sounds) and visual (objects, color, natural scenes, faces, action) events. The extracted 13 key-frames out of 512 of the original sequence (i.e., summarization percentage 2.5%) based on audiovisual saliency information, summarize the important visual scenes, some of which were selected based on the presence of important, aural attention-triggering audio events.

[4] http://www.firstdescentmovie.com

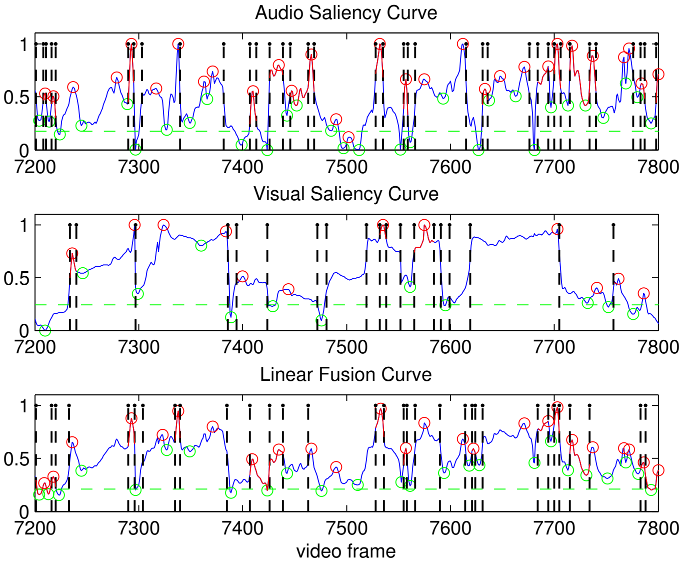


Fig. 1.8. Curves and detected features for audio saliency (top), video saliency (middle) and audiovisual saliency (bottom). The frame sequence was from the movie "Lord of the Rings I".

1.5.3 Automated Saliency-based Annotation

A method to derive automatic saliency-based annotation of audiovisual streams is by applying appropriate heuristically defined or learned thresholds on the audiovisual attention curves. The level sets of the curves thus define indicator functions of salient activity; see Fig. 1.9. A comparison against the available ground-truth is not a straight-forward task. On performing annotation, the human sensory system is able to almost automatically integrate and detect salient audiovisual information across many frames. Thus, such results are not directly comparable to the automatic annotation, since the audio part depends on the processing frame length and shift and the spatiotemporal nature of the visual part depends highly on the chosen frame neighborhood rather than on biological evidence.

Comparison against the ground-truth turns into a problem of tuning two different parameters, namely the extent (filter length) w of a smoothing operation and the threshold T that decides the salient versus the non-salient curve parts, and detecting the optimal point of operation. Perceptually, these two parameters are related, since a mildly smoothed curve (high peaks) should be accompanied by a high threshold, while a strongly smoothed curve (lower peaks) by a lower threshold. We relate these parameters using an exponential function

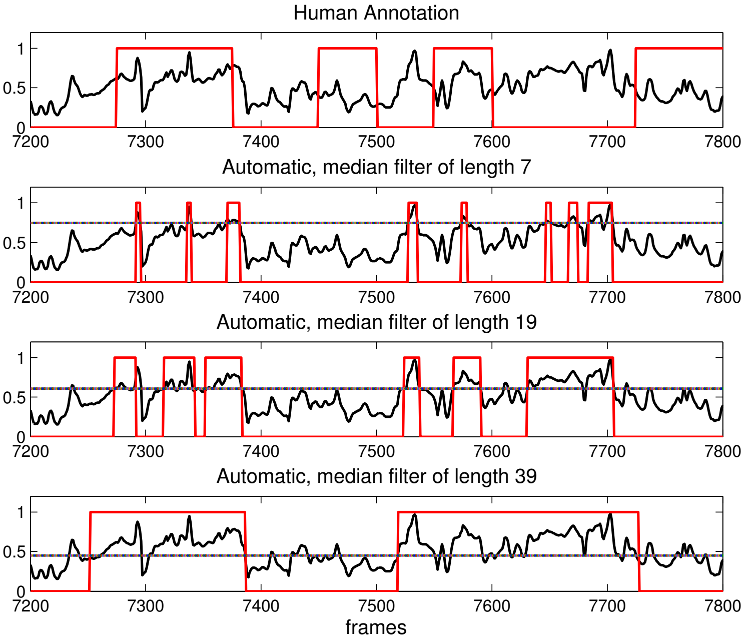


Fig. 1.9. Human and automated saliency-based annotations. Top row: Audiovisual saliency curve and manual annotation by inspection superimposed. Saliency indicator functions obtained with a median filter of variable size (7, 19, 39) in all other plots. The frame sequence was the same as in Fig. 1.8.

$$T(w) = \exp(-w/b), \tag{1.18}$$

where b is a scale factor, set to $b = 0.5$ in our experiments. Thus, a variable sized median filter is used for smoothing the audiovisual curve.

Fig. 1.9 shows a snapshot of the audiovisual curve for a sequence of 600 frames, the ground-truth, and the corresponding indicator functions and threshold levels computed by (1.18) for three different median filter lengths. We derive a precision/recall value for each filter length as shown in Fig. 1.10 for the whole duration of the two reference movie clips. Values on the horizontal x- axis relate to the size of the filter. As expected, the recall value is continuously increasing, since the thresholded, smoothed audiovisual curve tends to include an ever bigger part of the ground-truth. As already mentioned, the ability of the human eye to integrate information across time makes direct comparisons difficult. The varying smoothness imposed by the median filter simulates this integration ability in order to provide a more fair comparison. Note that although all presented experiments with the audiovisual saliency curve used the linear scheme for combining the audio features, prior to audio-

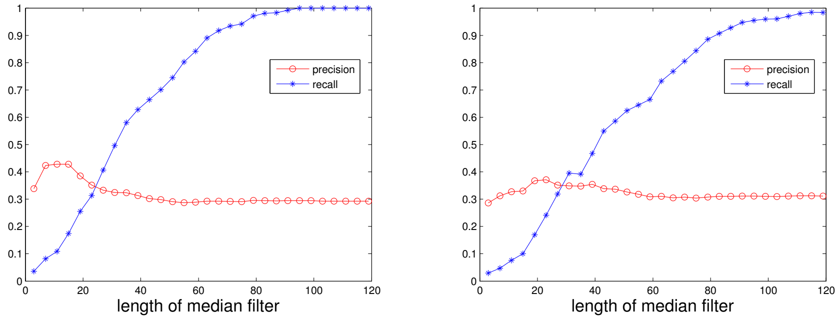


Fig. 1.10. Precision/Recall plots using human ground truth labeling on two film video segments. Left: “300”, right: “Lord of the Rings I”

visual integration, preliminary experiments with the non-linear fusion scheme (1.6) for the audio saliency yielded similar performance in the precision/recall framework.

1.5.4 Video Summarization

The dynamic summarization of video sequences involves reducing the content of the initial video using a seamless selection of audio and video subclips. The selection here is based on the attentional importance given by the associated audiovisual saliency curve. In order for the resulting summary to be perceptible, informative and enjoyable by the user, the video subsegments should follow a smooth transition, the associated audio clips should not be truncated and important audiovisual events should be included. One approach to creating summaries is to select, based on a user- or application- defined skimming index, portions of video around the previously detected key frames and align the corresponding “audio sentences” [21].

Here, summaries were created using a predefined *skimming percentage* c. In effect, a smoother attention curve is created using median filtering from the initial audiovisual saliency curve, since information from key-frames or saliency boundaries is not necessary. A saliency threshold T_c is selected so that the required *percent of summarization* c is achieved. Frames m with audiovisual saliency value $S_{\mathrm{av}}[m] > T_c$ 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 $D = \{m : S_{\mathrm{av}}[m] > T_c\}$ is 20% of the total number of frames. The result from this leveling step is a video frame indicator function I_c for the desired level of summarization c. The indicator function 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. The total effect is equivalent to 1D morphological filtering operations on the binary indicator function, where the filter's length is related to the minimum number of allowed frames in a skim and the distance between skims that are to be merged.

The movie summaries, obtained by skimming 2, 3 and 5 times faster than real time, were subjectively evaluated in terms of informativeness and enjoyability by 10 naive subjects. Preliminary average results indicate that the summaries obtained by the above procedure are well informative and enjoyable. However, more work is needed to improve the "smoothness" of the summary to improve the quality and enjoyability of the created skims.

1.6 Conclusions

In this chapter we have presented efficient audio and image processing algorithms to compute audio and visual saliency curves, respectively, from the aural and visual streams of videos and explored the potential of their integration for summarization and saliency-based annotation. The involved audio and image saliency detection modules attempt to capture the perceptual human ability to automatically focus on salient events. A simple fusion scheme was employed to create audiovisual saliency curves that were applied to movie summarization (detecting static key-frames and create video skims). This revealed that successful video summaries can be formed using saliency-based models of perceptual attention. The selected key-frames described the shots or different scenes in a movie, while the formed skims were intelligible and enjoyable, when viewed by different users. In a task of saliency-based video annotation, the audiovisual saliency curve correlated adequately well with the decisions of human observers.

Future work involves mainly two directions, namely more sophisticated fusion methods and improved techniques to create video summarization. Fusion schemes should be explored both for intra-modality integration (audio to audio, video to video) to create the individual saliency curves and inter-modality integration for the audiovisual curve. Different techniques may proven to be appropriate for the audio and visual parts, like the non-linear audio saliency scheme described herein. To develop more efficient summarization schemes, attention should be paid to the effective segmentation and selection of the video frames, aligned with the flow of audio sentences like dialogues or music parts. Summaries can be enhanced by including other cues besides saliency, related to semantic video content.

References

1. B. Adams et al., “IBM research TREC-2002 video retrieval system,” in *Proc. Text Retrieval Conference*, 2002.
2. Y. Avrithis, A. Doulamis, N. Doulamis, and S. Kollias, “Summarization of videotaped presentations: automatic analysis of motion and gesture,” *Computer Vision and Image Understanding*, vol. 75, no. 12, pp. 3–24, 1998.
3. 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.
4. J. J. Clark and A. L. Yuille, *Data Fusion for Sensory Information Processing.* Kluwer Academic Publ., 1990.
5. D. Dimitriadis, P. Maragos, and A. Potamianos, “Auditory Teager energy cepstrum coefficients for robust speech recognition,” in *Proc. Int’l Conf. on Speech Communication and Technology*, Lisboa, Portugal, Sep 2005.
6. A. Doulamis, N. Doulamis, Y. Avrithis, and S. Kollias, “A fuzzy video content representation for video summarization and content-based,” *Signal Processing*, vol. 80, no. 6, pp. 1049–1067, Jun 2000.
7. G. Evangelopoulos and P. Maragos, “Multiband modulation energy tracking for noisy speech detection,” *IEEE Trans. Audio, Speech, and Language Processing*, vol. 14, no. 6, pp. 2024–2038, Nov 2006.
8. W. T. Freeman and E. Adelson, “The design and use of steerable filters,” *IEEE Trans. Pattern Anal. Mach. Intell.*, no. 9, pp. 891–906, 1991.
9. D. Gabor, “Theory of communication,” *Journal Inst. of Elec. Eng. London*, vol. 93, no. III, pp. 429–457, 1946.
10. A. Girgensohn, J. Boreczky, and L. Wilcox, “Keyframe-based user interfaces for digital video,” *IEEE Computer*, vol. 34, no. 9, pp. 61–67, Sep 2001.
11. A. Hauptmann, “Lessons for the future from a decade of Informedia video analysis research,” in *Proc. ACM Int’l Conference on Image and Video Retrieval*, vol. 3568, 2005, pp. 1–10.
12. A. Hauptmann, R. Yan, T. Ng, W. Lin, R. Jin, D. Christel, M. Chen, and R. Baron, “Video classification and retrieval with the Informedia digital video library system,” in *Proc. Text Retrieval Conference*, Gaithersburg, MD, USA, Nov 2002.

13. S. X. Ju, M. J. Black, S. Minneman, and D. Kimber, “Summarization of videotaped presentations: automatic analysis of motion and gesture,” *IEEE Trans. Circuits Syst. Video Technol.*, vol. 8, no. 5, pp. 686–696, 1998.
14. J. Kaiser, “On a simple algorithm to calculate the ‘energy’ of a signal,” in *Proc. IEEE Int’l Conf. Acous., Speech, and Signal Processing*, Albuquerque N.M., Apr 1990, pp. 381–384.
15. ——, “Construction and evaluation of a robust multifeature speech/music discriminator,” in *Proc. IEEE Int’l Conf. Acous., Speech, and Signal Processing*, 1997, pp. 1331–1334.
16. E. Kandel, J. Schwartz, and T. Jessell, *Principles of Neural Science.* Stamford, Connecticut: McGraw-Hill, 4 edition, 2000.
17. 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.
18. 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.
19. L. Lu, H. Zhang, and H. Jiang, “Content analysis for audio classification and segmentation,” *IEEE Trans. Speech Audio Process.*, vol. 10, no. 7, pp. 504–516, 2002.
20. Y. Ma, L. Lu, H. Zhang, and M. Li, “A user attention model for video summarization,” in *Proc. ACM Int’l Conference on Multimedia*, 2002.
21. Y.-F. Ma, X.-S. Hua, L. Lu, and H.-J. Zhang, “A generic framework of user attention model and its application in video summarization,” *IEEE Trans. Multimedia*, vol. 7, pp. 907–919, 2005.
22. 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.
23. S. McAdams, “Recognition of auditory sound sources and events,” in *Thinking in Sound:The Cognitive Psychology of Human Audition.* Oxford University Press, 1993.
24. G. Monaci and P. Vandergheynst, “Audiovisual gestalts,” in *Proc. IEEE Int’l Conf. on Computer Vision and Pattern Recognition Workshop.* New York, NY: IEEE Computer Society, 2006, p. 200.
25. A. Potamianos and P. Maragos, “Speech formant frequency and bandwidth tracking using multiband energy demodulation,” *J. of the Acous. Soc. Am.*, vol. 99, no. 6, pp. 3795–3806, Jun 1996.
26. S. Raaijmakers, J. Den Hartog, and J. Baan, “Multimodal topic segmentation and classification of news video,” in *Proc. Text Retrieval Conference*, vol. 2, 2002, pp. 33–36.
27. L. Rabiner and B.-H. Juang, *Fundamentals of Speech Recognition.* NJ, USA: Prentice-Hall, 1993.
28. K. Rapantzikos, N. Tsapatsoulis, Y. Avrithis, and S. Kollias, *Signal Processing: Image Communication*, 2007, submitted for publication.
29. K. Rapantzikos and M. Zervakis, “Robust optical flow estimation in MPEG sequences,” in *Proc. IEEE Int’l Conf. Acous., Speech, and Signal Processing*, Mar 2005.
30. K. Ratakonda, M. Sezan, and R. Crinon, “Hierarchical video summarization,” in *Proc. SPIE, Visual Comm. and Image Proc.*, vol. 3653, Dec 1998, pp. 1531–1541.

31. M. Rautiainen et al., “TREC 2002 video track experiments at MediaTeam Oulu and VTT,” in *Proc. Text Retrieval Conference*, 2002.
32. M. Smith and T. Kanade, “Video skimming and characterization through the combination of image and language understanding techniques,” in *Proc. IEEE Int’l Conf. on Computer Vision and Pattern Recognition*, 1997, p. 775.
33. X. Sun and M. Kankanhalli, “Video summarization using R-sequences,” *Real-time imaging*, vol. 6, no. 6, pp. 449–459, Dec 2000.
34. H. Teager and S. Teager, “Evidence of nonlinear sound production mechanisms in the vocal tract,” in *Speech Production and Speech Modelling.* Kluwer Academic, 1990, pp. 241–261.
35. A. Treisman and G. Gelade, “A feature integration theory of attention,” *Cognit. Psychology*, vol. 12, no. 1, pp. 97–136, 1980.
36. N. Tsingos, E. Gallo, and G. Drettakis, “Perceptual audio rendering of complex virtual environments,” in *Proc. ACM Int’l conference on Computer Graphics and Interactive Techniques*, 2004.
37. S. Uchihashi, J. Foote, A. Girgensohn, and J. Boreczky, “Video manga: generating semantically meaningful video summaries,” in *Proc. ACM Int’l Conference on Multimedia*, 1999, pp. 383–392.
38. L. Ying, S.-H. Lee, C.-H. Yeh, and C.-C. Kuo, “Techniques for movie content analysis and skimming,” in *IEEE Signal Process. Mag.*, vol. 23, no. 2, Mar 2006, pp. 79–89.
39. J. Zacks and B. Tversky, “Event structure in perception and conception,” *Psychological Bulletin*, no. 127, pp. 3–21, 2001.
40. Y. Zhuang, Y. Rui, T. Huang, and S. Mehrotra, “Adaptive key frame extraction using unsupervised clustering,” in *Proc. IEEE Int’l Conf. on Image Processing*, 1998, pp. 866–870.

Index