Content-based Video Parsing and Indexing based on Audio-Visual Interaction

Sofia Tsekeridou and Ioannis Pitas
Department of Informatics
Aristotle University of Thessaloniki
Box 451, 54006 Thessaloniki, Greece
e-mail: {*sofia,pitas*}*@zeus.csd.auth.gr*

Abstract

A content-based video parsing and indexing method is presented in this paper, which analyzes both information sources (auditory and visual) and accounts for their inter-relations and synergy to extract high-level semantic information. Both frame-based and object-based access to the visual information is employed. The aim of the method is to extract semantically meaningful video scenes and assign semantic label(s) to them. Due to the temporal nature of video, time has to be accounted for. Thus, time-constrained video representations and indices are generated. The current approach searches for specific types of content information relevant to the presence or absence of speakers or persons. Audio source parsing and indexing leads to the extraction of a speaker label mapping of the source over time. Video source parsing and indexing results in the extraction of a talking face shot mapping over time. Integration of the audio and visual mappings constrained by interaction rules leads to higher levels of video abstraction and even partial detection of its context.

Keywords

Content-based video indexing, video parsing, video retrieval, multimedia database, audio-visual interaction, speaker recognition, face detection, facial feature extraction, scene identification

I. Introduction

Content-based video parsing, indexing and retrieval have recently grown to active research topics due to the enormous amount of *unstructured* video data available nowadays, the spread of its use as a data source in many applications and the increasing difficulty in its manipulation and retrieval of the material of interest. Many application areas have emerged for such tasks such as multimedia publishing, distributed multimedia systems, video-on-demand (VOD) [1], news-on-demand, digital video libraries [2]-[4], production, entertainment, interactive video [5], broadcasting, education [6], surveillance systems, videotelephony and games. The need for content-based indexing and coding has also been foreseen by ISO/MPEG that has founded the basis for the definition of two new coding standards: MPEG-4 [7] and MPEG-7 [8]. MPEG-4 will specify a content-based audiovisual coding standard where data are understood as a composition of objects, separately coded, thus allowing the independent access and manipulation of each object, whereas MPEG-7 will specify a standardized description of various types of multimedia information, associated with the content itself, to allow its indexing and thus the fast and efficient search, filtering and retrieval of the material of interest to the user [9].

Content-based video parsing involves temporal segmentation of the video into elementary units (smallest video indexing units) and content extraction of those units based on visual and/or audio semantic primitives. Thus a structured organization of the video source based on content is achieved.

Content-based video indexing is the task of tagging semantically meaningful video units obtained during parsing to enable easy and efficient retrieval of the material of interest with respect to user queries on content.

Although content understanding might be an easy task for a human by perceiving the intra- and inter-relations of many sources of information available (auditory, visual, textual), it is proven to be a very complicated process for a computer due to its lack of perception under unconstrained environments and to the unstructured nature of raw video data. To make matters worse, no concrete definition of video content can be available for easy and efficient parsing and indexing, since content definition depends on the needs and desires of the person who searches the database. Manual video segmentation and indexing based on semantic features is a very tedious and time consuming process and becomes impractical when an immense amount of video data has to be handled (e.g. during the population of digital video libraries). In summary, content descriptors have to be found and automatic or semi-automatic methods should be used for the structured organization of video based on content. Due to the inability for a universal content definition to account for all cases, different content types have been addressed by specific applications (application-oriented indexing), e.g. topic discrimination [10], violence detection [11], [12], speaker identification [13]-[16], scene change detection [17]-[21], news indexing [22]-[24], object-of-interest location [25], [26], keyword spotting [4], [23], thus limiting the effectiveness of the indexing process. However, the combination of the different single application content description and indexing schemes under the umbrella of a universal video indexing method that will offer ways of content retrieval based on arbitrary user requests presents a solution to a more extended, flexible and efficient system. Limited efforts have been made to apply general descriptors and description schemes based on high level semantics and operating on a predefined set of content categories thus offering high level domain-independent video abstraction [27]-[29]. Such content-based indexing methods provide tools for queries based on high level semantics without providing examples to the system. Another approach to video parsing and indexing attempts to provide visual or iconic summaries to the user organized in a hierarchical manner in order to facilitate browsing [30]-[37]. Compact visual representations are created to serve this purpose such as: micons, video posters, mosaic representations. A third approach to video indexing aims at serving user queries by example [2], [31], [38], [39]. A visual query is submitted by the user and similarity measures are defined to retrieve relative information. Visual representations are based on single or a group of keyframes or key-objects extracted from the input source and indexed by low level quantitative static (color, shape, texture) or dynamic features (motion or geometry). In this case, existing retrieval techniques for image databases may be applied for video retrieval as well (keyframe manipulation only).

Existing parsing and indexing methods have mainly been applied to a single data source, audio or visual. Only few publications have addressed both sources. It is evident that, by extracting semantics from both sources, exploiting their inter-relations, using a knowledge-based model and imposing constraints leads to more efficient indexing and to higher levels of video abstraction. Even human perception is improved when both sources are available. To probe further, content extraction is a very difficult task to be performed automatically by image and signal processing tools. The common approach is to firstly extract low-level content-free features. Semantic features are then derived by knowledge-based models or a degree of user interaction and imposed constraints. Existing video parsing and indexing methods are distinguished to two categories with respect to access for feature extraction: a) shot-based and b) object-based. The first category uses features characterizing frames or shots whereas the second one uses object-based features. Shot-based indexing methods are capable of extracting more abstract content information. However, object-based ones can give answers to more specific user queries involving single objects in a scene. Consequently, combination of both categories allows multi-level video abstraction and, thus, more efficient indexing.

A content-based video parsing and indexing method is presented in this paper, which analyzes both information sources (auditory and visual) and exploits their inter-relations to extract high-level semantic information. It can be classified as a first approach indexing method, as it searches for specific content information based on the existence of faces or speakers in the identified shots. Both frame-based and object-based access to the visual information is employed. The video parsing and indexing process is performed off-line suitable for applications such as digital video library population. The aim of the method is to extract semantically meaningful video scenes and assign semantic label(s) to them with respect to certain speaker presence/absence or (talking) face shot presence/absence. Due to the temporal nature of video, time has to be accounted for. Thus, novel time-constrained video representations and indices are generated in the form of time-dependent semantically labeled mapping functions, which allow both sequential and random access to the video data. Audio source parsing and indexing leads to the extraction of a speaker label mapping of the source over time. Robust thresholds are set for endpoint detection leading to the extraction of words, phrases or entire sentences. Parallel modeling of speaker classes results in better recognition results. Novel recognition techniques are presented by introducing variable length segment-based distance metrics which exploit the results obtained by the endpoint detection task. Video source parsing and indexing results in the extraction of a talking face shot mapping over time. Shot boundary detection is enhanced by post-processing frame color differences located inside a temporal one-dimensional window. During the face localization process, a robust range of the HSV color space is used by the color segmentation

task for skin-colored region detection. Facial mirror-symmetry and biometric analogy principles are exploited during face detection and facial feature detection. The use of the newly-introduced weighted block matching approach for the accurate estimation of eye locations leads to remarkable results. This is partly due to the use of a weighting function which presents a simple model of the eye structure. Mouth template tracking errors and their shot-based variance evaluation aids in decisions on talking face shot presence. The semantic labels employed at this stage are: (a) for audio: speech, silence (or noise), speaker identity (initial user guidance in supplying training sets is required for this), (b) for video: shot, no face shot, face shot (containing a person), talking face shot (containing a talking person). Additional semantic labels can be inserted if further processing is performed. The interaction of audio and visual labels supplies the user with more abstract information and a degree of its context, e.g. person X talking and being present in scene Y (characterized by its duration) or a news story is shown in scene Z (characterized by its duration) while reporter W is telling the story without being present in the scene. Keyword spotting and speech transcriptions [3], [23] or closed-caption processing and use of OCR techniques have been used by other approaches but they are not addressed in the current work. The presented method is applicable in video sequences with shots of single faces in frontal views (although extension of the method to deal with more than one faces is straightforward), where silence presence in the audio source is usually an indication of shot break. Such sequences are news sequences, dialogues, talk shows, or even close-ups in movies with video structure similar to our method.

The outline of the paper is as follows. Section II presents all the processing methods involved in extracting semantic information from the audio source. Initially, low level audio processing is performed. Content-free feature vectors are derived and used subsequently by a classifier to derive speaker models based on predefined training data sets (instance of user interaction). Novel methods are presented for speaker recognition that lead to the final time-dependent audio mapping with respect to speaker identities (a first attempt can be found in [15]). In section III, the main tasks involved for visual parsing and indexing are presented. Initially, shot boundaries are estimated, faces are searched for in each shot and finally mouth regions are detected and tracked in each face shot. The purpose of this analysis is to obtain automatically a time-dependent visual mapping with respect to the existence of talking faces in shots. Integration of audio and visual mappings, in order to achieve more efficient content-based video indexing, and interaction rules are discussed in section IV. Certain principles and aspects involved in content-based retrieval with respect to the proposed indexing schemes are briefly outlined in section V. Experimental results obtained on two news sequences are reported in section VI and conclusions are drawn in section VII.

II. Audio Source Parsing and Indexing Module

A. Audio Analysis

Low-level audio analysis involves the tasks of speech-silence discrimination, voiced-unvoiced discrimination, linear predictive (LP) analysis and derivation of LP cepstral coefficients of the speech-voiced classified frames, which compose the low-level audio feature vectors (Figure 1). In the sequel, the term *audio frame* refers to the shortest in duration audio part used in short-time audio analysis, whereas the term *segment* refers to a group of a variable number of successive frames. Generally, grouping is performed on the basis of successive frames exhibiting a similar attribute, which in our case is their pre-classification as speech frames. Each task involved in the audio analysis process will be successively described in more detail.

A.1 Speech-Silence Discrimination

Initially, *silence detection* is performed to discard silence periods from further analysis. Silence frames are audio frames of background noise with a relatively low energy level and high zero crossing rate compared to voice segments. In fact, unvoiced speech frames attain similar characteristics to silence frames but usually exhibit higher energy levels than background noise. They are also located in the neighbourhood of voiced ones which exhibit high energy values and low zero crossing rates. Consequently, in order to distinguish speech from silence, the average magnitude (instead of energy) and zero crossing rate are exploited [40]. The short-time signal average magnitude M_t and zero crossing rate Z_t of an M-sample audio frame $x_t(n)$, $n = 0, .., M-1$ are defined by:

$$M_t = \sum_{k=0}^{M-1} |x_t(k)| \quad (1)$$

$$Z_t = \frac{1}{2M} \sum_{k=1}^{M} |sgn(x_t(k)) - sgn(x_t(k-1))| \quad (2)$$

$t = 0, .., N-1$, where N is the total number of audio frames and M is the number of samples in the audio frame. Non-overlapping audio frames of 10msec duration are employed in the speech-silence detection process. A convenient approach to robust speech-silence discrimination is endpoint detection [40], [41], which determines the beginning and end of words, phrases or sentences so that subsequent processing ignores the surrounding background noise. Average magnitude and zero crossing thresholds are chosen relative to background noise characteristics evaluated on an apriori known audio interval corresponding to background noise, its short-time average magnitude and zero crossing rate functions being $M_{t,n}$ and $Z_{t,n}$ respectively. The average magnitude thresholds used by the endpoint detection process are set equal to $M_{thr,up} = \frac{E[M_t] - M_{thr,low}}{2}$ and $M_{thr,low} = \max(M_{t,n})$. The purpose of setting an upper threshold is to increase the distinguishability between silence and unvoiced frames

bearing in mind that unvoiced frames are located in the neighbourhood of voiced ones. Search for boundaries is initialized from the points where magnitude values are at least equal to $M_{thr,up}$. The zero crossing rate threshold is set equal to $Z_{thr} = \max(Z_{t,n})$. Such threshold selection proves robust and endpoint detection is correctly performed. Boundaries of words, phrases or entire sentences are satisfactorily estimated, a useful outcome that is subsequently exploited in the speaker recognition phase. It is understandable that, in continuous speech (natural speaking style), it is difficult to extract the boundaries of single words. Instead, it is easier to extract the boundaries of phrases or entire sentences. In cases that no speech recognition follows, such endpoint detection does not limit the effectiveness of subsequent processing. At this stage, a first level of audio abstraction is achieved, namely time-dependent speech-silence mapping. Silence segments are a clue for possible scene change as will be discussed later on.

A.2 Voiced-Unvoiced Discrimination

Low-level feature vectors are evaluated on voiced frames only [42]. Thus *voiced-unvoiced discrimination* prior to feature extraction should be performed. The discrimination process is based on the fact that voiced sounds exhibit low frequency content whereas unvoiced sounds are characterized by significant high frequency content. Thus, the energy distributions of the short-time Fourier transform of the frame signal are evaluated in the lower and upper frequencies (the boundary in our case is set at 2kHz with a sampling rate of 22kHz of the input signal, denoting a frequency range of 0-11kHz). The value of their ratio, which lies in the range $[0, 1]$, determines the nature of the speech frame signal with respect to voicing. Values greater than 0.25 imply unvoiced sounds, which are ignored in subsequent processing, as can be seen in Figure 1.

A.3 Audio Feature Extraction

For *audio feature extraction* in voiced frames, the speech signal is initially pre-emphasized by an FIR filter with transfer function $H(z) = 1 - 0.95z^{-1}$. The main purpose of pre-emphasis is to flatten the speech signal spectrally, in order to increase the relative energy of its high-frequency spectrum [42]. Furthermore, pre-emphasis ensures prevention of numerical instability in the case when linear predictive (LP) analysis follows. Speech frames of 30msec duration each are extracted with an overlap of 20msec. Each frame is windowed by a Hamming window of size M in order to minimize the signal discontinuities at its beginning and end. LP analysis follows and the LP parameter set $\mathbf{a}$ is later on converted to the LP cepstral coefficient vector $\mathbf{c}$. The latter has proved to be a more robust parameter set for speaker recognition than the LP coefficients themselves. Given the LP parameter set of a voiced audio frame, $\mathbf{a} = \{\alpha_m, 1 \leq m \leq p\}$, ($p$: order of the LP model), the LP cepstral coefficients

$\mathbf{c} = \{c_m, 1 \leq m \leq p\}$ [41], [42] are defined by:

$$c_m = \alpha_m + \sum_{k=1}^{m-1} (\frac{k}{m}) c_k a_{m-k}, \qquad 1 \leq m \leq p \tag{3}$$

B. Speaker Classification

After the generation of the sequence of feature vectors on voiced frames, speaker modeling and recognition succeeds aiming at extracting a time-constrained speaker identity mapping as shown in Figure 2.

B.1 Speaker Modeling

Speaker modeling is performed by a learning vector quantization (LVQ3) classifier [43]. Speaker recognition based on vector quantization is a conventional and successful method even with a limited number of training vectors. All speaker codebooks are trained in parallel using training vectors from apriori known time periods in which a certain speaker speaks and misclassification cases for the training vectors are compensated for. Initialization of the codebooks, each composed by 32 codevectors, is performed by the LBG algorithm. The size of the codebook has been experimentally determined as the one leading to the best classification results (other sizes tested were 16 and 64). This selection has been made bearing in mind that only voiced feature vectors are considered and that the training periods are rather short. [42] indicates that a codebook size of 64 is adequate for enhanced recognition results when a significant amount of training vectors is available and all kinds of sounds are considered. Classification of a training vector $\mathbf{c}$ to a speaker codebook $\mathbf{m}_i$, $i = 1, .., N_{speakers}$, is performed judging from Mahalanobis codevector distance values $d(\mathbf{m}_{ik}, \mathbf{c})$, $k = 1, .., 32$:

$$d(\mathbf{m}_{ik}, \mathbf{c}) = (\mathbf{m}_{ik} - \mathbf{c})^T \mathbf{R}_i^{-1} (\mathbf{m}_{ik} - \mathbf{c}) \tag{4}$$

where $\mathbf{m}_{ik}$ represents the k-th codevector of the speaker codebook $\mathbf{m}_i$ and $\mathbf{R}_i$ denotes the global covariance matrix derived from the training vectors of $\mathbf{m}_i$. The codevectors of every class which exhibit the minimum codevector distances from the training feature vector $\mathbf{c}$ are selected. Finally, the training vector $\mathbf{c}$ is classified to that codevector $\mathbf{m}_{ik}$ of the speaker codebook $\mathbf{m}_i$, which produces the minimal codevector distance $d(\mathbf{m}_{ik}, \mathbf{c})$ among those produced by the preselected codevectors of all speaker codebooks.

In succession, correct classification and misclassification cases are appropriately used for speaker codebook updating and total distortion minimization. If a training vector $\mathbf{c}$, known to belong to the codevector $\mathbf{m}_{jk}$ of the speaker codebook $\mathbf{m}_j$, is classified at iteration t to the codevector $\mathbf{m}_{ik}$ of the speaker codebook $\mathbf{m}_i$ and the training vector $\mathbf{c}$ lies in a predefined "window", then updating of the

respective codevectors of both speaker classes at iteration $t+1$ is performed in the following way:

$$\begin{aligned} \mathbf{m}_{ik}(t+1) &= \mathbf{m}_{ik}(t) - \alpha(t)[\mathbf{c} - \mathbf{m}_{ik}(t)] \\ \mathbf{m}_{jk}(t+1) &= \mathbf{m}_{jk}(t) + \alpha(t)[\mathbf{c} - \mathbf{m}_{jk}(t)] \end{aligned} \tag{5}$$

A training vector $\mathbf{c}$ is assumed to lie in the "window" if $\frac{s_j - s_i}{s_j} > w$ [43], where w is a constant set equal to 0.25 and s_j and s_i are the Mahalanobis distances of vector $\mathbf{c}$ from the codevector of the codebook $\mathbf{m}_j$ that it belongs to and from the codevector of the codebook $\mathbf{m}_i$ that it has been classified to respectively. In the case that the training vector $\mathbf{c}$ is classified correctly to the codevector $\mathbf{m}_{jk}$ of the speaker codebook $\mathbf{m}_j$ then updating at iteration $t+1$ is performed by:

$$\mathbf{m}_{jk}(t+1) = \mathbf{m}_{jk}(t) + \epsilon\alpha(t)[\mathbf{c} - \mathbf{m}_{jk}(t)] \tag{6}$$

It is noted that updating is performed only to the appropriate codevector $\mathbf{m}_{jk}$. $\alpha(t)$ is a linearly decreasing function depending on the number of iterations which controls the learning rate. The constant ϵ is usually set equal to a small value (0.25 in our case). Training is performed iteratively until the overall average distortion falls below a small value or the maximum number of iterations has been reached.

B.2 Speaker Recognition

During the *speaker recognition* process, the minimal codevector distance $s_i(\mathbf{c}_t)$ between an input test vector $\mathbf{c}_t$ and the codevectors $\mathbf{m}_{ik}$ of the speaker codebook $\mathbf{m}_i$ is calculated by:

$$s_i(\mathbf{c}_t) = \min_{\mathbf{m}_{ik} \in \mathbf{m}_i} [d(\mathbf{m}_{ik}, \mathbf{c}_t)], k = 1, .., 32 \tag{7}$$

Index i spans the number of speaker classes. $d(\mathbf{m}_{ik}, \mathbf{c}_t)$ is evaluated by (4). The global covariance matrix $\mathbf{R}_i$ has been evaluated during the training process.

In our approach, instead of assigning each input test vector separately to a speaker class, variable length groups of input test vectors corresponding to entire speech segments are all together classified. Such an approach is adopted based on the fact that only one speaker is bound to be talking between two silence frames. Thus, the average distortion of the entire speech segment is evaluated by:

$$S_i = \frac{1}{T} \sum_{t=1}^{T} s_i(\mathbf{c}_t) \tag{8}$$

where T is the number of speech frames included in the current speech segment and $s_i(\mathbf{c}_t)$ are given by (7). The unknown speaker is identified as the reference speaker whose model gives the minimum average distortion:

$$\text{Speaker Label} = \arg \min_{1 \leq i \leq N_{Speakers}} S_i \tag{9}$$

Finally, the time interval that corresponds to the speech segment determined beforehand is assigned to the identified speaker. Thus, *a speaker label mapping over time* is made possible offering a second level of audio abstraction as can be seen in Figure 2. Information about the presence of a speaker in certain parts of the video source can thus be indexed and retrieved. Speaker alternation can also be monitored and retrieved.

III. Visual Source Parsing and Indexing Module

A. Visual Analysis

The visual analysis module is shown in Figure 3. The input $v(n)$ is the entire video sequence. Initially, shot boundary (or scene cut) detection and video segmentation into shots is performed. Its output is a time dependent function $S(t)$ mapping the different shot boundaries detected (first level of video abstraction, content-free visual representation). The term *shot* denotes a sequence of successive frames that corresponds to a single camera start and end session. The term *scene* refers to one or more successive shots combined together due to exhibiting the same semantically meaningful concept, e.g. a scene that addresses the same topic although many shots may be involved. Scene characterization is different when different content is searched for. The first frame of each detected shot is subsequently processed by a face detection algorithm to take decisions on the existence of faces in the respective shot. A second time-dependent mapping function $F(t)$ is thus obtained on identified face shots (second level of video abstraction, content-based visual representation). The first frames of face shots is further analyzed by a facial feature extraction module aiming at extracting mouth locations. The face detection and facial feature extraction approach is similar to the one presented in [44]. A mouth template is thus obtained and used for mouth tracking in the remaining frames of the current face shot. Template matching errors are estimated by this process. Post-processing of these errors leads to conclusions on the existence of talking faces in the shot, described by the time-dependent mapping function $M(t)$ (third level of video abstraction, content-based visual representation). The processing details involved in the fulfillment of the above-mentioned tasks are subsequently outlined.

A.1 Shot Boundary Detection

Shot boundary detection is performed by pair-wise comparison between successive frames. Greyscale or color difference metrics are quite sensitive to abrupt object or camera motion, when not in combination with appropriate thresholding that is adaptive to local variations caused by such changes in the visual content. Consequently, a color frame difference metric is calculated by the current approach and is appropriately thresholded in such a way that camera motion or simple edit effects, like zoom-ins/outs, are not mistaken for possible scene cuts. Color instead of greyscale information is considered

because it is thought to convey better discriminant capabilities. Pair-wise color differences, $FD(t)$, are defined by:

$$FD(t) = \frac{1}{3X_{dim} \times Y_{dim}} \sum_{\mathbf{x}} ||\mathbf{I}(\mathbf{x};t) - \mathbf{I}(\mathbf{x};t-1)||_1 \tag{10}$$

where $\mathbf{I}(\mathbf{x};t) = [I_r(\mathbf{x};t)I_g(\mathbf{x};t)I_b(\mathbf{x};t)]^T$ represents the vector-valued pixel intensity function composed of the three color components: $I_r(\mathbf{x};t)$, $I_g(\mathbf{x};t)$ and $I_b(\mathbf{x};t)$. By $||.||_1$ the L_1-vector norm metric is denoted. Vector $\mathbf{x} = (x, y)$ spans the spatial dimension of the sequence (each frame is of size $X_{dim} \times Y_{dim}$) whereas scalar t its temporal one. Frame differencing is computationally intensive but no limitations on processing time are imposed when the task is performed off-line. Such a metric has been selected for shot boundary detection because it can account for gradual camera motion (in combination with the thresholding task) or zoom-ins/outs and it alleviates the drawback of false break detections when color histogram differencing is used due to lightning changes, camera motion or edit effects.

In order to detect possible shot breaks, local color frame difference mean values on a one-dimensional temporal window W of size 3 are obtained ($FD_{m,t\in W} = E[FD(t)|_{t\in W}]$). Then, the current value of $FD(t)$ is examined to specify whether it is the maximum value of those inside the temporal window (possibility of a shot break at t). If this is the case and $FD(t)/FD_{m,t\in W} > 2$, a scene cut is declared at t. Such window-based thresholding offers the means of adaptive thresholding according to local content and proves flexible and efficient in gradual camera movements and simple edit effects as zoom-ins and outs (no false positives are reported in such cases and, thus, over-segmentation is avoided). Abrupt changes are directly recognized.

A.2 Face Shot Determination

In order to *detect and isolate face shots*, the first frame of every shot is processed by a *face localization algorithm*, which is based on the assumption that faces could be characterized by their skin-like color, their elliptical shape and the x/y ratio of their surface. Since skin-like colors are better represented in the HSV color space, color transformation from the RGB to the HSV color space is performed [44], [45]. Pixels are afterwards selected having skin-like colors by setting appropriate thresholds to hue, saturation and value values, specifically:

$$\begin{aligned} 0^0 \leq H \leq 25^0, \quad 335^0 \leq H \leq 360^0 \\ 0.2 \leq S \leq 0.6 \\ V \geq 0.4 \end{aligned} \tag{11}$$

This hue range corresponds to reddish colors. The saturation range in (12) ensures exclusion of pure red or very dark red colors. Both account for small variations in lighting conditions. Value thresholding

is introduced to discard dark colors (such as dark hair). This color segmentation process proves to be adequately robust even with complex background or slightly varying lightning conditions. Color segmentation is performed at a coarse frame resolution to create an initial map of all pixels having colors close to the ones corresponding to skin. Refinement of this map is successively attempted in order to eliminate spurious pixels.

The next task is to *detect different skin-colored objects* in the frame. Object determination is performed by an iterative merging algorithm which considers 4-neighbourhood connectivity and which attempts to fill the "holes" that possibly exist in a compound object (e.g. the regions of eyes that color segmentation has discarded). The merging algorithm starts from the left-most assigned pixel of the currently considered object and proceeds by assigning the next pixels of the same row to the current object, if they are connected with previously assigned to this object pixels. Start and end points are saved and search in the next row is initialized from these points onwards. The algorithm is performed iteratively until no more pixels connected to the current object are detected. The process is computationally efficient and requires only a few iterations.

After the different skin-colored objects have been extracted, *a decision is made on whether one of them represents a face*. The decision is based on the size of the object with respect to the size of the frame (very small objects, even faces, are of no significance) and (mainly) on the ratio of its width to height which must be within a certain range of the facial x/y ratio. After a face object is detected, all other objects are discarded from the initial map, leaving only the face region. This final map is called *face region map*.

A.3 Facial Analysis

After a face has been detected in the first frame of the current shot and a face map has been evaluated, *coarse facial feature extraction* is pursued by detecting only the eyes, mouth and chin aiming at locating the mouth region. Facial feature candidates are initially extracted by minima analysis of the vertical face profile estimated on the face region map only. Information from all three color channels is used, which proves to enhance the discrimination between the features. Minima locations in the y direction are extracted using the first and second derivatives of the vertical face profile. At this point, our aim is to roughly extract the y-position of the eyes. Eyes are searched for in the upper part of the face considering facial biometric analogies. Once a rough estimate of the y-position of the eyes is derived, *finer estimation of eye locations* is attempted by a weighted block matching approach based on facial mirror-symmetry principles in frontal or semi-frontal views. Specifically, initial blocks of size $b_x \times b_y$ relevant to the face object size are centered around the initial coarse eye locations (l_x, l_y) and (r_x, r_y) respectively from either side of the center between the eyes, set equal to the middle-x point of the face

object. Search regions of size $(2s_x + 1) \times (2s_y + 1)$ are subsequently defined around the initial block locations. Finer eye locations are determined by minimizing the following weighted block matching error, $E(dx, dy)$:

$$E(dx, dy) = \sum_{x=0}^{b_x} \sum_{y=0}^{b_y} \sum_{i \in \{r,g,b\}} [|W(x, y) I_i(l_x - \frac{b_x}{2} + x + dx, l_y - \frac{b_y}{2} + y + dy) + \quad (12)$$

$$W(b_x - x, y) I_i(r_x + \frac{b_x}{2} - x - dx, r_y - \frac{b_y}{2} + y + dy)|] \quad (13)$$

where $dx \in [-s_x, s_x]$, $dy \in [-s_y, s_y]$. The two-dimensional weighting function $W(x, y)$ employed in this process is defined by:

$$W(x, y) = \frac{\sin((x^2 + y^2)/c)}{(x^2 + y^2)/c} \quad (14)$$

attempting to emphasize the center regions, which in the case of eyes represent the iris, and inverse weight the surrounding regions, thus leading to a simple model of the eye structure. Factor c is set proportional to the eye block width.

Accurate estimation of eye locations leads to precise evaluation of the head rotation angle by calculating the angle between the eye block centers, i.e.

$$\theta = \arctan(D_y / D_x) \quad (15)$$

where D_x and D_y are the pixel distances between the eye block centers in the x- and y-directions, respectively.

Head rotation compensation and re-estimation of the vertical face profile follows in order to obtain a better estimation of the exact location of the mouth area. Eye positions and face center are both known in the new coordinate system (rotated accordingly). Minima analysis of the vertical face profile is again employed in order to estimate the mouth center y-location. For this reason, minima are searched for in the lower part of the face by selecting the ones lying in a predefined eye distance range with respect to the face size based on facial biometric analogies. The mouth center in the x direction is set equal to the x position of the face center, a valid assumption for vertically aligned faces. Mouth horizontal profiles are then extracted in the area of the mouth center to locate the edges of the mouth and thus be able to initialize a rectangular mouth template with respect to the mouth width. All feature locations are subsequently rotated back to the initial coordinate system. Thus exact location of the mouth region is estimated and a rectangular mouth template is extracted having dimensions proportional to the mouth width to be used for mouth template tracking along the remaining frames of the face shot. It is obvious that compensation for head rotation enhances the accuracy of facial feature extraction.

A.4 Mouth Template Tracking

A mouth template tracking process follows aiming at extracting information on whether the face detected in the current shot is talking or not. For this reason, a template matching approach is adopted. Mouth locations in successive frames are estimated by minimizing L_1 norms between the current frame and the reference template. Search regions are appropriately initialized to account for motion of the head besides the mouth movement. After mouth regions have been tracked throughout the entire face shot, mean square errors between successive tracked mouth regions are evaluated to further detect mouth movements. Such movement implies a talking face. Large error variance (per shot) denotes mouth movement. Obviously, errors in mouth region estimation would result in errors concerning the decisions for talking or non-talking faces. The output of the visual analysis module is a time-dependent mapping function $M(t)$ (Figure 3) containing information about detected talking or non-talking face shots.

IV. Audio-Visual Interaction

Multimodal interaction can serve two purposes, as shown in Figure 4:

- enhance the "content findings" of one source by using similar content knowledge extracted from the other source(s)
- offer a more detailed content description about the same video instances by combining the content descriptors (semantic labels) of all data sources based on interaction rules and coincidence concepts.

The refinement process in our case involves estimation of speaker presence likelihood in every face shot. Speaker recognition performed on the audio source is prone to recognition errors which can be limited by utilizing the related visual information, e.g. presence of talking people, as described subsequently. The presence likelihood of speaker SP_i in a face shot FS_k is evaluated by:

$$P(SP_i|FS_k) = \frac{M_i}{L_k} \tag{16}$$

where M_i is the number of speech frames assigned to speaker SP_i and L_k is the total number of speech frames in face shot FS_k. The speaker that exhibits the maximum presence likelihood is the winner. All speech frames in face shot FS_k are indexed with the winner's identity. This approach enables a quite accurate detection of transition points between speakers in different successive face shots and limits the recognition errors introduced by the audio source. Of course, refinement of the time dependent speaker label indexing function is performed only in face shots with talking faces which contain respective content (speaker / talking face). Refinement of the shot boundary detection can be performed based on the fact that silence usually accompanies them. Therefore, silence segments are

a strong indication of shot boundaries and sometimes of topic changes [17].

The interaction of audio and visual semantic labels (speech, silence, speaker identity, face presence, face absence, talking face presence) supplies the user with more detailed information and a degree of the underlying context, e.g. person X talking and being present in scene Y (characterized by its duration) or a news story is shown in scene Z (characterized by its duration) while reporter W is telling the story without being present in the scene. Separate audio and visual maps containing different content information extracted from the respective sources can be combined in different ways based on such interaction rules. Another example is the combination of successive shots into a single story unit by observing that the same speaker speaks throughout the duration of those shots (extraction of the entire video segment that corresponds to the speaker presence evaluated from the respective mapping function). Appropriate visual summary representations of such a video segment by one of the existing methods can further guide the user to the topic of the story unit. Furthermore, dialogue situations can be detected by setting appropriate interaction rules, such as the frequent alternation of the same speakers.

V. Content-based Retrieval Aspects

It is evident that the presented content-based video indexing scheme defines the semantically labeled time dependent mapping functions of both sources as content description schemes (scene-based indexing). Such time-dependent mappings can be modeled as time intervals (referring to one or more successive video shots) assigned with a semantic label (a text string) included in a preset menu available to the end user. Overlapping of time intervals conveys inter-relations between the content conveyed by the different semantic labels and an indication to context. Therefore, conventional relational database management systems can be used to store the indexing information along with pointers to the raw data or a collection of appropriately constructed visual summaries of the respective video segments. Inter-relations are accounted for by the logical operations supported by the offered search utilities of a database system to the user.

VI. Simulation Results

The content-based video parsing and indexing scheme has been tested on many news sequences. Two of them are shown as examples in this paper, of approximately 6min and 4min duration respectively. The story units from the second news sequence have been cut out leaving only the anchorperson units (two reporters are present) in order to evaluate the performance of the face localization and mouth tracking methods involved in the visual parsing process. In the first news sequence, five different speakers exist but the faces of only three of them appear (two reporters in a studio on a chroma-

keyed background and one reporter outdoors). Two of the speakers appear in different parts of the sequence. The video sequences have been digitized at a frame rate of 25fps at PAL resolution. The audio signal has been sampled at 22kHz and each sample is a 16bit signed integer. Synchronization of the audio and video sources was attained during recording. Evaluation of the different tasks involved in the audio and visual source parsing and indexing modules will be successively reported. Firstly, the performance of specific tasks in these modules which are not directly involved in (actually precede) the determination of the respective mapping function will be examined. The study will later on focus on the derivation of the audio, visual or audio-visual mapping functions in order to better exhibit the high degree of their inter-relations and the modes of their final interaction which leads to the desired indexing.

Initially, speech-silence discrimination exhibits a remarkable performance achieving to locate entire phrases or sentences. An example of such discrimination can be viewed in Figure 5. Positive values of the discrimination function represent speech detected regions whereas negative values denote silence intervals. Rare occasions of unvoiced speech parts being classified as silence frames have only been observed. There was no case of background noise being classified as speech. For the speaker classification task, training periods for every speaker present in each news sequence have been set to 30secs of continuous speech. The novel recognition method evaluating variable-length segment-based distances from speaker classes leads to remarkable recognition results for both news sequences exhibiting only 0.19% and 0.068% recognition error rates respectively. Face detection results for three different reporters in the considered news sequences shown in Figures 6(a), 7(a) and 8(a) can be viewed in Figures 6(b)-(e), 7(b)-(e) and 8(b)-(e) respectively. It can be seen that in all cases the face region is correctly detected. There is only a sole instance of failure to detect an existing face due to bad lightning conditions. Normalization methods to deal with this problem could be incorporated in the method. Facial feature extraction, head rotation compensation and mouth region estimation results are illustrated in Figures 6(f)-(k), 7(f)-(i) and 8(f)-(k). For the second reporter, the rotation angle has been estimated equal to 0^0. Consequently, no results on head rotation compensation or rotated face vertical profile are given. Eye locations are in all cases correctly detected, proving the efficiency of the weighted block matching approach adopted for accurate eye location estimation. Head rotation angles are, consequently, estimated correctly and mouth region templates correctly determined. In order to establish the satisfactory performance of the face localization and facial feature extraction method experiments have been performed in our previous work on a database of 37 faces [44]. As far as mouth template tracking is concerned, some examples are given in Figure 9. The illustrated mouth regions are several frames apart from each other. Although the approach does not account for head

rotation, tracking is well performed.

The extraction of the different source mapping functions and the performance of some of the tasks involved in the process will now be examined. The shot boundary detection method is initially studied leading to a shot-based representation of the visual source. In Figure 10(a) the first frame of all detected shots of the first news sequence are illustrated (the order is from top-left to bottom-right). It can be seen that successive shots are dissimilar and thus no over-segmentation has been performed. The following analysis will justify this performance and the elimination of under-segmentation cases as well. In Figure 10(b) the frame color difference function employed in the shot boundary detection task can be viewed. One may conclude that appropriate thresholding of this error function will lead to correct estimation of shot boundaries. Post-processing of this function by temporal windowing and thresholding in the way already discussed leads to the shot boundary mapping function illustrated in Figure 10(c). Abrupt changes are directly recognized as shot breaks. Significant frame difference variances inside a single shot due to camera movement or simple edit effects do not lead to over-segmentation. The method remains to be evaluated in the presence of edit effects such as dissolves or fades. After shot-based segmentation of the visual source, face shot determination follows by analyzing the first frame of each shot. The results obtained by this process have been already discussed. The method fails in detecting a single face due to bad lightning conditions. The next task is to determine which of these face shots include a talking face. Mouth template error functions per face shot are evaluated, the variance of which leads to conclusions on the existence of talking faces in the current face shot. The result of this process is the extraction of a talking face shot mapping function as the one shown in Figure 10(d). The method achieves to locate all shots containing talking faces with the exception of the one that was missed in the previous stage of face shot classification. Finally, the remarkable speaker recognition performance, achieved by one of the tasks involved in the audio parsing system, leads to the estimation of the speaker label mapping function (from the audio source only) shown in Figure 10(e). In this figure, silence segments have been removed for better evaluation of the outcome (silence cases are represented by a zero value of the mapping function). For simplicity reasons, speakers are labeled with numbers. In reality, speaker identity labels are actually stored and later on retrieved. For the first news sequence, the only failures of the method were: (a) its inability to handle an unvoiced part of speech because it was surrounded by silence segments and (b) one case of speaker recognition error due to the short length of the respective segment. The first failure is a case where silence detection errors lead to propagation of errors to further classifications. Both error instances are shown in the side plot of Figure 10(e). In this side plot, silence segment cases are also illustrated for better result assessment.

Finally, the integration of the single-source derived mapping functions based on audio-visual interaction remains to be evaluated. As can be seen in Figure 10(f), refinement of the speaker label mapping function extracted from the audio source (Figure 10(e)) by exploiting concurrent relative visual content, i.e. the talking face shot mapping function (Figure 10(d)), in the way discussed in section IV, leads to elimination of the existing recognition errors and convergence to the correct mapping function. This is better established by observing the side plots of Figures 10(e)-(g), which illustrate the error cases encountered when only the audio source was processed and how they were eliminated based on audio-visual interaction principles. The vertical dashed lines in Figure 10 serve the purpose of illustrating time correspondences and, consequently, content inter-relation capabilities, between the different mapping functions.

Respective results on the extraction of the single-source time-dependent mapping functions and their integration have been obtained for the second news sequence as well. The first frames of every detected shot in this sequence are shown in 11(a). The frame difference function and the shot boundary mapping function are illustrated in Figures 11(b) and 11(c) respectively. Again, shot boundaries are correctly estimated. The talking face shot mapping function can be viewed in Figure 11(d). It is seen that all shots are correctly recognized as talking face shots (no news scenes exist). There are two cases of two successive face shots showing the same anchorperson with different scaling factors. These initially came from different topic units in the original news sequence. The abrupt change in the scaling factor enables discrimination of both shots. Color histograms in this case would fail to recognize the shot break. Furthermore, camera zoom ins/outs would occur gradually, thus leading to no sharp peak in the frame difference function. The speaker label mapping function as extracted by processing the audio source only is illustrated in Figure 11(e). A sole speaker recognition error case is only observed, as can be seen in the side plot of the latter figure. Audio-visual integration of the single source mapping functions leads to more accurate results (Figure 11(f)), which converge to the correct ones (Figure 11(g)).

Although simulations have been performed on news sequences, the presented method could be further applied on dialogue scenes, talk shows, movie scenes where face close-ups occur or generally on video segments where talking face(s) are present. The presented method does not exploit the inherent temporal and spatial structure of news sequences as news parsing and indexing methods do [22]-[24], [32]. For example, neither apriori knowledge on the alternation between anchorperson shots and news shots nor the fact that the background in anchorperson shots remains constant and that the reporter is actually located in the view center are exploited during parsing. The current approach only assumes the existence of a single face in frontal views (the extension of the face detection method

to identify existence of multiple faces is straightforward) and direct relation between shot breaks and silence presence in audio. These assumptions are actually valid in the mentioned video input cases. It is obvious though that the proposed method cannot be used for other than the afore-mentioned types of generic scenes.

An overview of similar to the presented approach methods that employ audio-visual interaction to extract semantic information will be subsequently given. [23] proposes use of speech to text transcripts and keyword spotting from the audio source and key-image detection mainly of faces and people from the video source. They finally associate the different source content findings by use of dynamic programming based on time durations. Some processing tasks are performed manually in this approach. Another method uses the audio source to enhance scene change detection in video sequences by identifying audio segments that are either silence, speech, music or noise and by relying on the assumption that audio silence segments may occur jointly with scene changes [17]. More accurate estimation of shot boundaries is also addressed by [18] which employs audio subband analysis to shift accordingly shot boundaries detected from visual analysis based on audio findings. The main assumption made is that silence periods exist between different advertisements. All these methods use both data sources in order to enhance the results obtained by a single source or improve the abstraction level of content findings similar to the current method. They are also based on a degree of domain knowledge or pre-specified assumptions about the video structure. The novelty of the current approach lies in content extraction based on speaker/talking face presence/absence and identification (if possible). It provides automatic answers to queries based on person identities, presence or absence.

VII. Conclusions

Content-based video parsing and indexing systems offer a flexible and efficient tool for further video retrieval and browsing. When such tasks combine semantic information from different data sources (auditory, visual, textual) through multimodal interaction concepts multi-level video abstraction is more efficiently performed. Besides, due to the temporal nature of video sources, time dependent representations and indices should also be a supplement to other kinds of indexing. The presented content-based parsing and indexing method aims at extracting semantically meaningful video scenes and assign semantic label(s) to them with respect to specific content. The semantic labels employed are: speech, silence, speaker identity, no face shot, face shot, talking face shot. Other semantic labels can be added to the system if further processing tools are used. The temporal coincidence or inter-relations of semantic labels can be captured by audio-visual interaction tasks based on certain rules.

References

[1] S-F. Chang, A. Eleftheriadis, and D. Anastasiou, "Development of columbia's video on demand testbed", *Signal Processing: Image Communication, Elsevier*, vol. 8, no. 3, pp. 191–207, 1996.

[2] M. Flickner, H. Sawhney, W. Niblack, J. Ashley, H. Qian, B. Dom, M. Gorkani, J. Hafner, D. Lee, D. Petkovic, D. Steele, and P. Yanker, "Query by image and video content: the qbic system", *IEEE Computer Magazine*, vol. 28, no. 9, pp. 23–32, Sept. 1995.

[3] H.D. Wactlar, T. Kanade, M.A. Smith, and S.M. Stevens, "Intelligent access to digital video: Informedia project", *IEEE Computer Magazine*, vol. 29, no. 5, pp. 46–52, May 1996.

[4] W. Li, S. Gauch, and J. Gauch amd K.M. Pua, "Vision: A digital video library", in *Proc. of 1st ACM Int. Conf. on Digital Libraries*, 1996, pp. 19–27.

[5] D.E. Blahut, T.E. Nichols, W.M. Schell, G.A. Story, and E.S. Szurkowski, "Interactive television", *Proceedings of the IEEE*, vol. 83, no. 7, pp. 1071–1085, July 1995.

[6] C. Shih-Fu, A. Eleftheriadis, and R. McClintock, "Next-generation content representation, creation, and searching for new-media applications in education", *Proceedings of the IEEE*, vol. 86, no. 5, pp. 884–904, May 1998.

[7] T. Sikora, "The MPEG-4 video standard verification model", *IEEE Trans. on Circuits and Systems for Video Technology*, vol. 7, no. 1, pp. 19–31, Feb. 1997.

[8] M.I. Sezan and R.J. Qian, "MPEG-7 standardization activities", in *Proc. of 1998 IEEE Int. Conf. on Image Processing*, Chicago, Illinois, USA, 4-7 Oct. 1998, vol. 3, pp. 517–520.

[9] P. Correia and F. Pereira, "The role of analysis in content-based video coding and indexing", *Signal Processing, Elsevier*, vol. 66, no. 2, pp. 125–142, 1998.

[10] T. Imai, R. Schwartz, F. Kubala, and L. Nguyen, "Improved topic discrimination of broadcast news using a model of multiple simultaneous topics", in *Proc. of 1997 IEEE Int. Conf. on Acoustics, Systems and Signal Processing*, Munich, Germany, 21-24 April 1997, pp. 727–730.

[11] J. Nam, M. Alghoniemy, and A. Tewfik, "Audio-visual content-based violent scene characterization", in *Proc. of 1998 IEEE Int. Conf. on Image Processing*, Chicago, Illinois, USA, 4-7 Oct. 1998, vol. 1, pp. 353–357.

[12] N. Vasconelos and A. Lippman, "Towards semantically meaningful feature spaces for the characterization of video content", in *Proc. of 1997 IEEE Int. Conf. on Image Processing*, 26-29 Oct. 1997.

[13] J. Nam, A.E. Cetin, and A. Tewfik, "Speaker identification and video analysis for hierarchical video shot classification", in *Proc. of 1997 IEEE Int. Conf. on Image Processing*, 26-29 Oct. 1997.

[14] D. Roy and C. Malamud, "Speaker identification based text to audio alignment for an audio retrieval system", in *Proc. of 1997 IEEE Int. Conf. on Acoustics, Systems and Signal Processing*, Munich, Germany, 21-24 April 1997, vol. II, pp. 1099–1102.

[15] S. Tsekeridou and I. Pitas, "Speaker identification for audio indexing applications", in *Proc. of IEEE Int. Conf. on Telecommunications*, Halkidiki, Greece, 22-25 Jun. 1998, vol. I, pp. 236–240.

[16] S. Tsekeridou and I. Pitas, "Speaker dependent video indexing based on audio-visual interaction", in *Proc. of 1998 IEEE Int. Conf. on Image Processing*, Chicago, Illinois, USA, 4-7 Oct. 1998, vol. 1, pp. 358–362.

[17] C. Saraceno and R. Leonardi, "Audio as a support to scene change detection and characterization of video sequences", in *Proc. of 1997 IEEE Int. Conf. on Acoustics, Systems and Signal Processing*, Munich, Germany, 21-24 April 1997, vol. IV, pp. 2597–2600.

[18] J. Nam and A. Tewfik, "Combined audio and visual streams analysis for video sequence segmentation", in *Proc. of*

1997 IEEE Int. Conf. on Acoustics, Systems and Signal Processing, Munich, Germany, 21-24 April 1997, vol. IV, pp. 2665–2668.

[19] T. Shin, J-G. Kim, H. Lee, and J. Kim, "Hierarchical scene change detection in an MPEG-2 compressed video sequence", in *Proc. of 1998 IEEE Int. Symp. on Circuits and Systems*, Monterey, California, USA, 31 May-3 June 1998.

[20] A.M. Alattar, "Wipe scene change detector for segmenting uncompressed video sequences", in *Proc. of 1998 IEEE Int. Symp. on Circuits and Systems*, Monterey, California, USA, 31 May-3 June 1998.

[21] J. Feng, K-T. Lo, and H. Mehrpour, "Scene change detection algorithm for MPEG video sequence", in *Proc. of 1996 IEEE Int. Conf. on Image Processing*, Lausanne, Switzerland, 16-19 Sept. 1996, vol. II, pp. 821–824.

[22] C.Y. Low, Q. Tian, and H. Zhang, "An automatic news video parsing, indexing and browsing system", in *Proc. of 1996 ACM Int. Conf. on Multimedia*, 1996, pp. 425–426.

[23] Y. Nakamura and T. Kanade, "Semantic analysis for video contents extraction - spotting by association in news video", in *Proc. of 1997 ACM Int. Conf. on Multimedia*, 1997, pp. 393–401.

[24] A. Hanjalic, R.L. Lagendijk, and J. Biemond, "Template-based detection of anchorperson shots in news programs", in *Proc. of 1998 IEEE Int. Conf. on Image Processing*, Chicago, Illinois, USA, 4-7 Oct. 1998, vol. 3, pp. 148–152.

[25] E. Stringa and C.S. Regazzoni, "Content-based and real time detection from video sequences acquired by surveillance systems", in *Proc. of 1998 IEEE Int. Conf. on Image Processing*, Chicago, Illinois, USA, 4-7 Oct. 1998, vol. 3, pp. 138–142.

[26] N. Herodotou, K.N. Plataniotis, and A.N. Venetsanopoulos, "A color segmentation and classification scheme for facial image and video retrieval", in *Proc. of EUSIPCO'98*, Rhodes, Greece, 8-11 September 1998, vol. III, pp. 1721–1724.

[27] M.R. Naphade, T. Kristjansson, B. Frey, and T.S. Huang, "Probabilistic multimedia objects (multijects): A novel approach to video indexing and retrieval in multimedia systems", in *Proc. of 1998 IEEE Int. Conf. on Image Processing*, Chicago, Illinois, USA, 4-7 Oct. 1998, vol. 3, pp. 536–540.

[28] T-S. Chua and L-Q. Ruan, "A video retrieval and sequencing system", *ACM Trans. on Information Systems*, vol. 13, no. 4, pp. 373–407, 1995.

[29] A.M. Ferman, A.M. Tekalp, and R. Mehrotra, "Effective content representation for video", in *Proc. of 1998 IEEE Int. Conf. on Image Processing*, Chicago, Illinois, USA, 4-7 Oct. 1998, vol. 3, pp. 521–525.

[30] B. Furht, *Multimedia Systems and Techniques*, Kluwer Academic Publishers, 1996.

[31] B. Furht, *Multimedia Tools and Applications*, Kluwer Academic Publishers, 1996.

[32] S.W. Smoliar and H. Zhang, "Content based video indexing and retrieval", *IEEE Multimedia Magazine*, vol. 1, no. 2, pp. 62–72, 1994.

[33] M. Irani and P. Anandan, "Video indexing based on mosaic representations", *Proceedings of the IEEE*, vol. 86, no. 5, pp. 905–921, May 1998.

[34] M. Irani, P. Anandan, J. Bergen, R. Kumar, and S. Hsu, "Mosaic representations of video sequences and their applications", *Signal Processing: Image Communication, Elsevier*, vol. 8, no. 4, pp. 327–351, 1996.

[35] M.M. Yeung and B-L. Yeo, "Video visualization for compact presentation and fast browsing of pictorial content", *IEEE Trans. on Circuits and Systems for Video Technology*, vol. 7, no. 5, pp. 771–785, Oct. 1997.

[36] H.S. Sawhney and S. Ayer, "Compact representations of videos through dominant and multiple motion estimation", *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 18, no. 8, pp. 814–830, Aug. 1996.

[37] B. Guensel, A.M. Ferman, and M.A. Tekalp, "Temporal video segmentation using unsupervised clustering and semantic object tracking", *SPIE Journal of Electronic Imaging*, vol. 7, no. 3, pp. 592–604, 1998.

[38] B. Guensel and A.M. Tekalp, "Content-based video abstraction", in *Proc. of 1998 IEEE Int. Conf. on Image Processing*, Chicago, Illinois, USA, 4-7 Oct. 1998, vol. 3, pp. 128–132.

[39] Y. Zhuang, Y. Rui, T.S. Huang, and S. Mehrotra, "Adaptive key frame extraction using unspupervised clustering", in *Proc. of 1998 IEEE Int. Conf. on Image Processing*, Chicago, Illinois, USA, 4-7 Oct. 1998, vol. 1, pp. 866–870.

[40] L. Rabiner and R.W. Schafer, *Digital Processing of Speech Signals*, Englewood Cliffs, N.J.: Prentice Hall, 1978.

[41] J.R. Deller, J.G. Proakis, and J.H.L. Hansen, *Discrete-Time Processing of Speech Signals*, Upper Saddle River, N.J.: Prentice Hall, 1987.

[42] L. Rabiner and B.-H. Juang, *Fundamentals of Speech Recognition*, Englewood Cliffs, N.J.: Prentice Hall, 1993.

[43] T. Kohonen, "The self-organizing map", *Proc. of the IEEE*, vol. 78, no. 9, pp. 1464–1480, September 1990.

[44] S. Tsekeridou and I. Pitas, "Facial feature extraction in frontal views using biometric analogies", in *Proc. of IX Eur. Signal Processing Conf.*, Rhodes, Greece, 8-11 Sept. 1998, vol. I, pp. 315–318.

[45] K. Sobottka and I. Pitas, "A novel method for automatic face segmentation, facial feature extraction and tracking", *Image Communication, Elsevier*, vol. 12, no. 3, pp. 263–281, 1997.

List of Figures

1 Audio analysis system 23

2 Speaker classification system 23

3 Visual analysis system 23

4 Audio-Visual interaction system 23

5 An example of speech-silence discrimination. Positive values of the discrimination function indicate speech detected regions whereas negative values represent silence segments. . . . 24

6 (a) Original frame of face shot. Results from: (b) skin color segmentation, (c) refinement process, (d) skin-colored object detection (different objects are assigned different graylevels), (e) face object extraction (face map), (f) vertical color profile on initial face object, (g) eye localization, (h) weighting function employed for eye location estimation, (i) head rotation compensation (the estimated mouth center is indicated with a white cross), (j) vertical color profile on rotated face object, and (k) final mouth region detection (its center and edges are shown by black crosses). 24

7 (a) Original frame of face shot. Results from: (b) skin color segmentation, (c) refinement process, (d) skin-colored object detection (different objects are assigned different graylevels), (e) face object extraction (face map), (f) vertical color profile on initial face object, (g) eye localization, (h) weighting function employed for eye location estimation, (i) final mouth region detection (its center and edges are shown by black crosses). The head rotation angle is evaluated to 0^0 in this case. 25

8 (a) Original frame of face shot. Results from: (b) skin color segmentation, (c) refinement process, (d) skin-colored object detection (different objects are assigned different graylevels), (e) face object extraction (face map), (f) vertical color profile on initial face object, (g) eye localization, (h) weighting function employed for eye location estimation, (i) head rotation compensation (the estimated mouth center is indicated with a white cross), (j) vertical color profile on rotated face object, and (k) final mouth region detection (its center and edges are shown by black crosses). 26

9 Examples of mouth template tracking. The illustrated tracked mouth regions are several frames apart from each other. 26

10 (a) First frames of all the detected shots in the first news sequence. The order is from top-left to bottom right. (b) Frame difference function, (c) Shot boundary mapping, (d) Talking face shot mapping, Speaker label mapping as obtained by (e) the audio source solely and (f) after applying audio-visual interaction rules. The correct speaker label mapping function is illustrated for evaluation in (g). The error cases encountered when processing the audio source only are illustrated in the side plots. 27

11 (a) First frames of all the detected shots in the second news sequence. The order is from top-left to bottom right. (b) Frame difference function, (c) Shot boundary mapping, (d) Talking face shot mapping, Speaker label mapping as obtained by (e) the audio source solely and (f) after applying audio-visual interaction rules. The correct speaker label mapping function is illustrated for evaluation in (g). The sole error case encountered when processing the audio source only is illustrated in the side plots. 28

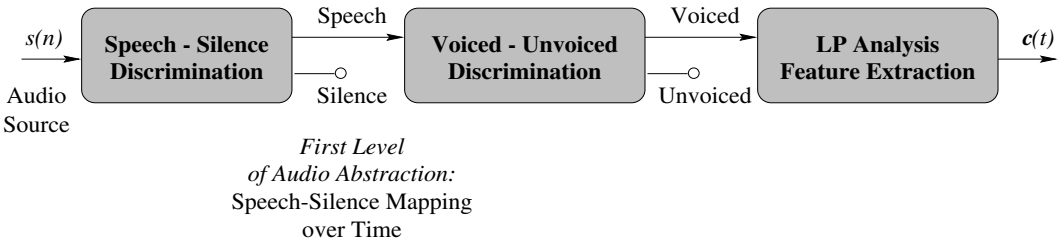


Fig. 1. Audio analysis system

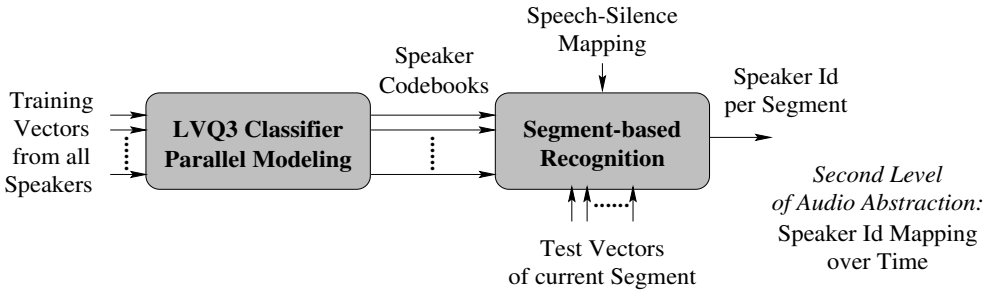


Fig. 2. Speaker classification system

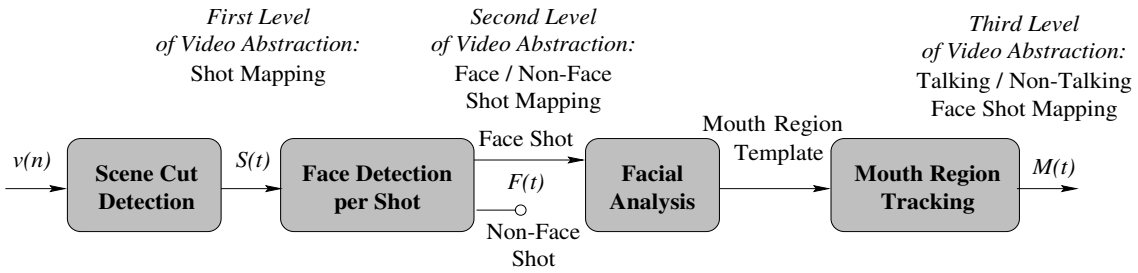


Fig. 3. Visual analysis system

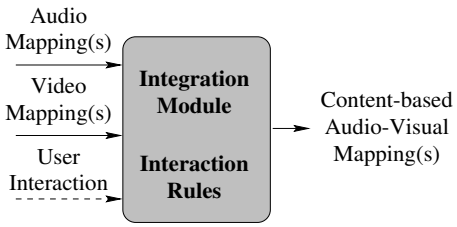


Fig. 4. Audio-Visual interaction system

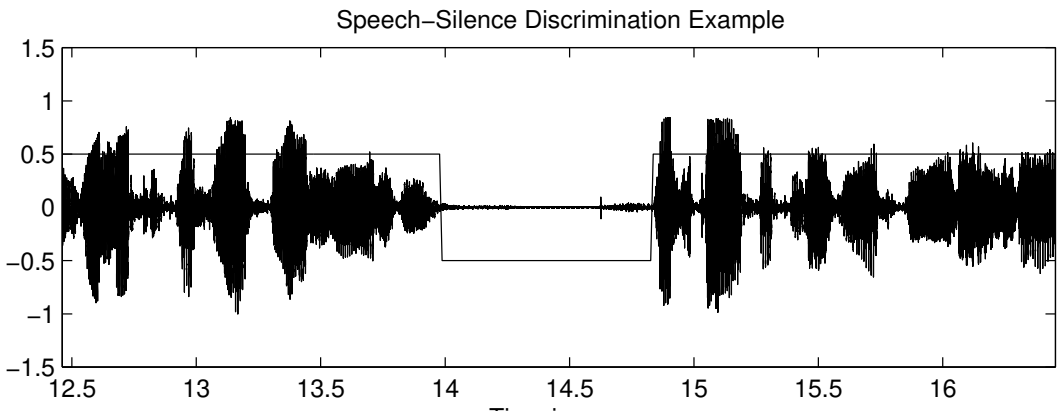


Fig. 5. An example of speech-silence discrimination. Positive values of the discrimination function indicate speech detected regions whereas negative values represent silence segments.

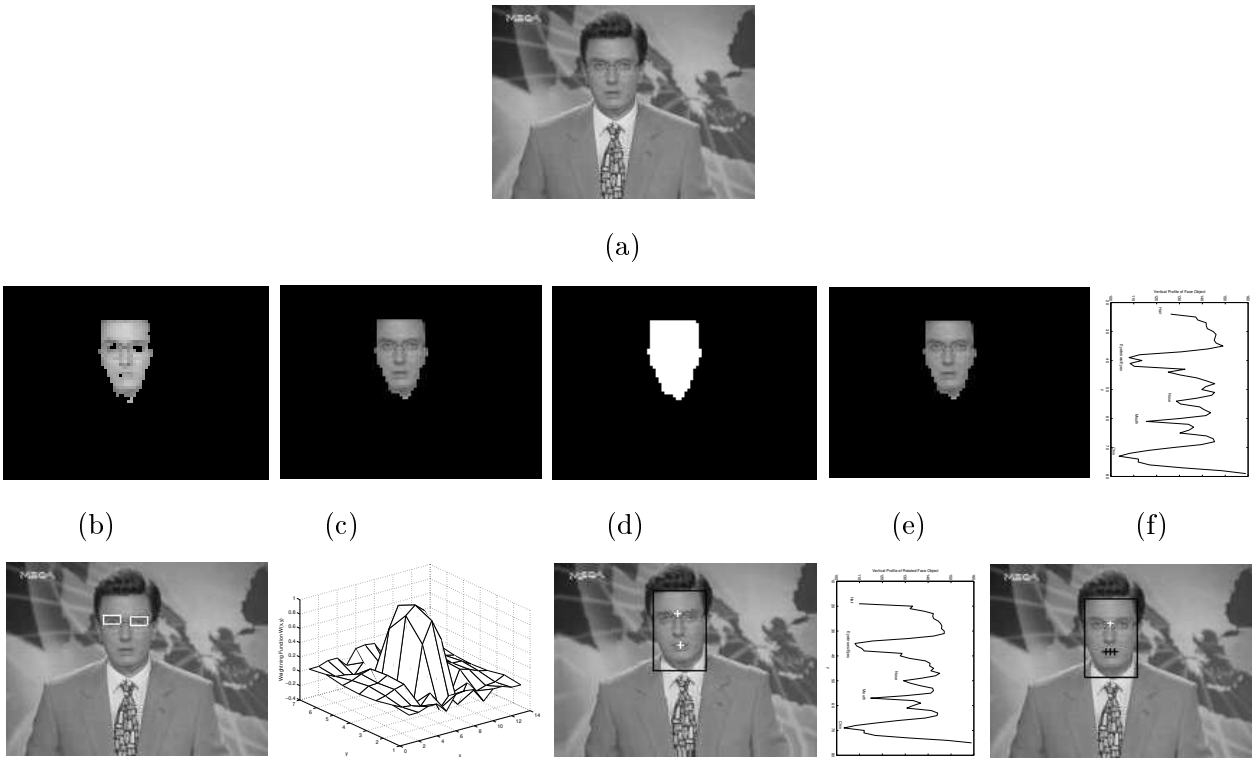


Fig. 6. (a) Original frame of face shot. Results from: (b) skin color segmentation, (c) refinement process, (d) skin-colored object detection (different objects are assigned different graylevels), (e) face object extraction (face map), (f) vertical color profile on initial face object, (g) eye localization, (h) weighting function employed for eye location estimation, (i) head rotation compensation (the estimated mouth center is indicated with a white cross), (j) vertical color profile on rotated face object, and (k) final mouth region detection (its center and edges are shown by black crosses).

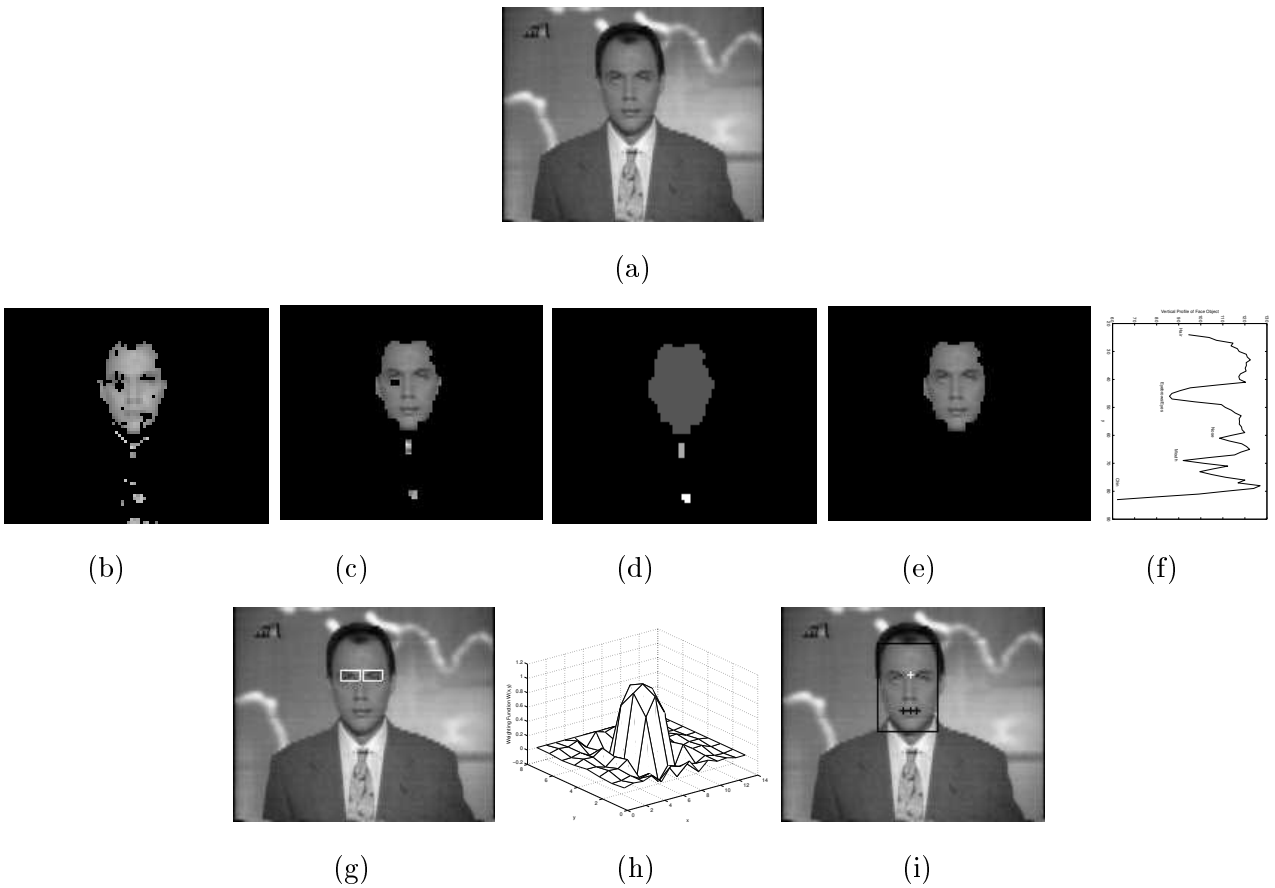
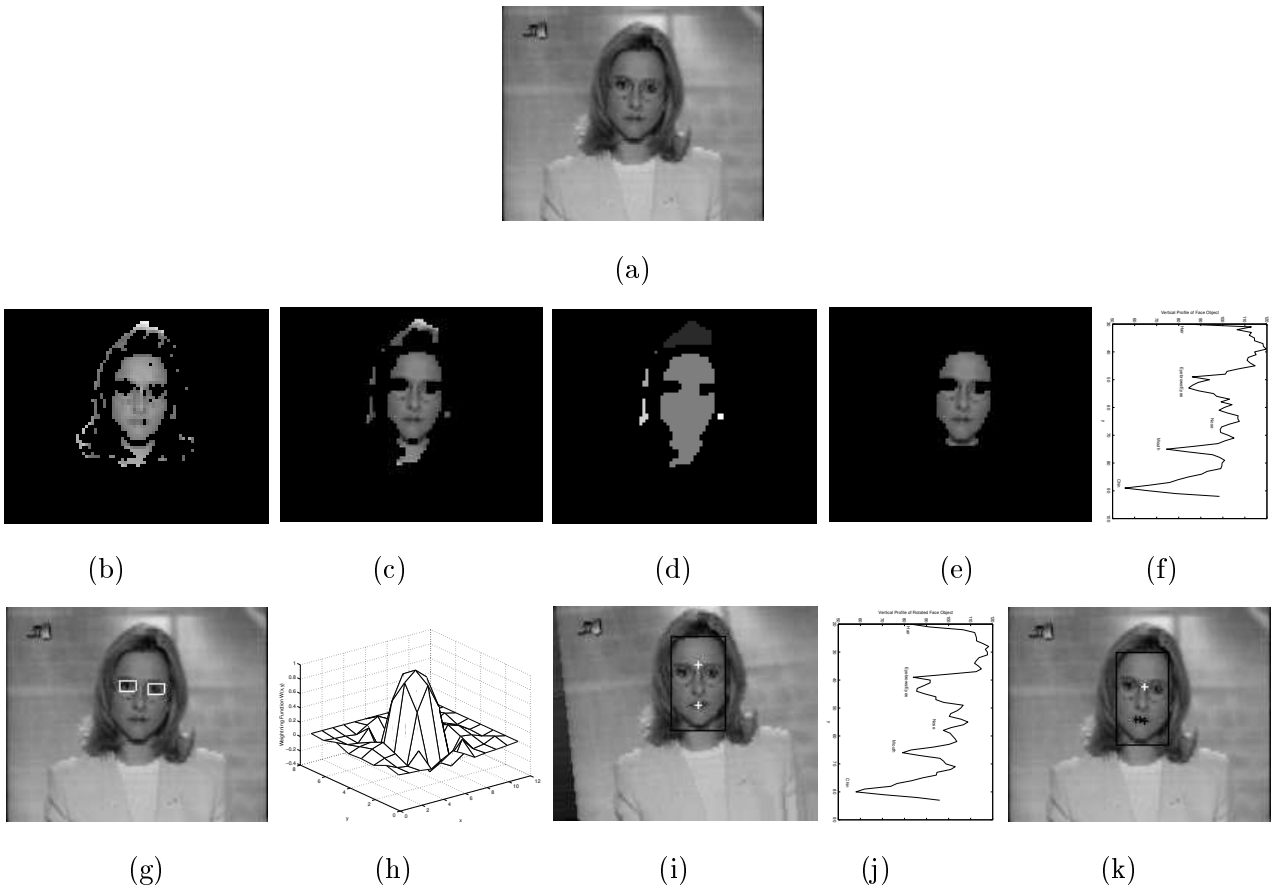


Fig. 7. (a) Original frame of face shot. Results from: (b) skin color segmentation, (c) refinement process, (d) skin-colored object detection (different objects are assigned different graylevels), (e) face object extraction (face map), (f) vertical color profile on initial face object, (g) eye localization, (h) weighting function employed for eye location estimation, (i) final mouth region detection (its center and edges are shown by black crosses). The head rotation angle is evaluated to 0^0 in this case.



(a) (b) (c) (d) (e) (f) (g) (h) (i) (j) (k)

Fig. 8. (a) Original frame of face shot. Results from: (b) skin color segmentation, (c) refinement process, (d) skin-colored object detection (different objects are assigned different graylevels), (e) face object extraction (face map), (f) vertical color profile on initial face object, (g) eye localization, (h) weighting function employed for eye location estimation, (i) head rotation compensation (the estimated mouth center is indicated with a white cross), (j) vertical color profile on rotated face object, and (k) final mouth region detection (its center and edges are shown by black crosses).

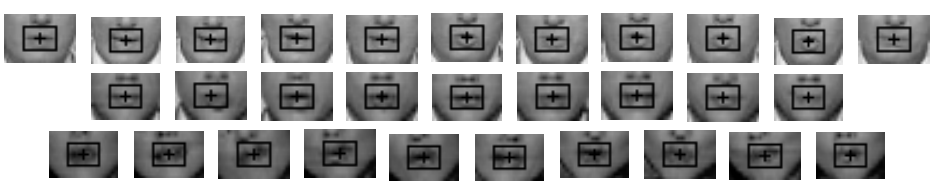


Fig. 9. Examples of mouth template tracking. The illustrated tracked mouth regions are several frames apart from each other.

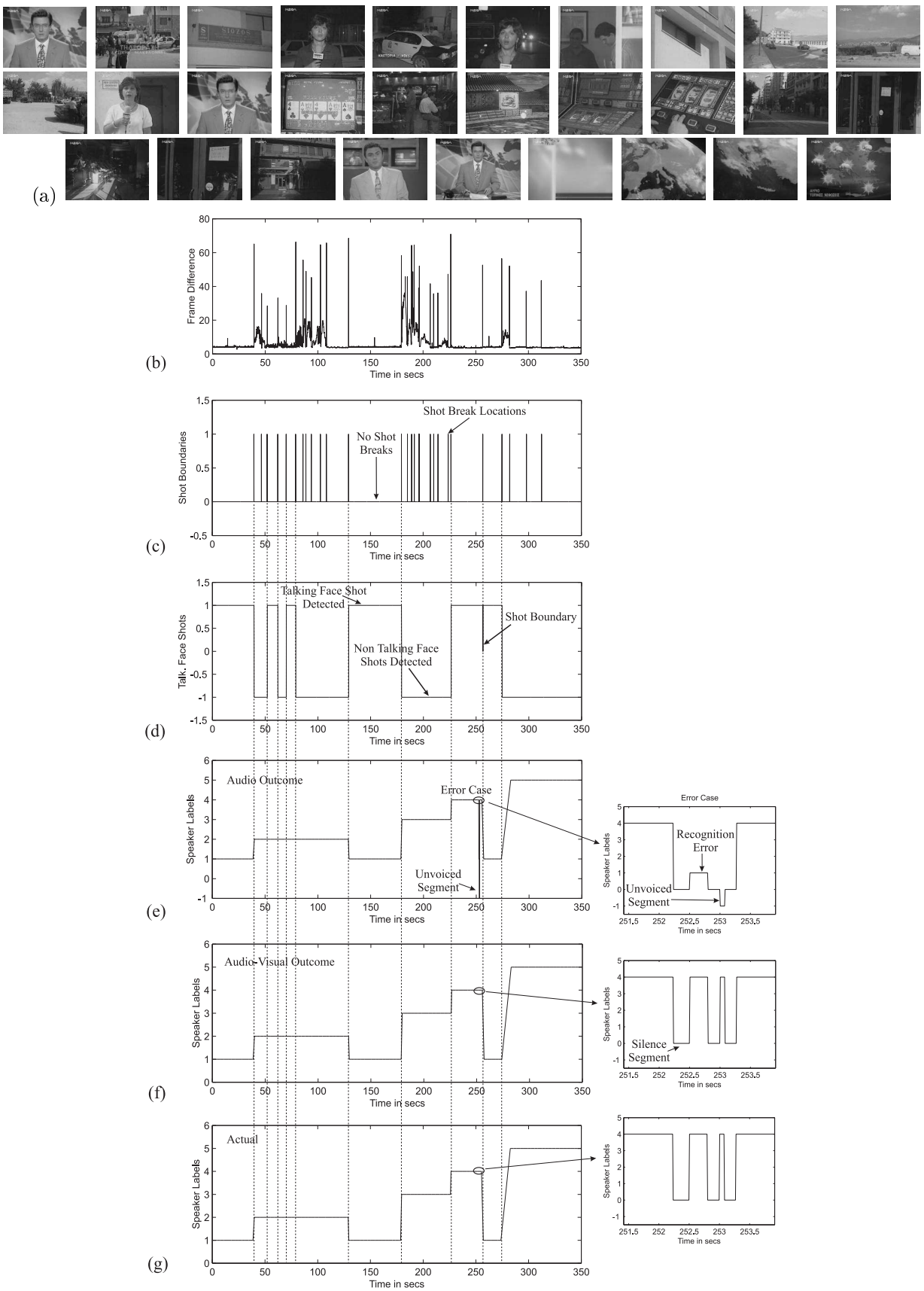


Fig. 10. (a) First frames of all the detected shots in the first news sequence. The order is from top-left to bottom right. (b) Frame difference function, (c) Shot boundary mapping, (d) Talking face shot mapping, Speaker label mapping as obtained by (e) the audio source solely and (f) after applying audio-visual interaction rules. The correct speaker label mapping function is illustrated for evaluation in (g). The error cases encountered when processing the audio source only are illustrated in the side plots.

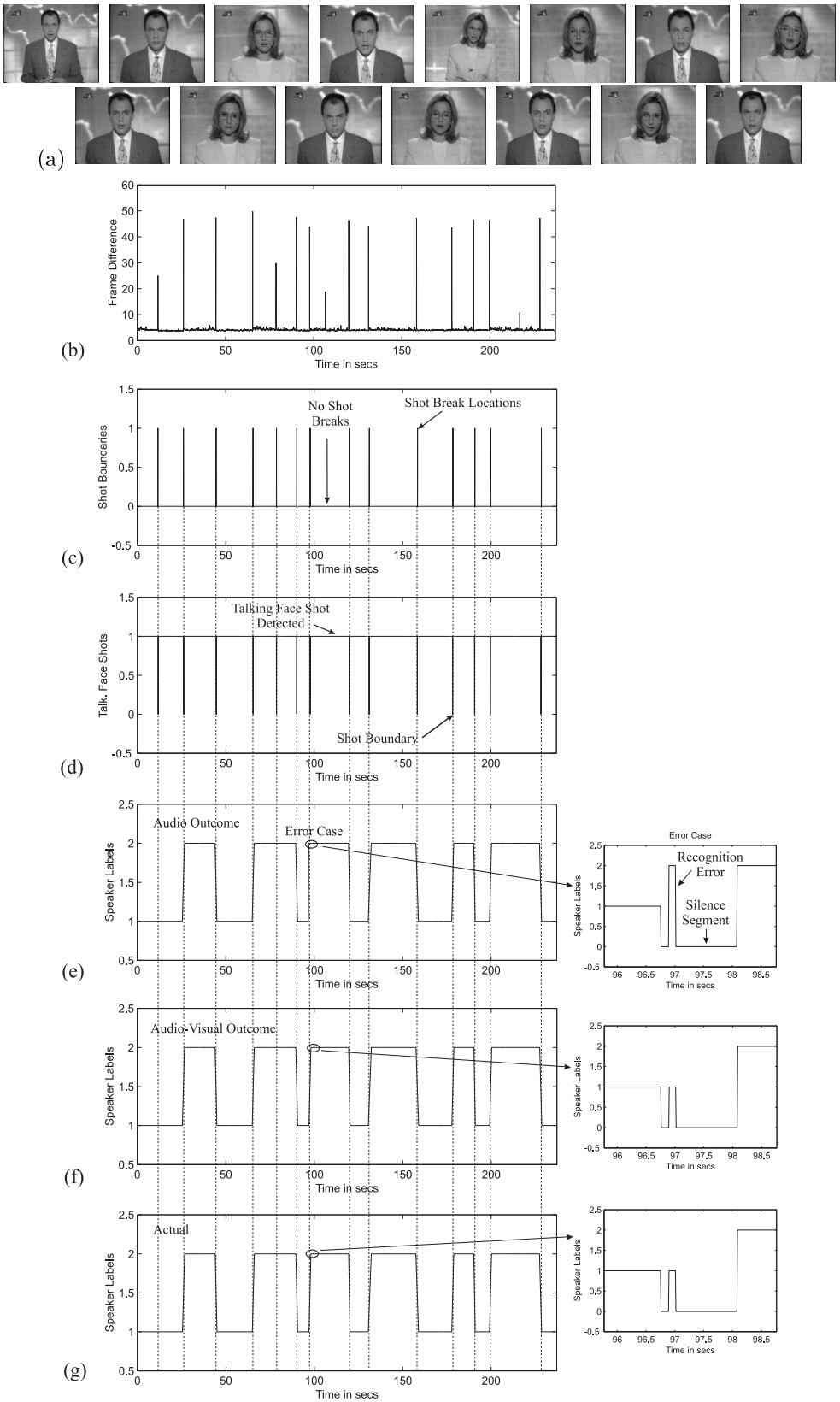


Fig. 11. (a) First frames of all the detected shots in the second news sequence. The order is from top-left to bottom right. (b) Frame difference function, (c) Shot boundary mapping, (d) Talking face shot mapping, Speaker label mapping as obtained by (e) the audio source solely and (f) after applying audio-visual interaction rules. The correct speaker label mapping function is illustrated for evaluation in (g). The sole error case encountered when processing the audio source only is illustrated in the side plots.