Multimedia Tools and Applications, 25, 5–35, 2005

Multimodal Video Indexing: A Review of the State-of-the-art

CEES G.M. SNOEK cgmsnoek@science.uva.nl
MARCEL WORRING worring@science.uva.nl
Intelligent Sensory Information Systems, Informatics Institute, University of Amsterdam, Kruislaan 403, 1098 SJ Amsterdam, The Netherlands

Abstract. Efficient and effective handling of video documents depends on the availability of indexes. Manual indexing is unfeasible for large video collections. In this paper we survey several methods aiming at automating this time and resource consuming process. Good reviews on single modality based video indexing have appeared in literature. Effective indexing, however, requires a multimodal approach in which either the most appropriate modality is selected or the different modalities are used in collaborative fashion. Therefore, instead of separately treating the different information sources involved, and their specific algorithms, we focus on the similarities and differences between the modalities. To that end we put forward a unifying and multimodal framework, which views a video document from the perspective of its author. This framework forms the guiding principle for identifying index types, for which automatic methods are found in literature. It furthermore forms the basis for categorizing these different methods.

Keywords: review, multimodal video indexing, video segmentation, multimodal integration, analysis framework

1. Introduction

For browsing, searching, and manipulating video documents, an index describing the video content is required. It forms the crux for applications like digital libraries storing multimedia data, or filtering systems [58] which automatically identify relevant video documents based on a user profile. To cater for these diverse applications, the indexes should be rich and as complete as possible.

Until now, construction of an index is mostly carried out by documentalists who manually assign a limited number of keywords to the video content. The specialist nature of the work makes manual indexing of video documents an expensive and time consuming task. Therefore, automatic classification of video content is necessary. This mechanism is referred to as video indexing and is defined as the process of automatically assigning content-based labels to video documents [30].

When assigning an index to a video document, three issues arise. The first is related to granularity and addresses the question: *what* to index, e.g., the entire document or single frames. The second issue is related to the modalities and their analysis and addresses the question: *how* to index, e.g., a statistical pattern classifier applied to the auditory content only. The third issue is related to the type of index one uses for labeling and addresses the question: *which* index, e.g., the names of the players in a soccer match, their time dependent position, or both.

Most solutions to video indexing address the *how* question with a unimodal approach, using the visual [16, 28, 63, 81, 84, 98, 102], auditory [22, 26, 47, 60, 61, 65, 92], or textual modality [12, 33, 103]. Good books [25, 32] and review papers [10, 13] on these techniques have appeared in literature. Instead of using one modality, multimodal video indexing strives to automatically classify (pieces of) a video document based on multimodal analysis. Only recently, approaches using combined multimodal analysis were reported [3, 5, 21, 35, 55, 66, 74] or commercially exploited, e.g., [17, 68, 89].

Ultimately the *which* question should be answered with content-based segment descriptors, for instance those proposed in the MPEG-7 standard [51, 52], that make a video document as accessible as a text document. However, the choice for an index is limited by the set of index terms for which automatic detectors can be realized.

One review of multimodal video indexing is presented in [90]. The authors focus on approaches and algorithms available for processing of auditory and visual information to answer the *how* and *what* question. We extend this by adding the textual modality, and by relating the *which* question to multimodal analysis. Moreover, we put forward a unifying and multimodal framework. Our work should therefore be seen as an extension to the work of [10, 13, 90]. Combined they form a complete overview of the field of multimodal video indexing.

The multimodal video indexing framework is defined in Section 2. We view a single video document from the perspective of its author, and discuss the different modalities and granularities involved in video indexing. This framework forms the basis for structuring the discussion on video document segmentation in Section 3. In Section 4 the role of conversion and integration in multimodal analysis is discussed. An overview of the index types that can be distinguished, together with some examples, will be given in Section 5. Finally, in Section 6 we end with a perspective on open research questions.

2. An author's perspective on video documents

In contrast to other frameworks, that view video documents from the (visual) data perspective, e.g., [2], we view a video document as a result of an authoring process. Consequence of this approach is that it allows for integration of different modalities more easily. To arrive at our framework for video indexing, we first consider video creation. In this survey we restrict ourselves to video made within a production environment, so excluding for example surveillance video. Video made within a production environment requires an author who conceives the idea for the video document and produces the final result, consisting of specific content and a layout. Therefore, we view a video document from an authors perspective.

An author uses visual, auditory, and textual channels to express his or her ideas. Hence, the content of a video is intrinsically multimodal. Let us make this more precise. In [57] multimodality is viewed from the system domain and is defined as "the capacity of a system to communicate with a user along different types of communication channels and to extract and convey meaning automatically". We extend this definition from the system domain to the video domain, by using an authors perspective as:

Definition 1 (Multimodality). The capacity of an author of the video document to express a predefined semantic idea, by combining a layout with a specific content, using at least two information channels.

We consider the following three information channels or modalities, within a video document:

- *Visual modality*: contains the *mise-en-scène*, i.e., everything, either naturally or artificially created, that can be seen in the video document;
- *Auditory modality*: contains the speech, music, and environmental sounds that can be heard in the video document;
- *Textual modality*: contains textual resources that describe the content of the video document;

For each of those modalities, Definition 1 naturally leads to a semantic perspective, a content perspective, and a layout perspective. We will now discuss each of the three perspectives involved. The important issue of combining modalities will be described later.

2.1. Semantic index

The first perspective expresses the intended semantic meaning of the author. Defined segments can have a different granularity, where granularity is defined as the descriptive coarseness of a meaningful unit of multimodal information [18]. To model this granularity, we define segments on five different levels within a semantic index hierarchy. The first three levels are related to the video document as a whole. The top level is based on the observation that an author creates a video with a certain purpose. We define:

- *Purpose*: set of video documents sharing similar intention;

The next two levels define segments based on consistent appearance of layout or content elements. We define:

- *Genre*: set of video documents sharing similar style;
- *Sub-genre*: a subset of a genre where the video documents share similar content;

The next level of our semantic index hierarchy is related to parts of the content, and is defined as:

- *Logical units*: a continuous part of a video document's content consisting of a set of named events or other logical units which together have a meaning;

Where named event is defined as:

- *Named events*: short segments which can be assigned a meaning that doesn't change in time;

Note that named events must have a non-zero temporal duration. A single image extracted from the video can have meaning, but this meaning will never be perceived by the viewer when this meaning is not consistent over a set of images.

At the first level of the semantic index hierarchy we defined purpose. According to [38], the purpose for which the video document is made is either entertainment, information, communication, or data analysis. Recall that we only consider video documents that are made within a production environment. Therefore, the purpose of data analysis is excluded. Genre examples range from feature films, news broadcasts, to commercials. This forms the second level. On the third level are the different sub-genres, which can be e.g., horror movie or ice hockey match. Examples of logical units, at the fourth level, are a dialogue in a drama movie, a first quarter in a basketball game, or a weather report in a news broadcast. Finally, at the lowest level, consisting of named events, examples can range from explosions in action movies, goals in soccer games, to a visualization of stock quotes in a financial news broadcast.

2.2. Content

The content perspective relates segments to elements that an author uses to create a video document. The following elements can be distinguished [9]:

- *Setting*: time and place in which the video's story takes place, can also emphasize atmosphere or mood;
- *Objects*: noticeable static or dynamic entities in the video document;
- *People*: human beings appearing in the video document;

Typically, setting is related to logical units. Objects and people are the main elements in named events. The appearance of the different content elements can be influenced by an author of the video document by using modality specific style elements. For the visual modality an author can use specific colors, lighting, camera angles, camera distance, and camera movement. Auditory style elements are the loudness, rhythmic, and musical properties. The textual appearance is determined by the style of writing and the phraseology. All these style elements contribute to expressing an author's intention.

2.3. Layout

The layout perspective considers the syntactic structure an author uses for the video document. In essence, the syntactic structure for each modality is a temporal sequence of *fundamental units*, which in itself do not have a temporal dimension. The nature of these units is the main factor discriminating the different modalities. The visual modality of a video document is a set of ordered images, or frames. So the fundamental units are the single image frames. Similarly, the auditory modality is a set of samples taken within a certain time span, resulting in audio samples as fundamental units. Individual characters form the fundamental units for the textual modality. Upon the fundamental units an aggregation is imposed, which is an artifact from creation. We refer to this aggregated fundamental

units as *sensor shots*, defined as a continuous sequence of fundamental units resulting from an uninterrupted sensor recording. For the visual and auditory modality this leads to:

- *Camera shots*: result of an uninterrupted recording of a camera;
- *Microphone shots*: result of an uninterrupted recording of a microphone;

For text, sensor recordings do not exist. In writing, uninterrupted textual expressions can be exposed on different granularity levels, e.g., word level or sentence level, therefore we define:

- *Text shots*: an uninterrupted textual expression;

Note that sensor shots are not necessarily aligned. Speech for example can continue while the camera switches to show the reaction of one of the actors. There are, however, situations where camera and microphone shots are recorded simultaneously, for example in live news broadcasts.

An author of the video document is also responsible for concatenating the different sensor shots into a coherent structured document by using *transition edits*. "He or she aims to guide our thoughts and emotional responses from one shot to another, so that the interrelationships of separate shots are clear, and the transitions between sensor shots are smooth" [9]. For the visual modality abrupt cuts, or gradual transitions,[1] like wipes, fades, or dissolves can be selected. This is important for visual continuity, but sound is also a valuable transitional device in video documents. Not only to relate shots, but also to make changes more fluid or natural. For the auditory transitions an author can have a smooth transition using music, or an abrupt change by using silence [9]. To indicate a transition in the textual modality, e.g., closed captions, an author typically uses ">>>", or different colors. They can be viewed as corresponding to abrupt cuts as their use is only to separate shots, not to connect them smoothly.

The final component of the layout are the optional visual or auditory *special effects*, used to enhance the impact of the modality, or to add meaning. Overlayed text, which is text that is added to video frames at production time, is also considered a special effect. It provides the viewer of the document with descriptive information about the content. Moreover, the size and spatial position of the text in the video frame indicate its importance to the viewer. "Whereas visual effects add descriptive information or stretch the viewer's imagination, audio effects add level of meaning and provide sensual and emotional stimuli that increase the range, depth, and intensity of our experience far beyond what can be achieved through visual means alone" [9]. Note that we don't consider artificially created content elements as special effects, as these are meant to mimic true settings, objects, or people.

Based on the discussion in this section we come to a unifying multimodal video indexing framework based on the perspective of an author. This framework is visualized in figure 1. It forms the basis for our discussion of state-of-the-art indexing techniques.

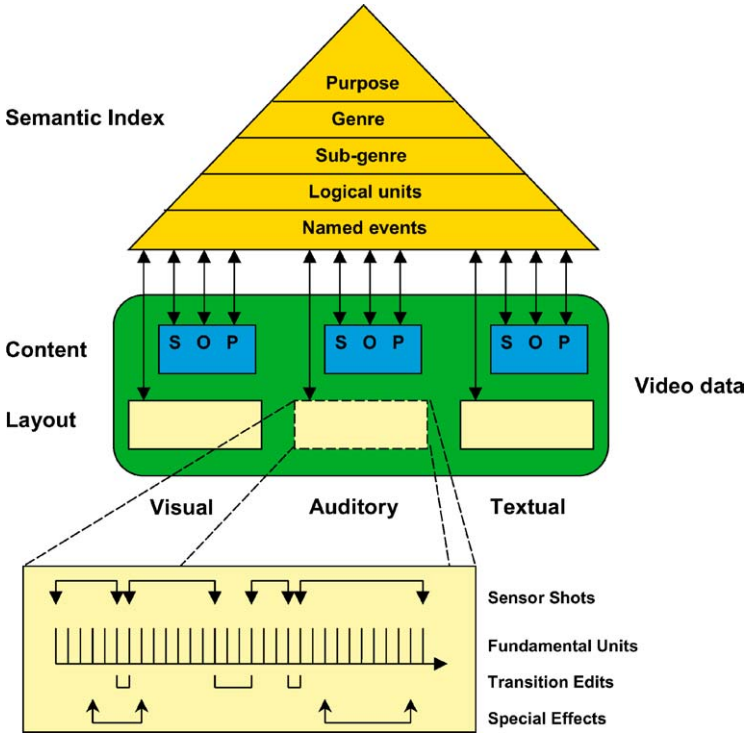

Figure 1. A unifying framework for multimodal video indexing based on an author's perspective. The letters S, O, P stand for setting, objects, and people. An example layout of the auditory modality is highlighted, the same holds for the others.

3. Video document segmentation

For analysis purposes the process of authoring should be reversed. To that end, first a segmentation should be made that decomposes a video document in its layout and content elements. Results can be used for indexing specific segments. In many cases segmentation can be viewed as a classification problem. In video indexing literature many heuristic methods are proposed. The more advanced techniques make explicit use of pattern recognition. Therefore, we will first discuss the different classification methods that are used in video indexing. Then, we will discuss reconstruction of the layout for each of the modalities. Finally, we will focus on segmentation of the content. The data flow necessary for analysis is visualized in figure 2.

3.1. Pattern recognition

In video indexing, patterns of interest need to be distinguished to make decisions about layout and content categories. These patterns can be, for example, sub images, samples,

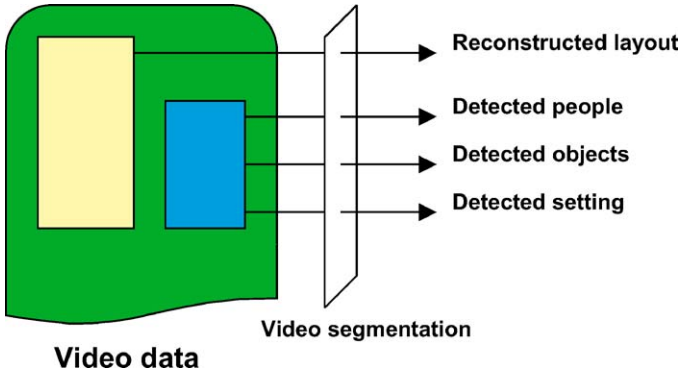

Figure 2. Data flow in unimodal video document segmentation.

or features derived from layout and content elements. According to [37] the four best approaches for pattern recognition are:

- *Template matching*: the pattern to be recognized is compared with a learned template, allowing changes in scale and pose;
- *Statistical classification*: the pattern to be recognized is classified based on the distribution of patterns in the space spanned by pattern features;
- *Syntactic or structural matching*: the pattern to be recognized is compared to a small set of learned primitives and grammatical rules for combining primitives;
- *Neural networks*: the pattern to be recognized is input to a network which has learned nonlinear input-output relationships;

Examples of those methods are found throughout our paper. The statistical approach is most frequently encountered in video indexing literature, especially the following four specific techniques:

- *Bayes classifier*: assigns a pattern to the class which has the maximum estimated posterior probability [37];
- *Decision tree*: assigns a pattern to a class based on a hierarchical division of feature space [37];
- *k-Nearest neighbor*: assigns a pattern to the majority class among the k patterns with smallest distance in feature space [37];
- *Hidden Markov model* (*HMM*): assigns a pattern to a class based on a sequential model of state and transition probabilities [46, 69];

Statistical classifiers are also well suited for multimodal classification. This aspect of pattern recognition will be highlighted in Section 4.2. We will now first discuss the reconstruction of layout and content elements.

3.2. *Layout reconstruction*

Layout reconstruction is the task of detecting the sensor shots and transition edits in the video data. For analysis purposes layout reconstruction is indispensable. Since the layout guides the spectator in experiencing the video document, it should also steer analysis.

For reconstruction of the visual layout, several techniques already exist to segment a video document on the camera shot level, known as *shot boundary detection*.[2] Various algorithms are proposed in video indexing literature to detect cuts in video documents, all of which rely on comparison of successive frames with some fixed or dynamic threshold on either pixel, edge, block, or frame level. Block level features can be derived from motion vectors, which can be computed directly from the visual channel, when coded in MPEG, saving decompression time. For an extensive overview of different cut detection methods we refer to the survey of Brunelli in [13] and the references therein.

Detection of transition edits in the visual modality can be done in several ways. Since the transition is gradual, comparison of successive frames is insufficient. The first researchers exploiting this observation where Zhang et al. [97]. They introduced the twin-comparison approach, using a dual threshold that accumulates significant differences to detect gradual transitions. For an extensive coverage of other methods we again refer to [13], we just summarize the methods mentioned. First, so called plateau detection uses every k-th frame. Another approach is based on effect modeling, where video production-based mathematical models are used to spot different edit effects using statistical classification. Finally, a third approach models the effect of a transition on intensity edges in subsequent frames.

Detection of abrupt cuts in the auditory layout can be achieved by detection of silences and transition points, i.e. locations where the category of the underlying signal changes. In literature different methods are proposed for their detection.

In [60] it is shown that average energy, E_n, is a sufficient measure for detecting silence segments. E_n is computed for a window, i.e., a set of n samples. If the average for all the windows in a segment are found lower than a threshold, a silence is marked. Another approach is taken in [99]. Here E_n is combined with the zero-crossing rate (ZCR), where a zero-crossing is said to occur if successive samples have different signs. A segment is classified as silence if E_n is consistently lower than a set of thresholds, or if most ZCRs are below a threshold. This method also includes unnoticeable noise.

Li et al. [43] use silence detection for separating the input audio segment into silence segments and signal segments. For the detection of silence periods they use a three-step procedure. First, raw boundaries between silence and signal are marked in the auditory data. In the succeeding two steps a fill-in process and a throwaway process are applied to the results. In the fill-in process short silence segments are relabeled signal and in the throwaway process low energy signal segments are relabeled silence.

Besides silence detection [43] also detects transition points in the signal segments by using break detection and break merging. They compute an onset and offset break to indicate a potential change in category of the underlying signal, by moving a window over the signal segment and compare E_n of different halves of the window at each sliding position. In the second step, adjacent breaks of the same type are merged into a single break.

In [99] music is distinguished from speech, silence, and environmental sounds based on features of the ZCR and the fundamental frequency. To assign the probability of being music to an audio segment, four features are used: the degree of being harmonic (based on fundamental frequency), the degree to which the fundamental frequency concentrates on certain values during a period of time, the variance of the ZCR, and the range of the amplitude of the ZCR.

The first step in reconstructing the textual layout is referred to as tokenization, in this phase the input text is divided into units called tokens or characters. Detection of text shots can be achieved in different ways, depending on the granularity used. If we are only interested in single words we can use the occurrence of white space as the main clue. However, this signal is not necessarily reliable, because of the occurrence of periods, single apostrophes and hyphenation [46]. When more context is taken into account one can reconstruct sentences from the textual layout. Detection of periods is a basic heuristic for the reconstruction of sentences, about 90% of periods are sentence boundary indicators [46]. Transitions are typically found by searching for predefined patterns.

Since layout is very modality dependent, a multimodal approach for its reconstruction won't be very effective. The task of layout reconstruction can currently be performed quite reliably. However, results might improve even further when more advanced techniques are used, for example methods exploiting the learning capabilities of statistical classifiers.

3.3. Content segmentation

In Section 2.2 we introduced the elements of content. Here we will discuss how to detect them automatically, using different detection algorithms exploiting visual, auditory, and textual information sources.

3.3.1. People detection. Detection of people in video documents can be done in several ways. They can be detected in the visual modality by means of their faces or other body parts, in the auditory modality by the presence of speech, and in the textual modality by the appearance of names. In the following, those modality specific techniques will be discussed in more detail. For an in-depth coverage of the different techniques we refer to the cited references.

Most approaches using the visual modality simplify the problem of people detection to detection of a human face. Face detection techniques aim to identify all image regions which contain a face, regardless of its three-dimensional position, orientation, and lighting conditions used, and if present return their image location and extents [95]. This detection is by no means trivial because of variability in location, orientation, scale, and pose. Furthermore, facial expressions, facial hair, glasses, make-up, occlusion, and lightning conditions are known to make detection error prone.

Over the years various methods for the detection of faces in images and image sequences are reported, see [95] for a comprehensive and critical survey of current face detection methods. From all methods currently available the one proposed by Rowley in [70] performs the best [67]. The neural network-based system is able to detect about 90% of all upright and frontal faces, and more important the system only sporadically mistakes non-face areas for faces.

When a face is detected in a video, face recognition techniques aim to identify the person. A common used method for face recognition is matching by means of *Eigenfaces* [64]. Here the matching is performed using single images, and the method is capable to recognize faces under varying pose. In [6] the authors demonstrate that by using *Fisherfaces* the error rates are lower for tests on certain face databases. Moreover the Fisherface method achieves better results when variations in lighting and expression are present simultaneously. A drawback of applying face recognition for video indexing, is its limited generic applicability [74]. Reported results [6, 64, 74] show that face recognition works in constrained environments, preferably showing a frontal face close to the camera. When using face recognition techniques in a video indexing context one should account for this limited applicability.

In [49] people detection is taken one step further, detecting not only the head, but the whole human body. The algorithm presented, first locates the constituent components of the human body by applying detectors for head, legs, left arm, and right arm. Each individual detector is based on the Haar wavelet transform using specific examples. After ensuring that these components are present in the proper geometric configuration, a second example-based classifier combines the results of the component detectors to classify a pattern as either a person or a non-person.

A similar part-based approach is followed in [24] to detect naked people. First, large skin-colored components are found in an image by applying a skin filter that combines color and texture. Based on geometrical constraints between detected components an image is labeled as containing naked people or not. Obviously this method is suited for specific genres only.

The auditory channel also provides strong clues for presence of people in video documents through speech in the segment. When layout segmentation has been performed, classification of the different signal segments as speech can be achieved based on the features computed. Again different approaches can be chosen.

In [99] five features are checked to distinguish speech from other auditory signals. First one is the relation between amplitudes of ZCR and energy curves. The second one is the shape of the ZCR curve. The third and fourth features are the variance and the range of the amplitude of the ZCR curve. The fifth feature is about the property of the short-time fundamental frequency. A decision value is defined for each feature. Based on these features, classification is performed using the weighted average of these decision values.

A more elaborated audio segmentation algorithm is proposed in [43]. The authors are able to segment not only speech but also speech together with noise, speech or music with an accuracy of about 90%. They compared different auditory feature sets, and conclude that temporal and spectral features perform bad, as opposed to Mel-frequency cepstral coefficients (MFCC) and linear prediction coefficients (LPC) which achieve a much better classification accuracy.

When a segment is labeled as speech, speaker recognition can be used to identify a person based on his or her speech utterance. Different techniques are proposed, e.g., [54, 61]. A generic speaker identification system consisting of three modules is presented in [61]. In the first module feature extraction is performed using a set of 14 MFCC from each window. In the second module those features are used to classify each moving window using a nearest neighbor classifier. The classification is performed using a ground truth. In the third module

results of each moving window are combined to generate a single decision for each segment. The authors report encouraging performance using speech segments of a feature film.

A strong textual cue for the appearance of people in a video document are words which are names. In [74], for example, natural language processing techniques using a dictionary, thesaurus, and parser are used to locate names in transcripts. The system calculates a grammatical, lexical, situational, and positional score for each word in the transcripts. A net likelihood score is then calculated which together with the name candidate and segment information forms the system's output. Related to this problem is the task of named entity recognition, which is known from the field of computational linguistics. Here one seeks to classify every word in a document into one of eight categories: person, location, organization, date, time, percentage, monetary value, or none of the above [8]. In the reference, name recognition is viewed as a classification problem, where every word is either part of some name, or not. The authors use a variant of an HMM for the name recognition task based on a bigram language model. Compared to any other reported learning algorithm, their name recognition results are consistently better.

In conclusion, people detection in video can be achieved using different approaches, all having limitations. Variance in orientation and pose, together with occlusion, make visual detection error prone. Speech detection and recognition is still sensitive to noise and environmental sounds. Also, more research on detection of names in text is needed to improve results. As the errors in different modalities are not necessarily correlated, a multimodal approach in detection of persons in video documents can be an improvement. Besides improved detection, fusion of different modalities is interesting with respect to recognition of specific persons.

3.3.2. Object detection. Object detection forms a generalization of the problem of people detection. Specific objects can be detected by means of specialized visual detectors, motion, sounds, and appearance in the textual modality. Object detection methods for the different modalities will be highlighted here.

Approaches for object detection based on visual appearance can range from detection of specific objects to detection approaches of more general objects. An example from the former is given in [76], where the presence of passenger cars in image frames is detected by using a product of histograms. Each histogram represents the joint statistics of a subset of wavelet coefficients and their position on the object. The authors use statistical modelling to account for variation, which enables them to reliably detect passenger cars over a wide range of points of view.

If we know what we are looking for, e.g., people or cars, the task is easier. If not, grouping based on motion is the best in absence of other knowledge. Moreover, since the appearance of objects might vary widely, rigid object motion detection is often the most valuable feature. Thus, when considering the approach for general object detection, motion is a useful feature. A typical method to detect moving objects of interest, starts with a segmentation of the image frame. Regions in the image frame sharing similar motion are merged in the second stage. Result is a motion-based segmentation of the video. In [56] a method is presented that segments a single video frame into independently moving visual objects. The method follows a bottom-up approach, starting with a color-based decomposition of the frame.

Regions are then merged based on their motion parameters via a statistical test, resulting in superior performance over other methods, e.g., [4, 93].

Specific objects can also be detected by analyzing the auditory layout segmentation of the video document. Typically, segments in the layout segmentation first need to be classified as environmental sounds. Subsequently, the environmental sounds are further analyzed for the presence of specific object sound patterns. In [92, 99] for example, specific object sound patterns e.g., dog bark, ringing telephones, and different musical instruments are detected using specific auditory features.

Detecting objects in the textual modality also remains a challenging task. A logical intermediate step in detecting objects of interest in the textual modality is part-of-speech tagging. The latter is the task of labeling each word in a sentence with its appropriate part of speech [46]. Though limited, the information we get from tagging is still quite useful. By extracting and analyzing the nouns in tagged text for example, one can make some assumptions about objects present. This technique is known as chunking [1]. To our knowledge chunking has not yet been used in combination with detection of objects in video documents. Its application however, might prove to be a valuable extension to unimodal object detection.

Successful detection of objects is limited to specific examples. A generic object detector still forms the holy grail in video document analysis. Therefore, multimodal object detection seems interesting. It helps if objects of interest can be identified within different modalities. Then the specific visual appearance, the specific sound, and its mentioning in the accompanying textual data can yield the evidence for robust recognition.

3.3.3. Setting detection. For the detection of setting, motion is not so relevant, as the setting is usually static. Therefore, techniques from the field of content-based image retrieval can be used. See [79] for a complete overview of this field. By using for example key frames, those techniques can easily be used for video indexing. We focus here on methods that assign a setting label to the data, based on analysis of the visual, auditory, or textual modality.

In [82] images are classified as either indoor or outdoor, using three types of visual features: one for color, texture, and frequency information. Instead of computing features on the entire image, the authors use a multi-stage classification approach. First, sub-blocks are classified independently, and afterwards another classification is performed using the k-nearest neighbor classifier.

Outdoor images are further classified into city and landscape images in [87]. Features used are color histograms, color coherence vectors, Discrete Cosine Transform (DCT) coefficients, edge direction histograms, and edge direction coherence vectors. Classification is done with a weighted k-nearest neighbor classifier with leave-one out method. Reported results indicate that the edge direction coherence vector has good discriminatory power for city vs. landscape. Furthermore, it was found that color can be an important cue in classifying natural landscape images into forests, mountains, or sunset/sunrise classes. By analyzing sub-blocks, the authors detect the presence of sky and vegetation in outdoor image frames in another paper. Each sub-block is independently classified, using a Bayesian classification framework, as sky vs. non-sky or vegetation vs. non-vegetation based on color, texture, and position features [86].

Detecting setting based on auditory information, can be achieved by detecting specific environmental sound patterns. In [92] the authors reduce an auditory segment to a small set of parameters using various auditory features, namely loudness, pitch, brightness, bandwidth, and harmonicity. By using statistical techniques over the parameter space the authors accomplish classification and retrieval of several sound patterns including laughter, crowds, and water. In [99] classes of natural and synthetic sound patterns are distinguished by using an HMM, based on timbre and rhythm. The authors are capable of classifying different environmental setting sound patterns, including applause, explosions, rain, river flow, thunder, and windstorm.

The transcript is used in [15] to extract geographic reference information for the video document. The authors match named places to their spatial coordinates. The process begins by using the text metadata as the source material to be processed. A known set of places along with their spatial coordinates, i.e., a gazetteer, is created to resolve geographic references. The gazetteer used consists of approximately 300 countries, states and administrative entities, and 17000 major cities worldwide. After post processing steps, e.g., including related terms and removing stop words, the end result are segments in a video sequence indexed with latitude and longitude.

We conclude that the visual and auditory modality are well suited for recognition of the environment in which the video document is situated. By using the textual modality, a more precise (geographic) location can be extracted. Fusion of the different modalities may provide the video document with semantically interesting setting terms such as: outside vegetation in Brazil near a flowing river. Which can never be derived from one of the modalities in isolation.

4. Multimodal analysis

After reconstruction of the layout and content elements, the next step in the inverse analysis process is analysis of the layout and content to extract the semantic index. At this point the modalities should be integrated. However, before analysis, it might be useful to apply modality conversion of some elements into more appropriate form. The role of conversion and integration in multimodal video document analysis will be discussed in this section, and is illustrated in figure 3.

4.1. Conversion

For analysis, conversion of elements of visual and auditory modalities to text is most appropriate.

A typical component we want to convert from the visual modality is overlayed text. Video Optical Character Recognition (OCR) methods for detection of text in video frames can be divided into component-based, e.g., [78], or texture-based methods, e.g., [44]. A method utilizing the DCT coefficients of compressed video was proposed in [101]. By using Video OCR methods, the visual overlayed text object can be converted into a textual format. The quality of the results of Video OCR vary, depending on the kind of characters used, their color, their stability over time, and the quality of the video itself.

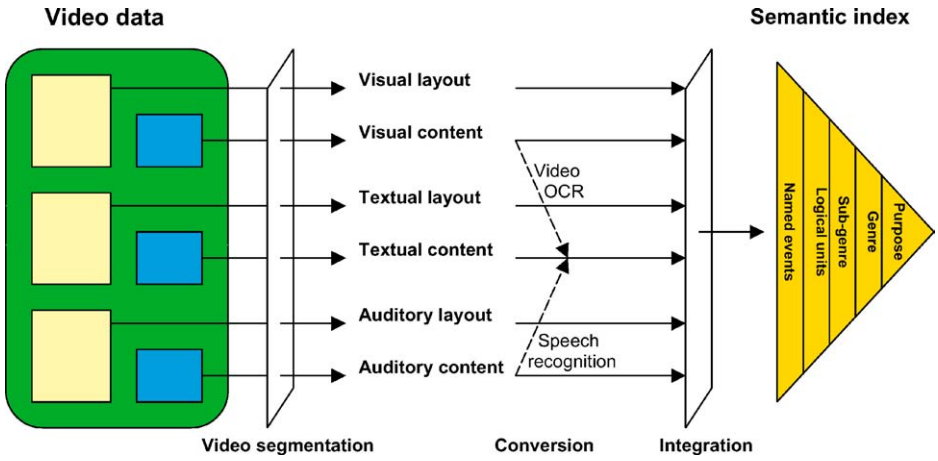

Figure 3. Role of conversion and integration in multimodal video document analysis.

From the auditory modality one typically wants to convert the uttered speech into transcripts. Available speech recognition systems are known to be mature for applications with a single speaker and a limited vocabulary. However, their performance degrades when they are used in real world applications instead of a lab environment [13]. This is especially caused by the sensitivity of the acoustic model to different microphones and different environmental conditions. Since conversion of speech into transcripts still seems problematic, integration with other modalities might prove beneficial.

Note that other conversions are possible, e.g., computer animation can be viewed as converting text to video. However, these are relevant for presentation purposes only.

4.2. Integration

The purpose of integration of multimodal layout and content elements is to improve classification performance. To that end the addition of modalities may serve as a verification method, a method compensating for inaccuracies, or as an additional information source.

An important aspect, indispensable for integration, is synchronization and alignment of the different modalities, as all modalities must have a common timeline. Typically the time stamp is used. We observe that in literature modalities are converted to a format conforming to the researchers main expertise. When audio is the main expertise, image frames are converted to (milli)seconds, e.g., [35]. In [3, 21] image processing is the main expertise, and audio samples are assigned to image frames or camera shots. When a time stamp isn't available, a more advanced alignment procedure is necessary. Such a procedure is proposed in [39]. The error prone output of a speech recognizer is compared and aligned with the accompanying closed captions of news broadcasts. The method first finds matching sequences of words in the transcript and closed caption by performing a dynamic-programming based alignment between the two text strings. Segments are then selected when sequences of three or more words are similar in both resources.

Table 1. An overview of different integration methods.

	Content segmentation		Classification method		Processing cycle	
	Symmetric	Asymmetric	Statistical	Knowledge	Iterated	Non-iterated
[3]	✓		✓			✓
[5]		✓		✓	✓	
[14]	✓		✓			✓
[20]	✓		✓			✓
[21]	✓		✓			✓
[23]	✓			✓		✓
[35]	✓		✓			✓
[35]		✓	✓			✓
[39]	✓		✓			✓
[40]	✓		✓			✓
[53]	✓			✓		✓
[55]	✓		✓		✓	
[66]	✓			✓		✓
[73]	✓			✓		✓
[74]	✓		✓			✓
[80]		✓		✓	✓	
[85]	✓			✓		✓
[91]	✓		✓			✓

To achieve the goal of multimodal integration, several approaches can be followed. We categorize those approaches by their distinctive properties with respect to the processing cycle, the content segmentation, and the classification method used. The processing cycle of the integration method can be iterated, allowing for incremental use of context, or non-iterated. The content segmentation can be performed by using the different modalities in a symmetric, i.e., simultaneous, or asymmetric, i.e., ordered, fashion. Finally, for the classification one can choose between a statistical or knowledge-based approach. An overview of the different integration methods found in literature is in Table 1.

Most integration methods reported are symmetric and non-iterated. Some follow a knowledge-based approach for classification of the data into classes of the semantic index hierarchy [23, 53, 66, 73, 85]. In [85] for example, the auditory and visual modality are integrated to detect speech, silence, speaker identities, no face shot, face shot, and talking face shot using knowledge-based rules. First, talking people are detected by detecting faces in the camera shots, subsequently a knowledge-based measure is evaluated based on the amount of speech in the shot.

Many methods in literature follow a statistical approach [3, 14, 20, 21, 35, 39, 40, 55, 74, 91]. An example of a symmetric, non-iterated statistical integration method is the Name-It system presented in [74]. The system associates detected faces and names, by calculating

a co-occurrence factor that combines the analysis results of face detection and recognition, name extraction, and caption recognition.

Hidden Markov Models are frequently used as a statistical classification method for multimodal integration [3, 20, 21, 35]. A clear advantage of this framework is that it is not only capable to integrate multimodal features, but is also capable to include sequential features. Moreover, an HMM can also be used as a classifier combination method.

When modalities are independent, they can easily be included in a product HMM. In [35] such a classifier is used to train two modalities separately, which are then combined symmetrically, by computing the product of the observation probabilities. It is shown that this results in significant improvement over a unimodal approach.

In contrast to the product HMM method, a neural network-based approach doesn't assume features are independent. The approach presented in [35], trains an HMM for each modality and category. A three layer perceptron is then used to combine the outputs from each HMM in a symmetric and non-iterated fashion.

Another advanced statistical classifier for multimodal integration was recently proposed in [55]. A probabilistic framework for semantic indexing of video documents based on so called multijects and multinets is presented. The multijects model content elements which are integrated in the multinets to model the relations between objects, allowing for symmetric use of modalities. For the integration in the multinet the authors propose a Bayesian belief network [62]. Significant improvements of detection performance is demonstrated. Moreover, the framework supports detection based on iteration. Viability of the Bayesian network as a symmetric integrating classifier was also demonstrated in [40], however that method doesn't support iteration.

In contrast to the above symmetric methods, an asymmetric approach is presented in [35]. A two-stage HMM is proposed which first separates the input video document into three broad categories based on the auditory modality, in the second stage another HMM is used to split those categories based on the visual modality. A drawback of this method is its application dependency, which may result in less effectiveness in other classification tasks.

An asymmetric knowledge-based integration method, supporting iteration, was proposed in [5]. First, the visual and textual modality are combined to generate semantic index results. Those form the input for a post-processing stage that uses those indexes to search the visual modality for the specific time of occurrence of the semantic event.

For exploration of other integration methods, we again take a look in the field of content-based image retrieval. From this field methods are known to integrate the visual and textual modality by combining images with associated captions or HTML tags. Early reported methods used a knowledge base for integration, e.g., the Piction system [80]. This system uses modalities asymmetrically, it first analyzes the caption to identify the expected number of faces and their expected relative positions. Then a face detector is applied to a restricted part of the image, if no faces are detected an iteration step is performed that relaxes the thresholds. More recently, Latent Semantic Indexing (LSI) [19] has become a popular means for integration [14, 91]. LSI is symmetric and non-iterated and works by statistically associating related words to the conceptual context of the given document. In effect it relates documents that use similar terms, which for images are related to features in the image. In [14] LSI is used to capture text statistics in vector form from an HTML document. Words

with specific HTML tags are given higher weights. In addition, the position of the words with respect to the position of the image in the document is also accounted for. The image features, that is the color histogram and the dominant orientation histogram, are also captured in vector form and combined they form a unified vector that the authors use for content-based search of a WWW-based image database. Reported experiments show that maximum improvement was achieved when both visual and textual information are employed.

In conclusion, video indexing results improve when a multimodal approach is followed. Not only because of enhancement of content findings, but also because more information is available. Most methods integrate in a symmetric and non-iterated fashion. Usage of incremental context by means of iteration can be a valuable addition to the success of the integration process. Usage of combined statistical classifiers in multimodal video indexing literature is still scarce, though various successful statistical methods for classifier combinations are known, e.g. bagging, boosting, or stacking [37]. So, probably results can be improved even more substantially when advanced classification methods from the field of statistical pattern recognition, or other disciplines are used, preferably in an iterated fashion.

5. Semantic video indexes

The methodologies described in Section 4 have been applied to extract a variety of the different video indexes described in Section 2.1. In this section we systematically report on the different indexes and the information from which they are derived. As methods for extraction of purpose are not mentioned in literature, this level is excluded. Figure 4 presents an overview of all indexes and the methods in literature which can derive them.

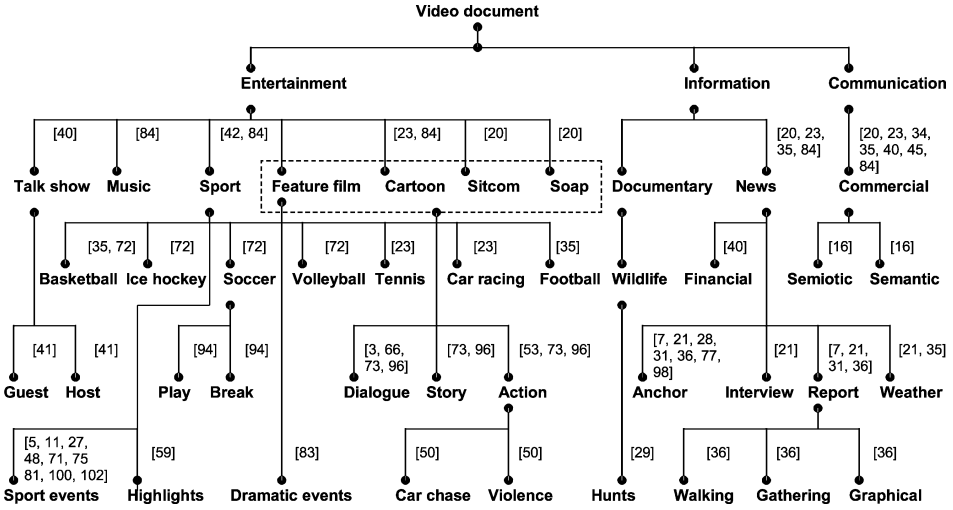

Figure 4. Semantic index hierarchy with instances as found in literature. From top to bottom instances from genre, sub-genre, logical units, and named events. The dashed box is used to group similar nodes.

5.1. Genre

"Editing is an important stylistic element because it affects the overall rhythm of the video document" [9]. Hence, layout related statistics are well suited for indexing a video document into a specific genre. Most obvious element of this editorial style is the average shot length. Generally, the longer the shots, the slower the rhythm of the video document.

The rate of shot changes together with the presence of black frames is used in [34] to detect commercials within news broadcast. The rationale behind detection of black frames is that they are often broadcasted for a fraction of a second before, after, and between commercials. However, black frames can also occur for other reasons. Therefore, the authors use the observation that advertisers try to make commercials more interesting by rapidly cutting between different shots, resulting in a higher shot change rate. A similar approach is followed in [45], for detecting commercials within broadcasted feature films. Besides the detection of monochrome frames and shot change rate, the authors use the edge change ratio and motion vector length to capture high action in commercials.

Average shot length, the percentage of different types of edit transitions, and six visual content features, are used in [84] to classify a video document into cartoons, commercials, music, news and sports video genres. As a classifier the C4.5 decision tree is used.

In [20] the observation is made that different genres exhibit different temporal patterns of face locations. They furthermore observe that the temporal behavior of overlayed text is genre dependent. In fact the following genre dependent functions can be identified:

- *News*: annotation of people, objects, setting, and named events;
- *Sports*: player identification, game related statistics;
- *Movies/TV series*: credits, captions, and language translations;
- *Commercials*: product name, claims, and disclaimers;

Based on results of face and text tracking, each frame is assigned one of 15 labels, describing variations on the number of appearing faces and/or text lines together with the distance of a face to the camera. These labels form the input for an HMM, which classifies an input video document into news, commercials, sitcoms, and soaps based on maximum likelihood.

Detection of generic sport video documents seems almost impossible due to the large variety in sports. In [42], however, a method is presented that is capable of identifying mainstream sports videos. Discriminating properties of sport videos are the presence of slow-motion replays, large amounts of overlayed text, and specific camera/object motion. The authors propose a set of eleven features to capture these properties, and obtain 93% accuracy using a decision tree classifier. Analysis showed that motion magnitude and direction of motion features yielded the best results.

Methods for indexing video documents into a specific genre using a multimodal approach are reported in [23, 35, 40]. In [35] news reports, weather forecasts, commercials, basketball, and football games are distinguished based on audio and visual information. The authors compare different integration methods and classifiers and conclude that a product HMM classifier is most suited for their task, see also Section 4.2.

The same modalities are used in [23]. The authors present a three-step approach. In the first phase, content features such as color statistics, motion vectors and audio statistics

are extracted. Secondly, layout features are derived, e.g., shot lengths, camera motion, and speech vs. music. Finally, a style profile is composed and an educational guess is made as to the genre in which a shot belongs. They report promising results by combining different layout and content attributes of video for analysis, and can find five (sub)genres, namely news broadcasts, car racing, tennis, commercials, and animated cartoons.

Besides auditory and visual information, [40] also exploits the textual modality. The segmentation and indexing approach presented uses three layers to process low-, mid-, and high-level information. At the lowest level features such as color, shape, MFCC, ZCR, and the transcript are extracted. Those are used in the mid-level to detect faces, speech, keywords, etc. At the highest level the semantic index is extracted through the integration of mid-level features across the different modalities, using Bayesian networks, as noted in Section 4.2. In its current implementation the presented system classifies segments as either part of a talk show, commercial or financial news.

5.2. Sub-genre

Research on indexing sub-genres, or specific instances of a genre, has been geared mainly towards sport videos [23, 35, 72] and commercials [16]. Obviously, future index techniques may also extract other sub-genres, for example westerns, comedies, or thrillers within the feature film genre.

Four sub-genres of sport video documents are identified in [72]: basketball, ice hockey, soccer, and volleyball. The full motion fields in consecutive frames are used as a feature. To reduce the feature, Principal Component Analysis is used. For classification two different statistical classifiers were applied. It was found that a continuous observation density Markov model gave the best results. The sequences analyzed were post-edited to contain only the play of the sports, which is a drawback of the presented system. For instance, no crowd scenes or time outs were included. Some sub-genres of sport video documents are also detected in [23, 35], as noted in Section 5.1.

An approach to index commercial videos based on semiotic and semantic properties is presented in [16]. Semiotics classifies commercials into four different sub-genres that relate to the narrative of the commercial. The following four sub-genres are distinguished: practical, critical, utopic, and playful commercials. Perceptual features e.g., saturated colors, horizontal lines, and the presence or absence of recurring colors, are mapped onto the semiotic categories. Based on research in the marketing field, the authors also formalized a link between editing, color, and motion effects on the one hand, and feelings that the video arouses in the observer on the other. Characteristics of a commercial are related to those feelings and have been organized in a hierarchical fashion. A main classification is introduced between commercials that induce feelings of *action* and those that induce feelings of *quietness*. The authors subdivide action further into suspense and excitement. Quietness is further specified in relaxation and happiness.

5.3. Logical units

Detection of logical units in video documents is extensively researched with respect to the detection of scenes or Logical Story Units (LSU) in feature films and sitcoms. An overview

and evaluation of such methods is presented in [88]. However, detection of LSU boundaries alone is not enough. For indexing, we are especially interested in its accompanying label.

A method that is capable of detecting dialogue scenes in movies and sitcoms, is presented in [3]. Based on audio analysis, face detection, and face location analysis the authors generate output labels which form the input for an HMM. The HMM outputs a scene labeled as either, establishing scene, transitional scene, or dialogue scene. According to the results presented, combined audio and face information gives the most consistent performance of different observation sets and training data. However, in its current design, the method is incapable of differentiating between dialogue and monologue scenes.

A technique to characterize and index violent scenes in general TV drama and movies is presented in [53]. The authors integrate cues from both the visual and auditory modality symmetrically. First, the spatio-temporal dynamic activity of each video shot is computed as a measure of action. This is combined with detection of flames and blood using a predefined color table. The corresponding audio information provides supplemental evidence for the identification of violent scenes. The focus is on the abrupt change in energy level of the audio signal, computed using the energy entropy criterion. As a classifier the authors use a knowledge-based combination of feature values on scene level.

By utilizing a symmetric and non-iterated multimodal integration method four different types of scenes are identified in [73]. The audio signal is segmented into silence, speech, music, and miscellaneous sounds. This is combined with a visual similarity measure, computed within a temporal window. Dialogues are then detected based on the occurrence of speech and an alternated pattern of visual labels, indicating a change of speaker. When the visual pattern exhibits a repetition the scene is labeled as story. When the audio signal isn't labeled as speech, and the visual information exhibits a progressive pattern, with contrasting visual content, the scene is labeled as action. Finally, scenes that don't fit in the aforementioned categories are indexed as generic scenes.

In contrast to [73], a unimodal approach based on the visual information source is used in [96] to detect dialogues, actions, and story units. Shots that are visually similar and temporally close to each other are assigned the same (arbitrary) label. Based on the patterns of labels in a scene, it is indexed as either dialogue, action, or story unit.

A scheme for reliably identifying logical units which clusters sensor shots according to detected dialogues, similar settings, or similar audio is presented in [66]. The method starts by calculating specific features for each camera and microphone shot. Auditory, color, and orientation features are supported as well as face detection. Next an Euclidean metric is used to determine the distance between shots with respect to the features, resulting in a so called distance table. Based on the distance tables, shots are merged into logical units using absolute and adaptive thresholds.

News broadcasts are far more structured than feature films. Researchers have exploited this to classify logical units in news video using a model-based approach. Especially anchor shots are easy to model and therefore easy to detect. Since there is only minor body movement they can be detected by comparison of the average difference between (regions in) successive frames. This difference will be minimal. This observation is used in [28, 77, 98]. In [28, 77] also the restricted position and size of detected faces is used.

Another approach for the detection of anchor shots is taken in [7, 31, 36]. Repetition of visually similar anchor shots throughout the news broadcast is exploited. To refine the classification of the similarity measure used [7], requires anchor shots candidates to have a motion quantity below a certain threshold. Each shot is classified as either anchor or report. Moreover, textual descriptors are added based on extracted captions and recognized speech. To classify report and anchor shots, the authors in [36] use face and lip movement detection. To distinguish anchor shots, the aforementioned classification is extended with the knowledge that anchor shots are graphically similar and occur frequently in a news broadcast. The largest cluster of similar shots is therefore assigned to the class of anchor shots. Moreover, the detection of a title caption is used to detect anchor shots that introduce a new topic. In [31] anchor shots are detected together with silence intervals to indicate report boundaries. Based on a topics database the presented system finds the most probable topic per report by analyzing the transcribed speech. Opposed to [7, 36], final descriptions are not added to shots, but to a sequence of shots that constitute a complete report on one topic. This is achieved by merging consecutive segments with the same topic in their list of most probable topics.

Besides the detection of anchor persons and reports, other logical units can be identified. In [21] six main logical units for TV broadcast news are distinguished, namely, begin, end, anchor, interview, report, and weather forecast. Each logical unit is represented by an HMM. For each frame of the video one feature vector is calculated consisting of 25 features, including motion and audio features. The resulting feature vector sequence is assigned to a logical unit based on the sequence of HMMs that maximizes the probability of having generated this feature vector sequence. By using this approach parsing and indexing of the video is performed in one pass through the video only.

Other examples of highly structured TV broadcasts are talk and game shows. In [41] a method is presented that detects guest and host shots in those video documents. The basic observation used is that in most talk shows the same person is host for the duration of the program but guests keep on changing. Also host shots are typically shorter since only the host asks questions. For a given show, the key frames of the N shortest shots containing one detected face are correlated in time to find the shot most often repeated. The key host frame is then compared against all key frames to detect all similar host shots, and guest shots.

In [94] a model for segmenting soccer video into the logical units break and play is given. A grass-color ratio is used to classify frames into three views according to video shooting scale, namely global, zoom-in, and close-up. Based on segmentation rules, the different views are mapped. Global views are classified as play and close-ups as breaks if they have a minimum length. Otherwise a neighborhood voting heuristic is used for classification.

5.4. Named events

Named events are at the lowest level in the semantic index hierarchy. For their detection different techniques have been used.

A three-level event detection algorithm is presented in [29]. The first level of the algorithm extracts generic color, texture, and motion features, and detects moving object blobs. The

mid-level employs a domain dependent neural network to verify whether the moving blobs belong to objects of interest. The generated shot descriptors are then used by a domain-specific inference process at the third level to detect the video segments that contain events of interest. To test the effectiveness of the algorithm the authors applied it to detect animal hunt events in wildlife documentaries.

Violent events and car chases in feature films are detected in [50], based on analysis of environmental sounds. First, low level sounds as engines, horns, explosions, or gunfire are detected, which constitute part of the high level sound events. Based on the dominance of those low level sounds in a segment it is labeled with a high level named event.

Walking shots, gathering shots, and computer graphics shots in broadcast news are the named events detected in [36]. A walking shot is classified by detecting the up and down oscillation of the bottom of a facial region. When more than two similar sized facial regions are detected in a frame, a shot is classified as a gathering shot. Finally, computer graphics shots are classified by a total lack of motion in a series of frames.

The observation that authors use lightning techniques to intensify the drama of certain scenes in a video document is exploited in [83]. An algorithm is presented that detects flash-lights, which is used as an identifier for dramatic events in feature films, based on features derived from the average frame luminance and the frame area influenced by the flashing light. Five types of dramatic events are identified that are related to the appearance of flashlights, i.e., supernatural power, crisis, terror, excitement, and generic events of great importance.

Whereas a flashlight can indicate a dramatic event in feature films, slow motion replays are likely to indicate semantically important events in sport video documents. In [59] a method is presented that localizes such events by detecting slow motion replays. The slow-motion segments are modeled and detected by an HMM.

One of the most important events in a sport video document is a score. In [5] a link between the visual and textual modalities is made to identify events that change the score in American football games. The authors investigate whether a chain of keywords, corresponding to an event, is found from the closed caption stream or not. In the time frames corresponding to those keywords, the visual stream is analyzed. Key frames of camera shots in the visual stream are compared with predefined templates using block matching based on the color distribution. Finally, the shot is indexed by the most likely score event, for example a touchdown.

Besides American football, methods for detecting events in tennis [48, 81, 100], soccer [11, 27], baseball [71, 100] and basketball [75, 102] are reported in literature. Commonly, the methods presented exploit domain knowledge and simple (visual) features related to color, edges, and camera/object motion to classify typical sport specific events e.g., smashes, corner kicks, and dunks using a knowledge-based classifier. An exception to this common approach is [71], which presents an algorithm that identifies highlights in baseball video by analyzing the auditory modality only. Highlight events are identified by detecting excited speech of the commentators and the occurrence of a baseball pitch and hit.

Besides semantic indexing, detection of named events also forms a great resource for reuse of video documents. Specific information can be retrieved and reused in different contexts, or reused to automatically generate summaries of video documents. This seems especially interesting for, but is not limited to, video documents from the sport genre.

5.5. Discussion

Now that we have described the different semantic index techniques, as encountered in literature, we are able to distinguish the most prominent content and layout properties per genre. As variation in the textual modality is in general too diverse for differentiation of genres, and more suited to attach semantic meaning to logical units and named events, we focus here on properties derived from the visual and auditory modality only. Though, a large amount of genres can be distinguished, we limit ourselves to the ones mentioned in the semantic index hierarchy in figure 4, i.e., talk show, music, sport, feature film, cartoon, sitcom, soap, documentary, news, and commercial. For each of those genres we describe the characteristic properties.

Most prominent property of the first genre, i.e., talk shows, is their well-defined structure, uniform setting, and prominent presence of dialogues, featuring mostly non-moving frontal faces talking close to the camera. Besides closing credits, there is in general a limited use of overlayed text.

Whereas talk shows have a well-defined structure and limited setting, music clips show great diversity in setting and mostly have ill-defined structure. Moreover, music will have many short camera shots, showing lots of camera and object motion, separated by many gradual transition edits and long microphone shots containing music. The use of overlayed text is mostly limited to information about the performing artist and the name of the song on a fixed position.

Sport broadcasts come in many different flavors, not only because there exist a tremendous amount of sport sub-genres, but also because they can be broadcasted live or in summarized format. Despite this diversity, most authored sport broadcasts are characterized by a voice over reporting on named events in the game, a watching crowd, high frequency of long camera shots, and overlayed text showing game and player related information on a fixed frame position. Usually sport broadcasts contain a vast amount of camera motion, objects, and players within a limited uniform setting. Structure is sport-specific, but in general, a distinction between different logical units can be made easily. Moreover, a typical property of sport broadcasts is the use of replays showing events of interest, commonly introduced and ended by a gradual transition edit.

Feature film, cartoon, sitcom, and soap share similar layout and content properties. They are all dominated by people (or toons) talking to each other or taking part in action scenes. They are structured by means of scenes. The setting is mostly limited to a small amount of locales, sometimes separated by means of visual, e.g., gradual, or auditory, e.g., music, transition edits. Moreover, setting in cartoons is characterized by usage of saturated colors, also the audio in cartoons is almost noise-free due to studio recording of speech and special effects. For all mentioned genres the usage of overlayed text is limited to opening and/or closing credits. Feature film, cartoon, sitcom, and soap differ with respect to people appearance, usage of special effects, presence of object and camera motion, and shot rhythm. Appearing people are usually filmed frontal in sitcoms and soaps, whereas in feature films and cartoons there is more diversity in appearance of people or toons. Special effects are most prominent in feature films and cartoons, laughter of an imaginary public is sometimes added to sitcoms. In sitcoms and soaps there is limited camera and object motion. In general

cartoons also have limited camera motion, though object motion appears more frequently. In feature films both camera and object motion are present. With respect to shot rhythm it seems legitimate to state that this has stronger variation in feature films and cartoons. The perceived rhythm will be slowest for soaps, resulting in more frequent use of camera shots with relative long duration.

Documentaries can also be characterized by their slow rhythm. Other properties that are typical for this genre are the dominant presence of a voice over narrating about the content in long microphone shots. Motion of camera and objects might be present in the documentary, the same holds for overlayed text. Mostly there is no well-defined structure. Special effects are seldom used in documentaries.

Most obvious property of news is its well-defined structure. Different news reports and interviews are alternated by anchor persons introducing, and narrating about, the different news topics. A news broadcast is commonly ended by a weather forecast. Those logical units are mostly dominated by monologues, e.g., people talking in front of a camera showing little motion. Overlayed text is frequently used on fixed positions for annotation of people, objects, setting, and named events. A report on an incident may contain camera and object motion. Similarity of studio setting is also a prominent property of news broadcasts, as is the abrupt nature of transitions between sensor shots.

Some prominent properties of the final genre, i.e., commercials, are similar to those of music. They have a great variety in setting, and share no common structure, although they are authored carefully, as the message of the commercial has to be conveyed in twenty seconds or so. Frequent usage of abrupt and gradual transition, in both visual and auditory modality, is responsible for the fast rhythm. Usually lots of object and camera motion, in combination with special effects, such as a loud volume, is used to attract the attention of the viewer. Difference with music is that black frames are used to separate commercials, the presence of speech, the superfluous and non-fixed use of overlayed text, a disappearing station logo, and the fact that commercials usually end with a static frame showing the product or brand of interest.

Due to the large variety in broadcasting formats, which is a consequence of guidance by different authors, it is very difficult to give a general description for the structure and characterizing properties of the different genres. When considering sub-genres this will only become more difficult. Is a sports program showing highlights of today's sport matches a sub-genre of sport or news? Reducing the prominent properties of broadcasts to instances of layout and content elements, and splitting of the broadcasts into logical units and named events seems a necessary intermediate step to arrive at a more consistent definition of genre and sub-genre. More research on this topic is still necessary.

6. Conclusion

Viewing a video document from the perspective of its author, enabled us to present a framework for multimodal video indexing. This framework formed the starting point for our review on different state-of-the-art video indexing techniques. Moreover, it allowed us to answer the three different questions that arise when assigning an index to a video document. The question *what to index* was answered by reviewing different techniques for

layout reconstruction. We presented a discussion on reconstruction of content elements and integration methods to answer the *how to index* question. Finally, the *which index* question was answered by naming different present and future index types within the semantic index hierarchy of the proposed framework.

At the end of this review we stress that multimodal analysis is the future. However, more attention, in the form of research, needs to be given to the following factors:

1. *Content segmentation.* Content segmentation forms the basis of multimodal video analysis. In contrast to layout reconstruction, which is largely solved, there is still a lot to be gained in improved segmentation for the three content elements, i.e., people, objects, and setting. Contemporary detectors are well suited for detection and recognition of content elements within certain constraints. Most methods for detection of content elements still adhere to a unimodal approach. A multimodal approach might prove to be a fruitful extension. It allows to take additional context into account. Bringing the semantic index on a higher level is the ultimate goal for multimodal analysis. This can be achieved by the integrated use of different robust content detectors or by choosing a constrained domain that ensures the best detection performance for a limited detector set.
2. *Modality usage.* Within the research field of multimodal video indexing, focus is still too much geared towards the visual and auditory modality. The semantic rich textual modality is largely ignored in combination with the visual or auditory modality. Specific content segmentation methods for the textual modality will have their reflection on the semantic index derived. Ultimately this will result in semantic descriptions that make a video document as accessible as a text document.
3. *Multimodal integration.* The integrated use of different information sources is an emerging trend in video indexing research. All reported integration methods indicate an improvement of performance. Most methods integrate in a symmetric and non-iterated fashion. Usage of incremental context by means of iteration can be a valuable addition to the success of the integration process. Most successful integration methods reported are based on the HMM and Bayesian network framework, which can be considered as the current state-of-the-art in multimodal integration. There seems to be a positive correlation between usage of advanced integration methods and multimodal video indexing results. This paves the road for the exploration of classifier combinations from the field of statistical pattern recognition, or other disciplines, within the context of multimodal video indexing.
4. *Technique taxonomy.* We presented a semantic index hierarchy that grouped different index types as found in literature. Moreover we characterized the different genres in terms of their most prominent layout and content elements, and by splitting its structure into logical units and named events. What the field of video indexing still lacks is a taxonomy of different techniques that indicates why a specific technique is suited the best, or unsuited, for a specific group of semantic index types.

The impact of the above mentioned factors on automatic indexing of video documents will not only make the process more efficient and more effective than it is now, it will also yield richer semantic indexes. This will form the basis for a range of new innovative applications.

Acknowledgments

The authors thank Jeroen Vendrig and Ioannis Patras from the University of Amsterdam for their valuable comments and suggestions. This research is sponsored by the ICES/KIS Multimedia Information Analysis project and TNO Institute of Applied Physics (TPD).

Notes

1. A gradual transition actually contains pieces of two camera shots, for simplicity we regard it as a separate entity.
2. As an ironic legacy from early research on video parsing, this is also referred to as scene-change detection.

References

1. S. Abney, "Part-of-speech tagging and partial parsing," in Corpus-Based Methods in Language and Speech Processing, S. Young and G. Bloothooft (Eds.), Kluwer Academic Publishers, Dordrecht, 1997, pp. 118–136.
2. S. Adali, K.S. Candan, S.S. Chen, K. Erol, and V.S. Subrahmanian, "The advanced video information system: Data structures and query processing," Multimedia Systems, Vol. 4, No. 4, pp. 172–186, 1996.
3. A.A. Alatan, A.N. Akansu, and W. Wolf, "Multi-modal dialogue scene detection using hidden markov models for content-based multimedia indexing," Multimedia Tools and Applications, Vol. 14, No. 2, pp. 137–151, 2001.
4. Y. Altunbasak, P.E. Eren, and A.M. Tekalp, "Region-based parametric motion segmentation using color information," Graphical Models and Image Processing, Vol. 60, No. 1, pp. 13–23, 1998.
5. N. Babaguchi, Y. Kawai, and T. Kitahashi, "Event based indexing of broadcasted sports video by intermodal collaboration," IEEE Transactions on Multimedia, Vol. 4, No. 1, pp. 68–75, 2002.
6. P.N. Belhumeur, J.P. Hespanha, and D.J. Kriegman, "Eigenfaces vs. fisherfaces: Recognition using class specific linear projection," IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 19, No. 7, pp. 711–720, 1997.
7. M. Bertini, A. Del Bimbo, and P. Pala, "Content-based indexing and retrieval of TV news," Pattern Recognition Letters, Vol. 22, No. 5, pp. 503–516, 2001.
8. D. Bikel, R. Schwartz, and R.M. Weischedel, "An algorithm that learns what's in a name," Machine Learning, Vol. 34, Nos. 1–3, pp. 211–231, 1999.
9. J.M. Boggs and D.W. Petrie, The Art of Watching Films, 5th edition, Mayfield Publishing Company: Mountain View, USA, 2000.
10. R.M. Bolle, B.-L. Yeo, and M.M. Yeung, "Video query: Research directions," IBM Journal of Research and Development, Vol. 42, No. 2, pp. 233–252, 1998.
11. A. Bonzanini, R. Leonardi, and P. Migliorati, "Event recognition in sport programs using low-level motion indices," in IEEE International Conference on Multimedia & Expo, Tokyo, Japan, 2001, pp. 1208–1211.
12. M. Brown, J. Foote, G. Jones, K. Sparck-Jones, and S. Young, "Automatic content-based retrieval of broadcast news," in ACM Multimedia 1995, San Francisco, USA, 1995.
13. R. Brunelli, O. Mich, and C.M. Modena, "A survey on the automatic indexing of video data," Journal of Visual Communication and Image Representation, Vol. 10, No. 2, pp. 78–112, 1999.
14. M. La Cascia, S. Sethi, and S. Sclaroff, "Combining textual and visual cues for content-based image retrieval on the world wide web," in IEEE Workshop on Content-Based Access of Image and Video Libraries, 1998.
15. M. Christel, A. Olligschlaeger, and C. Huang, "Interactive maps for a digital video library," IEEE Multimedia, Vol. 7, No. 1, pp. 60–67, 2000.
16. C. Colombo, A. Del Bimbo, and P. Pala, "Semantics in visual information retrieval," IEEE Multimedia, Vol. 6, No. 3, pp. 38–53, 1999.
17. Convera. `http://www.convera.com`.

18. G. Davenport, T. Aguierre Smith, and N. Pincever, "Cinematic principles for multimedia," in IEEE Computer Graphics & Applications, Vol. 11, No. 4, pp. 67–74, 1991.
19. S. Deerwester, S.T. Dumais, G.W. Furnas, T.K. Landauer, and R. Harshman, "Indexing by latent semantic analysis," Journal of the American Society for Information Science, Vol. 41, No. 6, pp. 391–407, 1990.
20. N. Dimitrova, L. Agnihotri, and G. Wei, "Video classification based on HMM using text and faces," in European Signal Processing Conference, Tampere, Finland, 2000.
21. S. Eickeler and S. Müller, "Content-based video indexing of TV broadcast news using hidden markov models," in IEEE International Conference on Acoustics, Speech, and Signal Processing, Phoenix, USA, 1999, pp. 2997–3000.
22. K. El-Maleh, M. Klein, G. Petrucci, and P. Kabal, "Speech/music discrimination for multimedia applications," in IEEE International Conference on Acoustics, Speech, and Signal Processing, Istanbul, Turkey, 2000, pp. 2445–2448.
23. S. Fischer, R. Lienhart, and W. Effelsberg, "Automatic recognition of film genres," in ACM Multimedia 1995, San Francisco, USA, 1995, pp. 295–304.
24. M.M. Fleck, D.A. Forsyth, and C. Bregler, "Finding naked people," in European Conference on Computer Vision, Cambridge, UK, 1996, Vol. 2, pp. 593–602.
25. B. Furht, S.W. Smoliar, and H.J. Zhang, Video and Image Processing in Multimedia Systems, 2nd edition, Kluwer Academic Publishers: Norwell, USA, 1996.
26. A. Ghias, J. Logan, D. Chamberlin, and B.C. Smith, "Query by humming—musical information retrieval in an audio database," in ACM Multimedia 1995, San Francisco, USA, 1995.
27. Y. Gong, L.T. Sin, and C.H. Chuan, "Automatic parsing of TV soccer programs," in IEEE International Conference on Multimedia Computing and Systems, 1995, pp. 167–174.
28. B. Günsel, A.M. Ferman, and A.M. Tekalp, "Video indexing through integration of syntactic and semantic features," in Third IEEE Workshop on Applications of Computer Vision, Sarasota, USA, 1996.
29. N. Haering, R. Qian, and I. Sezan, "A semantic event-detection approach and its application to detecting hunts in wildlife video," IEEE Transactions on Circuits and Systems for Video Technology, Vol. 10, No. 6, pp. 857–868, 2000.
30. A. Hampapur, R. Jain, and T. Weymouth, "Feature based digital video indexing," in IFIP 2.6 Third Working Conference on Visual Database Systems, Lausanne, Switzerland, 1995.
31. A. Hanjalic, G. Kakes, R.L. Lagendijk, and J. Biemond, "Dancers: Delft advanced news retrieval system," in IS&T/SPIE Electronic Imaging 2001: Storage and Retrieval for Media Databases 2001, San Jose, USA, 2001.
32. A. Hanjalic, G.C. Langelaar, P.M.B. van Roosmalen, J. Biemond, and R.L. Lagendijk, Image and Video Databases: Restoration, Watermarking and Retrieval, Elsevier Science: Amsterdam, The Netherlands, 2000.
33. A.G. Hauptmann, D. Lee, and P.E. Kennedy, "Topic labeling of multilingual broadcast news in the informedia digital video library," in ACM DL/SIGIR MIDAS Workshop, Berkely, USA, 1999.
34. A.G. Hauptmann and M.J. Witbrock, "Story segmentation and detection of commercials in broadcast news video," in ADL-98 Advances in Digital Libraries, Santa Barbara, USA, 1998, pp. 168–179.
35. J. Huang, Z. Liu, Y. Wang, Y. Chen, and E.K. Wong, "Integration of multimodal features for video scene classification based on HMM," in IEEE Workshop on Multimedia Signal Processing, Copenhagen, Denmark, 1999.
36. I. Ide, K. Yamamoto, and H. Tanaka, "Automatic video indexing based on shot classification," in First International Conference on Advanced Multimedia Content Processing, Vol. 1554 of Lecture Notes in Computer Science, Springer-Verlag: Osaka, Japan, 1999.
37. A.K. Jain, R.P.W. Duin, and J. Mao, "Statistical pattern recognition: A review," IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 22, No. 1, pp. 4–37, 2000.
38. R. Jain and A. Hampapur, "Metadata in video databases," ACM SIGMOD, Vol. 23, No. 4, pp. 27–33, 1994.
39. P.J. Jang and A.G. Hauptmann, "Learning to recognize speech by watching television," IEEE Intelligent Systems, Vol. 14, No. 5, pp. 51–58, 1999.
40. R.S. Jasinschi, N. Dimitrova, T. McGee, L. Agnihotri, J. Zimmerman, and D. Li, "Integrated multimedia processing for topic segmentation and classification," in IEEE International Conference on Image Processing, Thessaloniki, Greece, 2001, pp. 366–369.

41. O. Javed, Z. Rasheed, and M. Shah, "A framework for segmentation of talk & game shows," in IEEE International Conference on Computer Vision, Vancouver, Canada, 2001.
42. V. Kobla, D. DeMenthon, and D. Doermann, "Identification of sports videos using replay, text, and camera motion features," in SPIE Conference on Storage and Retrieval for Media Databases, Vol. 3972, pp. 332–343, 2000.
43. D. Li, I.K. Sethi, N. Dimitrova, and T. McGee, "Classification of general audio data for content-based retrieval," Pattern Recognition Letters, Vol. 22, No. 5, pp. 533–544, 2001.
44. H. Li, D. Doermann, and O. Kia, "Automatic text detection and tracking in digital video," IEEE Transactions on Image Processing, Vol. 9, No. 1, pp. 147–156, 2000.
45. R. Lienhart, C. Kuhmünch, and W. Effelsberg, "On the detection and recognition of television commercials," in IEEE Conference on Multimedia Computing and Systems, Ottawa, Canada, 1997, pp. 509–516.
46. C.D. Manning and H. Schütze, Foundations of Statistical Natural Language Processing. The MIT Press, Cambridge, USA, 1999.
47. K. Minami, A. Akutsu, H. Hamada, and Y. Tomomura, "Video handling with music and speech detection," IEEE Multimedia, Vol. 5, No. 3, pp. 17–25, 1998.
48. H. Miyamori and S. Iisaku, "Video annotation for content-based retrieval using human behavior analysis and domain knowledge," in IEEE International Conference on Automatic Face and Gesture Recognition, Grenoble, France, 2000, pp. 26–30.
49. A. Mohan, C. Papageorgiou, and T. Poggio, "Example-based object detection in images by components," IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 23, No. 4, pp. 349–361, 2001.
50. S. Moncrieff, C. Dorai, and S. Venkatesh, "Detecting indexical signs in film audio for scene interpretation," in IEEE International Conference on Multimedia & Expo, Tokyo, Japan, 2001, pp. 1192–1195.
51. F. Nack and A.T. Lindsay, "Everything you always wanted to know about MPEG-7: Part 1," IEEE Multimedia, Vol. 6, No. 3, pp. 65–77, 1999.
52. F. Nack and A.T. Lindsay, "Everything you always wanted to know about MPEG-7: Part 2," IEEE Multimedia, Vol. 6, No. 4, pp. 64–73, 1999.
53. J. Nam, M. Alghoniemy, and A.H. Tewfik, "Audio-visual content-based violent scene characterization," in IEEE International Conference on Image Processing, Chicago, USA, 1998, Vol. 1, pp. 353–357.
54. J. Nam, A. Enis Cetin, and A.H. Tewfik, "Speaker identification and video analysis for hierarchical video shot classification," in IEEE International Conference on Image Processing, Washington DC, USA, 1997, Vol. 2.
55. M.R. Naphade and T.S. Huang, "A probabilistic framework for semantic video indexing, filtering, and retrieval," IEEE Transactions on Multimedia, Vol. 3, No. 1, pp. 141–151, 2001.
56. H.T. Nguyen, M. Worring, and A. Dev, "Detection of moving objects in video using a robust motion similarity measure," IEEE Transactions on Image Processing, Vol. 9, No. 1, pp. 137–141, 2000.
57. L. Nigay and J. Coutaz, "A design space for multimodal systems: concurrent processing and data fusion." in INTERCHI'93 Proceedings, Amsterdam, the Netherlands, 1993, pp. 172–178.
58. D.W. Oard, "The state of the art in text filtering," User Modeling and User-Adapted Interaction, Vol. 7, No. 3, pp. 141–178, 1997.
59. H. Pan, P. Van Beek, and M.I. Sezan, "Detection of slow-motion replay segments in sports video for highlights generation," in IEEE International Conference on Acoustic, Speech and Signal Processing, 2001.
60. N.V. Patel and I.K. Sethi, "Audio characterization for video indexing," in Proceedings SPIE on Storage and Retrieval for Still Image and Video Databases, San Jose, USA, 1996, Vol. 2670, pp. 373–384.
61. N.V. Patel and I.K. Sethi, "Video classification using speaker identification," in IS&T SPIE, Proceedings: Storage and Retrieval for Image and Video Databases IV, San Jose, USA, 1997.
62. J. Pearl, Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference. Morgan Kaufmann: San Mateo, USA, 1988.
63. A.K. Peker, A.A. Alatan, and A.N. Akansu, "Low-level motion activity features for semantic characterization of video," in IEEE International Conference on Multimedia & Expo, New York City, USA, 2000.

64. A. Pentland, B. Moghaddam, and T. Starner, "View-based and modular eigenspaces for face recognition," in IEEE International Conference on Computer Vision and Pattern Recognition, Seattle, USA, 1994.
65. S. Pfeiffer, S. Fischer, and W. Effelsberg, "Automatic audio content analysis," in ACM Multimedia 1996, Boston, USA, 1996, pp. 21–30.
66. S. Pfeiffer, R. Lienhart, and W. Effelsberg, "Scene determination based on video and audio features," Multimedia Tools and Applications, Vol. 15, No. 1, pp. 59–81, 2001.
67. T.V. Pham and M. Worring, "Face detection methods: A critical evaluation," Technical Report 2000-11, Intelligent Sensory Information Systems, University of Amsterdam, 2000.
68. Praja. `http://www.praja.com`.
69. L.R. Rabiner, "A tutorial on hidden markov models and selected applications in speech recognition," Proceedings of the IEEE, Vol. 77, No. 2, pp. 257–286, 1989.
70. H.A. Rowley, S. Baluja, and T. Kanade, "Neural network-based face detection," IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 20, No. 1, pp. 23–38, 1998.
71. Y. Rui, A. Gupta, and A. Acero, "Automatically extracting highlights for TV baseball programs," in ACM Multimedia 2000, Los Angeles, USA, 2000, pp. 105–115.
72. E. Sahouria and A. Zakhor, "Content analysis of video using principal components," IEEE Transactions on Circuits and Systems for Video Technology, Vol. 9, No. 8, pp. 1290–1298, 1999.
73. C. Saraceno and R. Leonardi, "Identification of story units in audio-visual sequences by joint audio and video processing," in IEEE International Conference on Image Processing, Chicago, USA, 1998.
74. S. Satoh, Y. Nakamura, and T. Kanade, "Name-It: Naming and detecting faces in news videos," IEEE Multimedia, Vol. 6, No. 1, pp. 22–35, 1999.
75. D.D. Saur, Y.-P. Tan, S.R. Kulkarni, and P.J. Ramadge, "Automated analysis and annotation of basketball video," in SPIE's Electronic Imaging conference on Storage and Retrieval for Image and Video Databases V, San Jose, USA, 1997, Vol. 3022, pp. 176–187.
76. H. Schneiderman and T. Kanade, "A statistical method for 3D object detection applied to faces and cars," in IEEE Computer Vision and Pattern Recognition, Hilton Head, USA, 2000.
77. K. Shearer, C. Dorai, and S. Venkatesh, "Incorporating domain knowledge with video and voice data analysis in news broadcasts," in ACM International Conference on Knowledge Discovery and Data Mining, Boston, USA, 2000, pp. 46–53.
78. J. Shim, C. Dorai, and R. Bolle, "Automatic text extraction from video for content-based annotation and retrieval," in IEEE International Conference on Pattern Recognition, 1998, pp. 618–620.
79. A.W.M. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain, "Content based image retrieval at the end of the early years," IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 22, No. 12, pp. 1349–1380, 2000.
80. R.K. Srihari, "Automatic indexing and content-based retrieval of captioned images," IEEE Computer, Vol. 28, No. 9, pp. 49–56, 1995.
81. G. Sudhir, J.C.M. Lee, and A.K. Jain, "Automatic classification of tennis video for high-level content-based retrieval," in IEEE International Workshop on Content-Based Access of Image and Video Databases, in conjunction with ICCV'98, Bombay, India, 1998.
82. M. Szummer and R.W. Picard, "Indoor-outdoor image classification," in IEEE International Workshop on Content-based Access of Image and Video Databases, in conjunction with ICCV'98, Bombay, India, 1998.
83. B.T. Truong and S. Venkatesh, "Determining dramatic intensification via flashing lights in movies," in IEEE International Conference on Multimedia & Expo, Tokyo, Japan, 2001, pp. 61–64.
84. B.T. Truong, S. Venkatesh, and C. Dorai, "Automatic genre identification for content-based video categorization," in IEEE International Conference on Pattern Recognition, Barcelona, Spain, 2000.
85. S. Tsekeridou and I. Pitas, "Content-based video parsing and indexing based on audio-visual interaction," IEEE Transactions on Circuits and Systems for Video Technology, Vol. 11, No. 4, pp. 522–535, 2001.
86. A. Vailaya and A.K. Jain, "Detecting sky and vegetation in outdoor images," in Proceedings of SPIE: Storage and Retrieval for Image and Video Databases VIII, San Jose, USA, 2000, Vol. 3972.
87. A. Vailaya, A.K. Jain, and H.-J. Zhang, "On image classification: City images vs. landscapes," Pattern Recognition, Vol. 31, pp. 1921–1936, 1998.

88. J. Vendrig and M. Worring, "Systematic evaluation of logical story unit segmentation," IEEE Transactions on Multimedia, Vol. 4, No. 4, pp. 492–499, 2002.
89. Virage. http://www.virage.com.
90. Y. Wang, Z. Liu, and J. Huang, "Multimedia content analysis using both audio and visual clues," IEEE Signal Processing Magazine, Vol. 17, No. 6, pp. 12–36, 2000.
91. T. Westerveld, "Image retrieval: Content versus context," in Content-Based Multimedia Information Access, RIAO 2000 Conference, Paris, France, 2000, pp. 276–284.
92. E. Wold, T. Blum, D. Keislar, and J. Wheaton, "Content-based classification, search, and retrieval of audio," IEEE Multimedia, Vol. 3, No. 3, pp. 27–36, 1996.
93. L. Wu, J. Benois-Pineau, and D. Barba, "Spatio-temporal segmentation of image sequences for object-oriented low bit-rate image coding," Image Communication, Vol. 8, No. 6, pp. 513–544, 1996.
94. P. Xu, L. Xie, S.-F. Chang, A. Divakaran, A. Vetro, and H. Sun, "Algorithms and systems for segmentation and structure analysis in soccer video," in IEEE International Conference on Multimedia & Expo, Tokyo, Japan, 2001, pp. 928–931.
95. M.-H. Yang, D. Kriegman, and N. Ahuja, "Detecting faces in images: A survey," IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 24, No. 1, pp. 34–58, 2002.
96. M.M. Yeung and B.-L. Yeo, "Video content characterization and compaction for digital library applications," in IS&T/SPIE Storage and Retrieval of Image and Video Databases V, 1997, Vol. 3022, pp. 45–58.
97. H.-J. Zhang, A. Kankanhalli, and S.W. Smoliar, "Automatic partitioning of full-motion video," Multimedia Systems, Vol. 1, No. 1, pp. 10–28, 1993.
98. H.-J. Zhang, S.Y. Tan, S.W. Smoliar, and G. Yihong, "Automatic parsing and indexing of news video," Multimedia Systems, Vol. 2, No. 6, pp. 256–266, 1995.
99. T. Zhang and C.-C.J. Kuo, "Hierarchical classification of audio data for archiving and retrieving," in IEEE International Conference on Acoustics, Speech, and Signal Processing, Phoenix, USA, 1999, Vol. 6, pp. 3001–3004.
100. D. Zhong and S.-F. Chang, "Structure analysis of sports video using domain models," in IEEE International Conference on Multimedia & Expo, Tokyo, Japan, 2001, pp. 920–923.
101. Y. Zhong, H.-J. Zhang, and A.K. Jain, "Automatic caption localization in compressed video," IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 22, No. 4, pp. 385–392, 2000.
102. W. Zhou, A. Vellaikal, and C.-C.J. Kuo, "Rule-based video classification system for basketball video indexing," in ACM Multimedia 2000, Los Angeles, USA, 2000.
103. W. Zhu, C. Toklu, and S.-P. Liou, "Automatic news video segmentation and categorization based on closed-captioned text," in IEEE International Conference on Multimedia & Expo, Tokyo, Japan, 2001, pp. 1036–1039.

Cees G.M. Snoek received the M.Sc. degree in business information systems from the University of Amsterdam, The Netherlands, in October 2000. Since January 2001 he has been a Research Assistant at the Intelligent Sensory Information Systems group of the University of Amsterdam, where he is pursuing the Ph.D. degree in computer science. Snoek was a Visiting Scientist at Informedia, Carnegie Mellon University in 2003. His research interests focus on multimedia signal processing and analysis, statistical pattern recognition, and content-based information retrieval, especially when applied in combination for the purpose of semantic multimedia understanding.

Marcel Worring received his M.Sc. degree (honors) and Ph.D. degree, both in computer science, from, respectively, the Free University Amsterdam ('88) and the University of Amsterdam ('93), The Netherlands. He is currently an Associate Professor at the University of Amsterdam. His interests are in multimedia information analysis and systems. He leads several multidisciplinary projects covering knowledge engineering, pattern recognition, image and video analysis, and information space interaction, conducted in close cooperation with industry. In 1998, he was a Visiting Research Fellow at the University of California, San Diego. He has published over 50 scientific papers and serves on the program committee of several international conferences.