The Authoring Metaphor to Machine Understanding of Multimedia

Cees G.M. Snoek

Printing: Febodruk BV, Enschede, The Netherlands.

ISBN 90-5776-143-2

The Authoring Metaphor to Machine Understanding of Multimedia

ACADEMISCH PROEFSCHRIFT

ter verkrijging van de graad van doctor
aan de Universiteit van Amsterdam,
op gezag van de Rector Magnificus prof. mr. P.F. van der Heijden
ten overstaan van een door het College voor Promoties ingestelde commissie,
in het openbaar te verdedigen in de Aula der Universiteit
op woensdag 26 oktober 2005, te 10:00 uur

door

Cornelis Gerardus Maria Snoek

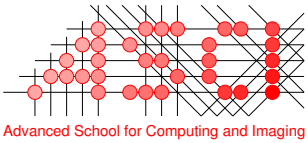
geboren te Purmerend

Promotiecommissie:

Promotor:	Prof. dr ir A. W. M. Smeulders
Co-promotor:	dr M. Worring
Overige leden:	Prof. dr A. Del Bimbo Prof. dr M. de Rijke Prof. dr A. Th. Schreiber dr Th. Gevers dr W. Kraaij

Faculteit: Faculteit der Natuurwetenschappen, Wiskunde en Informatica

The work described in this thesis was supported by the ICES/KIS MIA project and TNO.



The work described in this thesis has been carried out within the graduate school ASCI, at the Intelligent Sensory Information Systems group of the University of Amsterdam. ASCI dissertation series number 116.

Intelligent Sensory Information Systems
University of Amsterdam
The Netherlands

LIEVEN ARBEID

Die 't Ambacht wel verstaet
daer van hij leven moet,

En die 't, niet wel alleen,
maer wel en geerne doet,

Beleeft het grootst geluck
dat yemand kan begeeren.

Hij spoedt, en spoedt met vreugd,
hij wint, en wint met eeren.

O aller staeten staet,
daer voordeel gaet met lust,

En lof en danck met beid'
en werkcken self is rust!

Constantijn Huygens (1596-1687).

Contents

Chapter 1

Introduction

The analog world of the past era has evolved into a digital one. With the digital revolution came the opportunity to create, store, duplicate, and transmit an unprecedented amount of multimedia information, i.e. combinations of text, audio, imagery, and video. This opportunity has been taken with eagerness. It has resulted in huge archives of multimedia data items. In addition to the digital revolution, the Internet rebellion at the fin de siècle of the 20th century offered new ways to connect archives, share, and sell multimedia assets. The volume of unstructured multimedia bits in the digital world is already beyond human reach.

In contrast to humans, machines are experts in handling large quantities of data. While data processing capabilities of machines are superb compared to human standards, data interpretation skills are poor relative to human performance to say the least. For an archive of Hollywood movies, retrieving the *shower scene* from Hitchcock's *Psycho* is a non-trivial task requiring human intervention when the data has not been manually annotated. Since improvement in data processing capabilities of humans is mostly if not exclusively science fiction, advancement in interpretation skills of machines is required. We need to equip machines with understanding of multimedia to aid humans in their struggle to bring order to the digital chaos.

1.1 Motivation

Whether it is a descriptive caption added to holiday pictures or the spoken comment to be synchronized with the action in a soccer broadcast, any multimedia production originates from the mind of an author. An author crafts a multimedia document based on a certain semantic intention. While doing so, the author faces the *intention gap*. We define:

Definition 1.1.1 (Intention Gap) *The lack of coincidence between the information that an author can produce into the multimedia data and the interpretations the user may give to the data.*

Sesame Street

Sesame Street is one of the most successful television series ever produced. The show combines the use of muppets, animation, live action, special effects, text, and music to teach young children about symbolic representations, cognitive processes, and physical and social environments. The show was the first program ever that integrated tailor made multimedia content, children oriented production conventions, and empirical research into a television program [43]. The careful creation process was of major importance to the success of *Sesame Street*.

The gap may be broad for amateur videos, where it is often unclear to outsiders what the author intended with the recorded footage. The gap is narrow for authored multimedia productions that aim for mass communication in the form of storytelling. For these productions, both author and user rely on the professional habits originating from the field of film art [21, 24]. Thus, when an author relies on the guidelines and techniques known from film art for storytelling he or she may succeed in bridging the intention gap, ultimately resulting in an effective multimedia communication.

In this thesis we focus on such professional multimedia for which the intention gap is more or less closed, i.e. produced video. The success of a produced video is highly dependent on the authoring process, see the *Sesame Street* sidebar for an example. While the authoring process is a decisive factor in the human appreciation and understanding of any multimedia production, it is left largely unexplored in the analysis of multimedia by machinery. This thesis aims to fill this lacuna. We use the authoring-driven creation process of professional produced video as a metaphor for machine-based understanding.

1.2 Problem Statement

Understanding multimedia starts with indexing its data at a semantic level. We need to capture the conceptual knowledge present in produced video before we can reason about the author's intention. An author departs from a conceptual idea to produce a video document. Then the author exploits a set of professional conventions and techniques to combine multimedia data into a produced video document. We reverse the authoring process to arrive at a semantic index of produced video. We start with an analysis of multimedia data. We then combine the results by exploiting common style conventions from film art. Once this is achieved we can reason based on context. An authoring-driven analysis methodology ultimately yields an effective semantic index.

For machines, the difficulty in attaching a semantic index to produced video lies in the *semantic gap*. We adapt the definition of [136] and define:

Definition 1.2.1 (Semantic Gap) *The lack of coincidence between the information that machines can extract from the multimedia data and the interpretations the user may give to the data.*

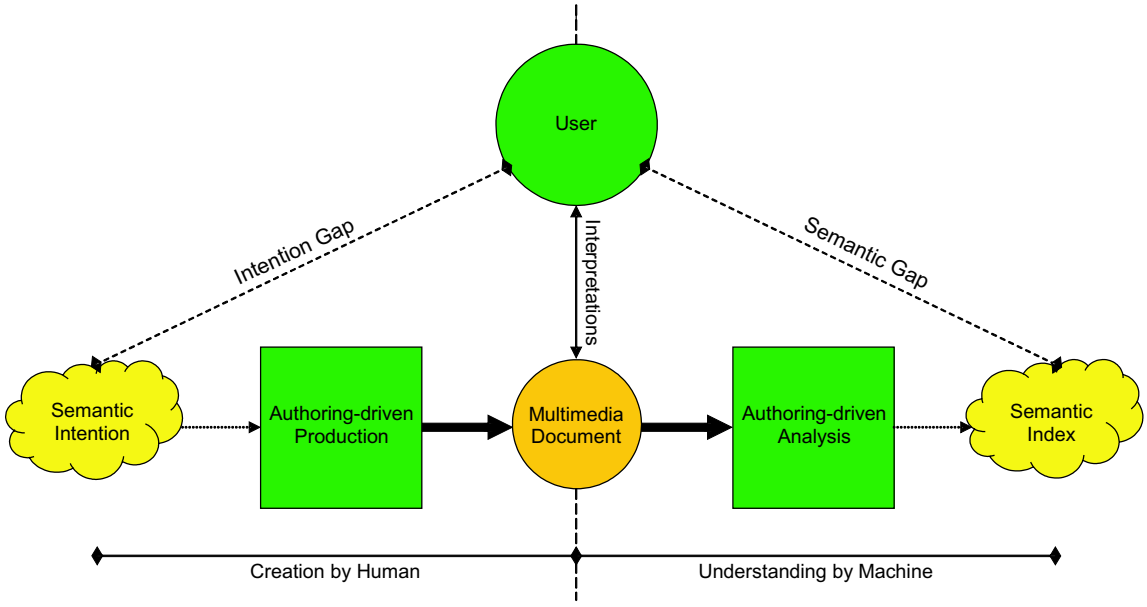


Figure 1.1: We reverse the authoring-driven process of multimedia production to arrive at machine-driven understanding.

The cause for the semantic gap lies in the fact that machines can only compute low level properties of data that have no clear relation with high level conceptual semantics. Hence, the fundamental question that is addressed in this thesis is:

How to bridge the semantic gap for produced video?

In our endeavor to machine understanding of multimedia we defy the rigors and hardships raised by the semantic gap using the authoring metaphor, this is illustrated in Fig. 1.1. Exploitation of the authoring metaphor for bridging the semantic gap raises a number of follow-up questions.

Early methods for semantic indexing of multimedia focused on single modality based analysis only. Such methods proved successful for classifying specific concepts in narrow domains based on a few simple rules. However, it soon became prevalent that scalability and robustness of unimodal rule-based approaches are limited. As a consequence the semantic gap remains. Because an author uses multiple media sources to convey meaning, the authoring metaphor dictates that analysis should exploit all information channels for semantic indexing. A multimodal analysis of produced video is a first step to obtain an effective semantic index. In addition, the complex thoughts of an author are not easily mapped on a few simple decision rules. Rather, this mapping requires quite a lot decision rules. To cope with this need, the use of advanced machine learning techniques is inevitable. Within the authoring metaphor, the first follow-up question is:

How to exploit multimodal analysis in combination with machine learning for multimedia understanding?

An analysis based on content, either in the visual channel, auditory channel, textual channel or combinations thereof using machine learning, throws a first bridge but is not enough to cover the semantic gap. Apart from content, an author uses specific techniques and conventions to compose raw material into an effective presentation. Hence, style conventions should be part of analysis as well. Within the authoring metaphor, the second question is:

How to include the notion of style into semantic multimedia analysis?

An author thinks in concepts, interacting them to strengthen semantic intention. Inclusion of context in the analysis was an important advancement in the research efforts to bridge the semantic gap [98]. Thus, in addition to content and style, context is a factor of importance in bridging the semantic gap. It is crucial to know how these three authoring elements relate to one another. Another issue, given thousands and thousands of concepts that might be present in a produced video, any semantic indexing method should be generic instead of specific. Therefore, the third question is:

How can a semantic analysis of content, style, and context be combined effectively for generic multimedia indexing?

Interpretation of semantics is user-dependent. Thus, eventually user involvement is inevitable for multimedia understanding. Moreover, the semantic gap dictates that only a limited lexicon of semantic concepts can be learned automatically. Hence, users should be offered other means, besides learning, to retrieve the semantics from multimedia data. In this respect, similarity is of interest. Interaction, learning, and similarity are identified in [136] as key techniques to bridge the semantic gap. We aim for their combination, which results in the fourth question:

How should we exploit the combination of learning, similarity, and interaction for effective multimedia retrieval?

Answering the four questions will help bridging the semantic gap. Then the problem of evaluation remains. The field of multimedia understanding in recent years has witnessed a proliferation of methods, often evaluated on specific and small data sets. As a result, experiments are non-repeatable; making it hard to judge whether approaches are truly promising. To counter this trend, the American National Institute of Science and Technology (NIST) initiated the TREC Video Retrieval Evaluation (TRECVID) [102]. The aim of the benchmark is to promote progress in content-based retrieval from digital video archives via open, metrics-based evaluation using a common large data set. We joined this initiative, by evaluating a substantial part of the authoring metaphor within the TRECVID benchmark. Because of benchmarks like TRECVID methods for multimedia understanding can be valued on their relative merit.

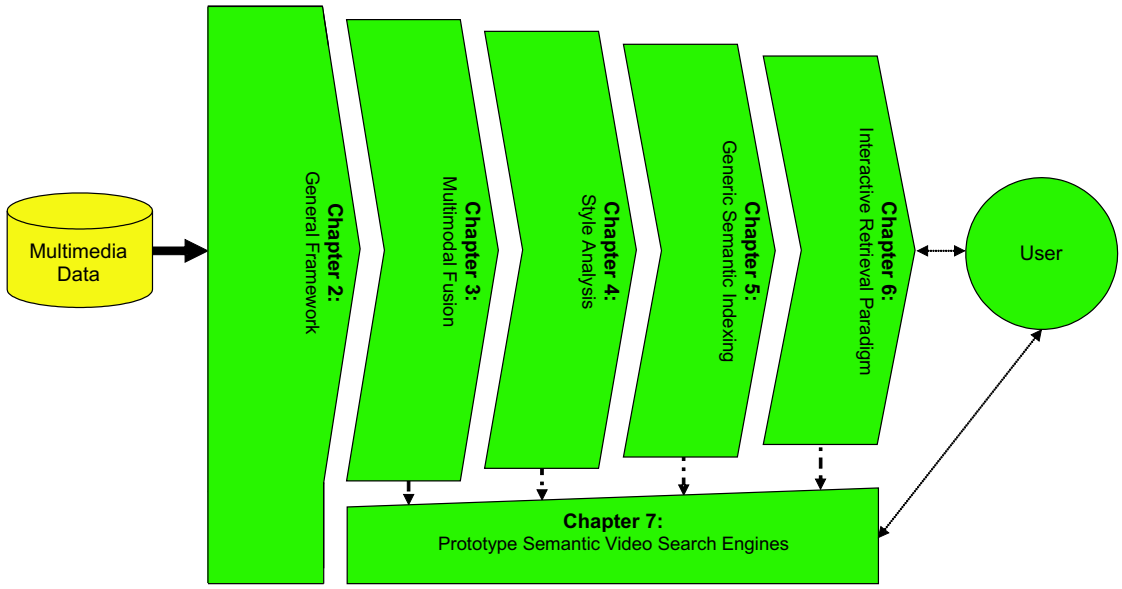


Figure 1.2: Outline of the work addressed in this thesis.

1.3 Organization

We present in this thesis a step-by-step extrapolation of the authoring metaphor to arrive at automatic indexing of semantic concepts in produced video archives. This process is illustrated in Fig. 1.2.

The thesis starts in Chapter 2 with the definition of a general framework for machine understanding of produced video. Viewing a video document as the result of an authoring process, allows for seamless integration of the different modalities involved. Moreover, it allows to structure the index methodologies and tools currently available in literature. We end the Chapter with a number of research issues that are addressed in the remainder of the thesis. Integration of modalities is a crucial part of any multimodal video indexing system. Chapter 3 addresses the specific problems for modality fusion, i.e. synchronization and the inclusion of temporal context. The proposed Time Interval Multimedia Event framework tackles these problems by using a representation based on time intervals. The representation allows integrating multiple modalities and facilitates the usage of several machine learning approaches for semantic indexing. In Chapter 4 we arrive at the heart of the authoring metaphor, where we introduce a framework for produced video indexing based on style analysis. By combining style detectors into a classifier ensemble, the framework facilitates robust classification of semantic concepts in produced video. A unifying architecture of our work addressed in Chapters 2, 3, and 4 together with recent advances in the field of multimedia understanding is presented in Chapter 5. The proposed semantic value chain follows the authoring metaphor by successively analyzing video on content, style, and context aspects. While doing so, it allows for semantic video indexing in a generic fashion. In Chapter 6 we present a paradigm for interactive video retrieval. We build the paradigm on three principles: learning of a lexicon of semantic

concepts, multimedia data similarity, and user interaction with a video search engine. Finally, in Chapter 7 we present four case studies that illustrate the practical use of the theory derived in this thesis into several semantic video search engine prototypes.

Chapter 2

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

Efficient and effective handling of video documents depends on the availability of indexes. Manual indexing is unfeasible for large video collections. In this Chapter 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.

*Published in *Multimedia Tools and Applications*, 25(1):5-35, 2005.

2.1 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 [103] 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 [55].

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 [32,53,108,149,153,175,181], auditory [40,50,87,105,106,110, 165], or textual modality [26,62,182]. Good books [46,58] and review papers [22,27] 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 [7, 13, 38, 65, 96, 111, 126] or commercially exploited, e.g. [33,114,160].

Ultimately the *which* question should be answered with content-based segment descriptors, for instance those proposed in the MPEG-7 standard [91,92], 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 [162]. 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 [22, 27, 162]. Combined they form a complete overview of the field of multimodal video indexing.

The multimodal video indexing framework is defined in Section 2.2. 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 2.3. In Section 2.4 the role of conversion and integration in multimodal analysis is discussed. An overview of the index types that can be distinguished, together with some exam-

ples, will be given in Section 2.5. Finally, in Section 2.6 we end with a perspective on open research questions.

2.2 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 [101] 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 2.2.1 (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 2.2.1 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.

2.2.1 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 [35]. 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 [69], 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.

2.2.2 Content

The content perspective relates segments to elements that an author uses to create a video document. The following elements can be distinguished [21]:

- *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.

2.2.3 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

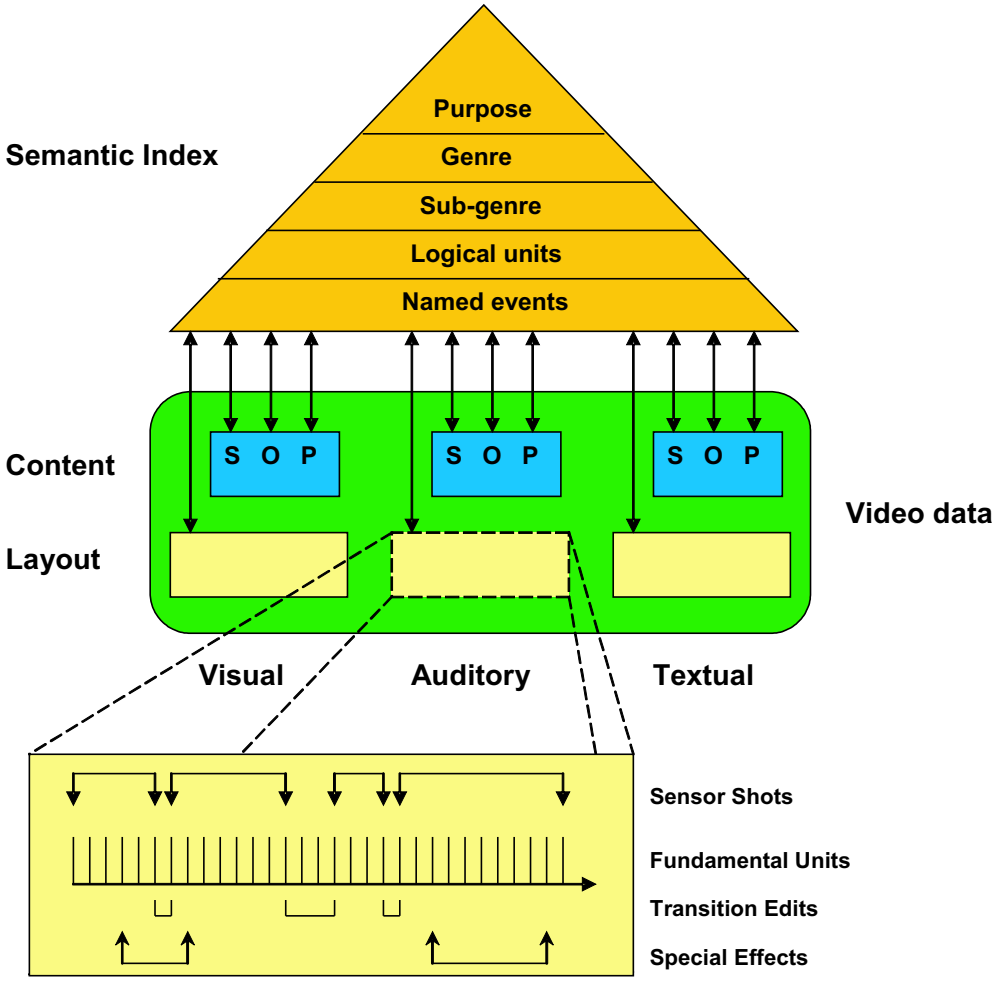


Figure 2.1: A unifying framework for multimodal video indexing based on an author's perspective. The letters S, O, P stand for setting, objects, and people. An example layout of the auditory modality is highlighted, the same holds for the others.

shots are smooth" [21]. For the visual modality abrupt cuts, or gradual transitions[†], 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 [21]. 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.

[†]A gradual transition actually contains pieces of two camera shots, for simplicity we regard it as a separate entity.

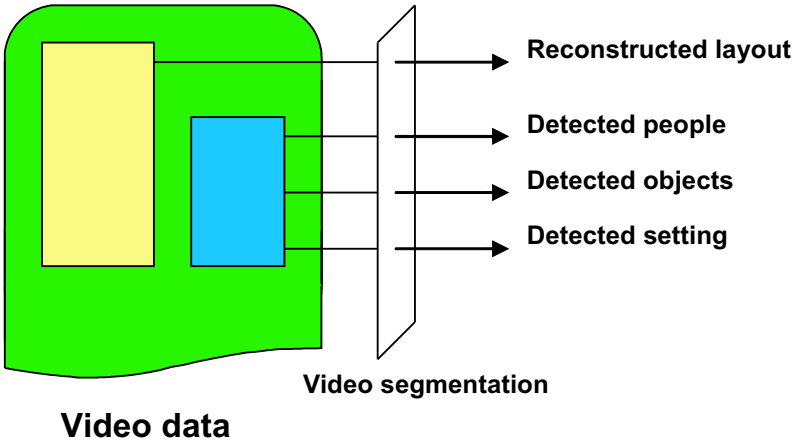


Figure 2.2: Data flow in unimodal video document segmentation.

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" [21]. 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 Fig. 2.1. It forms the basis for our discussion of state-of-the-art indexing techniques.

2.3 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 Fig. 2.2.

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, or features derived from layout and content elements. According to [68] 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 this chapter. 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 [68];
- *Decision Tree*: assigns a pattern to a class based on a hierarchical division of feature space [68];
- *k-Nearest Neighbor*: assigns a pattern to the majority class among the k patterns with smallest distance in feature space [68];
- *Hidden Markov Model (HMM)*: assigns a pattern to a class based on a sequential model of state and transition probabilities [86, 118];

Statistical classifiers are also well suited for multimodal classification. This aspect of pattern recognition will be highlighted in Section 2.4.2. We will now first discuss the reconstruction of layout and content elements.

2.3.2 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*[‡].

[‡]As an ironic legacy from early research on video parsing, this is also referred to as scene-change detection.

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 [27] 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 [174]. 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 [27], 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 [105] 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 [177]. 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 [80] 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 [80] 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 [177] 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 [86]. 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 [86]. 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.

2.3.3 Content Segmentation

In Section 2.2.2 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.

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 [171]. 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 [171] for a comprehensive and critical survey of current face detection methods. From all methods currently available the one proposed by Rowley in [120] performs the best [112]. 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 [109]. Here the matching is performed using single images, and the method is capable to recognize faces under varying pose. In [15] 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 [126]. Reported results [15,109,126] 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 [89] 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 [45] 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 [177] 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 [80]. 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. [94, 106]. A generic speaker identification system consisting of three modules is presented in [106]. 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 [126], 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 [19]. 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.

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 [129], 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 modeling 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 [100] 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. [9, 168].

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 [165, 177] 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 [86]. 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.

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 [136] 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 [150] 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 [157]. 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 [156].

Detecting setting based on auditory information, can be achieved by detecting specific environmental sound patterns. In [165] 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 [177] 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 [30] 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.

2.4 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 Fig. 2.3.

2.4.1 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

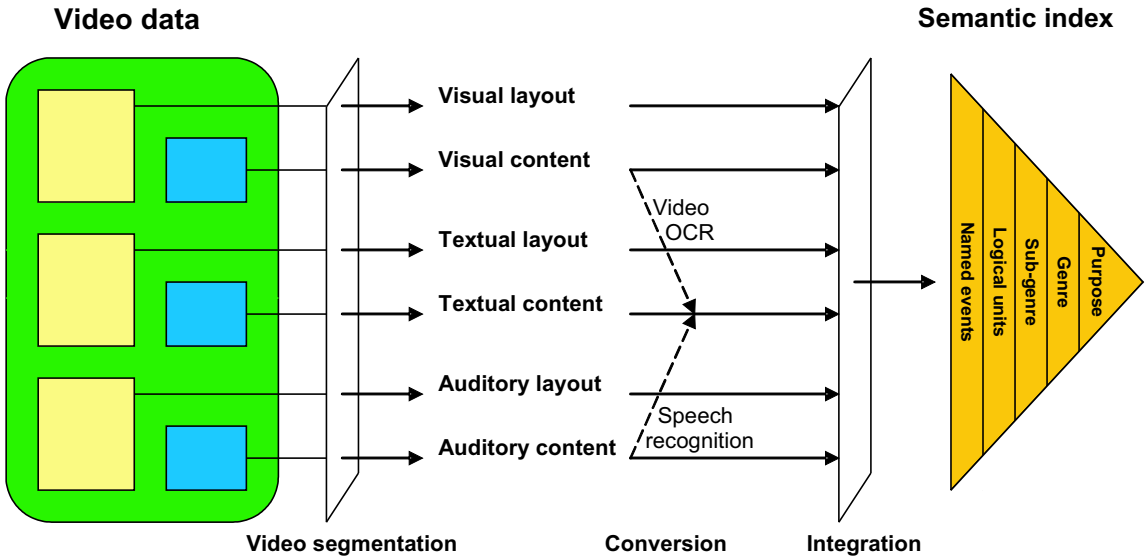


Figure 2.3: Role of conversion and integration in multimodal video document analysis.

text. Video Optical Character Recognition (OCR) methods for detection of text in video frames can be divided into component-based, e.g. [133], or texture-based methods, e.g. [81]. A method utilizing the DCT coefficients of compressed video was proposed in [179]. 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.

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 [27]. 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.

2.4.2 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 align-

Table 2.1: An overview of various integration methods.

	Content Segmentation		Classification Method		Processing Cycle	
	Symmetric	Asymmetric	Statistical	Knowledge	Iterated	Non-Iterated
[7]	✓		✓			✓
[13]		✓		✓	✓	
[37]	✓		✓			✓
[38]	✓		✓			✓
[44]	✓			✓		✓
[65]	✓		✓			✓
[65]		✓	✓			✓
[70]	✓		✓			✓
[71]	✓		✓			✓
[76]	✓		✓			✓
[93]	✓			✓		✓
[96]	✓		✓		✓	
[111]	✓			✓		✓
[124]	✓			✓		✓
[126]	✓		✓			✓
[148]		✓		✓	✓	
[154]	✓			✓		✓
[163]	✓		✓			✓

ment 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. [65]. In [7,38] 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 [70]. 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.

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 2.1.

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 [44, 93, 111, 124, 154]. In [154] 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 [7, 37, 38, 65, 70, 71, 76, 96, 126, 163]. An example of a symmetric, non-iterated statistical integration method is the Name-It system presented in [126]. 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 [7, 37, 38, 65]. 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 [65] 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 [65], 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 [96]. 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 [107]. 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 [71], however that method doesn't support iteration.

In contrast to the above symmetric methods, an asymmetric approach is presented in [65]. 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 [13]. 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 [148]. 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) [36] has become a popular means for integration [76, 163]. 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 [76] 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 [68]. 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.

2.5 Semantic Video Indexes

The methodologies described in Section 2.4 have been applied to extract a variety of the different video indexes described in Section 2.2.1. 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. Fig. 2.4 presents an overview of all indexes and the methods in literature which can derive them.

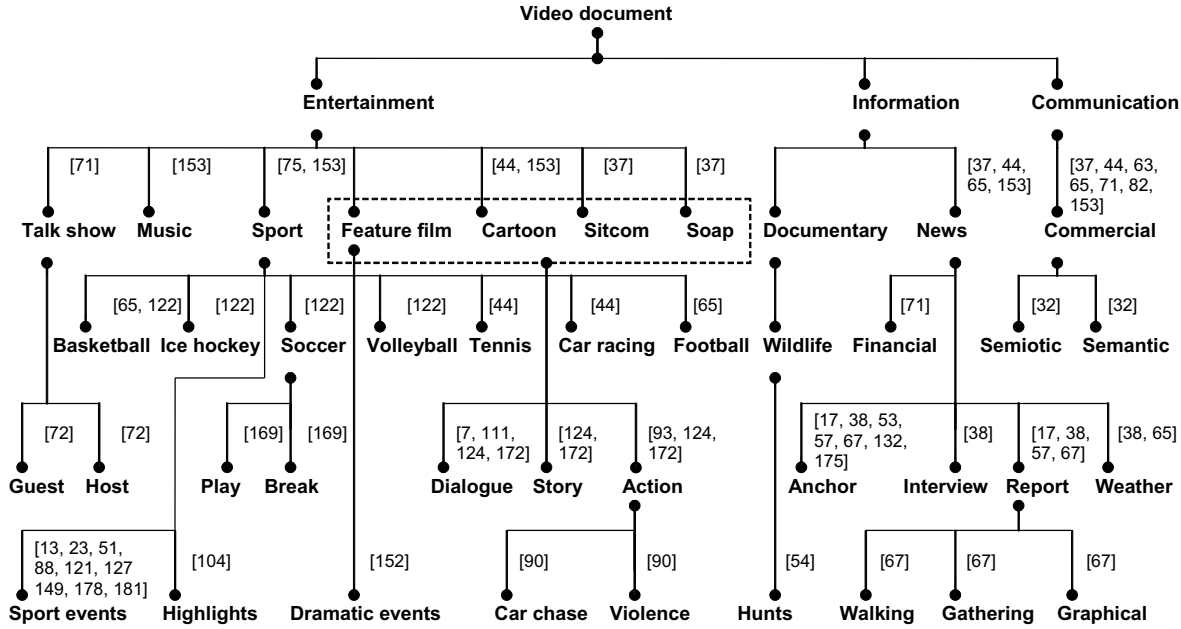


Figure 2.4: Semantic index hierarchy with instances as found in literature. From top to bottom instances from genre, sub-genre, logical units, and named events. The dashed box is used to group similar nodes.

2.5.1 Genre

"Editing is an important stylistic element because it affects the overall rhythm of the video document" [21]. 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 [63] 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 [82], 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 [153] 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 [37] 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 [75], 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 [44, 65, 71]. In [65] 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 2.4.2.

The same modalities are used in [44]. 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, [71] 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 2.4.2. In its current implementation the presented system classifies segments as either part of a talk show, commercial or financial news.

2.5.2 Sub-genre

Research on indexing sub-genres, or specific instances of a genre, has been geared mainly towards sport videos [44, 65, 122] and commercials [32]. 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 [122]: 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 [44, 65], as noted in Section 2.5.1.

An approach to index commercial videos based on semiotic and semantic properties is presented in [32]. 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.

2.5.3 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 [159]. 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 [7]. 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 [93]. 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 [124]. 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 [124], a unimodal approach based on the visual information source is used in [172] 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 [111]. 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 [53, 132, 175]. In [53, 132] also the restricted position and size of detected faces is used.

Another approach for the detection of anchor shots is taken in [17, 57, 67]. Repetition of visually similar anchor shots throughout the news broadcast is exploited. To refine the classification of the similarity measure used, [17] 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 [67] 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 [57] 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 [17,67], 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 [38] 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 [72] 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 [169] 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.

2.5.4 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 [54]. 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 [90], 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 [67]. 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 [152]. An algorithm is presented that detects flashlights, 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 [104] 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 [13] 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 [88, 149, 178], soccer [23, 51], baseball [121, 178] and basketball [127, 181] 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 [121], 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.

2.5.5 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 Fig. 2.4, 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. News reports and interviews are alternated by anchor persons introducing, and narrating about, the various 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.

2.6 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 explo-

ration 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.

Chapter 3

Multimedia Event-Based Video Indexing using Time Intervals*

We propose the Time Interval Multimedia Event (TIME) framework as a robust approach for classification of semantic events in multimodal video documents. The representation used in TIME extends the Allen temporal interval relations and allows for proper inclusion of context and synchronization of the heterogeneous information sources involved in multimodal video analysis. To demonstrate the viability of our approach, it was evaluated on the domains of soccer and news broadcasts. For automatic classification of semantic events, we compare three different machine learning techniques, i.e. C4.5 decision tree, Maximum Entropy, and Support Vector Machine. The results show that semantic video indexing results significantly benefit from using the TIME framework.

*Published in *IEEE Transactions on Multimedia*, 7(4):638-647, 2005.

3.1 Introduction

Management of digital video documents is becoming more and more problematic due to the ever growing size of content produced. For easy management a semantic index describing the different events in the content of the document is indispensable. Since manual annotation is unfeasible, because of its tedious and cumbersome nature, automatic video indexing methods are necessary.

In general, automatic indexing methods suffer from the *semantic gap* or the lack of coincidence between the extracted information and its interpretation by a user, as recognized for image indexing in [136]. Video indexing has the advantage that it can profit from combined analysis of visual, auditory, and textual information sources. For multimodal indexing, two problems have to be unravelled. Firstly, when integrating analysis results of different information channels, difficulties arise with respect to synchronization. The synchronization problem is typically solved by converting all modalities to a common layout scheme [142], e.g. camera shots, hereby ignoring the layout of the other modalities. This introduces the second problem, namely the difficulty to properly model context, i.e. how to include clues that do not occur at the exact moment of the semantic event of interest? When synchronization and context have been solved, multimodal video indexing might be able to bridge the semantic gap to some extent.

Existing methods for multimodal integration can be grouped into knowledge based approaches [13, 44] and statistical approaches [56, 65, 84, 96, 180]. The former approaches typically combine the output of different multimodal detectors into a rule based classifier. In [13] for example, the authors first analyze the textual channel for the occurrence of specific keywords that have a relation with a semantic event in American football. This results in a time interval where a possible event has taken place. The visual information of this time interval is then used for final classification. The drawback of this two stage approach is the dependence on the first stage. If the textual stream detector fails, no event is detected. To limit this model dependency, and improve the robustness, a statistical approach seems more promising. Various statistical frameworks can be exploited for multimodal integration. Recently there has been a wide interest in applying the Dynamic Bayesian Network (DBN) framework for multimodal integration [65, 96]. Other multimodal statistical frameworks that were proposed include the use of C4.5 decision trees [180], Maximum Entropy [56], and Support Vector Machines [84]. However, all of these frameworks suffer from the problems of synchronization and context, identified above. Furthermore, they lack satisfactory inclusion of the textual modality. Therefore, a new framework is needed.

In this Chapter we propose the Time Interval Multimedia Event (TIME) framework which explicitly handles context and synchronization and, as it is based on statistics, yields a robust approach for multimodal integration.

To demonstrate the viability of our approach for video indexing of semantic events we provide a systematic evaluation of three statistical classifiers, using TIME, and discuss their performance on the domains of soccer and news broadcasts. The soccer domain was chosen because events occur infrequently and in an unpredictable manner. Hence, contextual clues are important for reliable detection. In contrast to soccer, the

news domain is far more structured. Here, synchronization of the different information sources is more important than context for accurate event detection.

The rest of this Chapter is organized as follows. First we discuss related work, with respect to the domains we consider. Then we proceed with the introduction of the TIME framework in Section 3.3, discussing both representation and classification. In Section 3.4 we discuss the detectors used for classification of various semantic events in soccer and news video. Experimental results are presented in Section 3.5.

3.2 Related Work in Soccer and News Analysis

The classification methods introduced in the introduction have been used in various applications. For an extensive overview we refer to Chapter 2. We focus here on the soccer and news domain.

In literature several methods for automatic soccer analysis have been proposed, e.g. [11, 39, 79, 173]. Most methods are based on analysis of the visual modality only. One of the first reported methods was presented in [173]. The authors focus on visualization of ball and player tracks using mosaics. However, no experiments in semantic event detection were demonstrated. More recently, methods were proposed that try to narrow the semantic gap based on a correlation between advanced visual detectors and semantic concepts. In [11, 79] camera based detectors are proposed, exploiting the relation between the movement of the ball and the camera. A slow-motion replay detector, among others, is proposed in [39] as a strong indicator for an event of importance that happened beforehand. For combination of the visual detectors a statistical DBN is used in [11, 79], whereas [39] exploits a knowledge based approach.

In contrast to soccer event detection methods, which are still mainly based on visual analysis, the state-of-the-art in news analysis is already based on multimodal analysis [18, 38, 67, 84]. In [18] anchor shots and graphical shots are detected based on similarity and motion. The remaining shots are classified as news footage and are annotated with text extracted from a Video Optical Character Recognition module and a speech recognition module. A similar approach is proposed in [67], besides anchors, graphics, and report events, they detect gathering and walking events by exploiting face statistics. Manually added captions are processed with a named entity recognizer to attach more semantics to the detected events. By exploiting the fixed structure of a news broadcast in combination with similarity, motion, and audio detectors, the authors of [38] are able to detect anchors, monologues, report footage and weather forecasts. Weather reports are also detected in [84], the authors combine text and image detectors and exploit combination strategies to improve classification accuracy. For the integration phase, again, a differentiation between knowledge based [18, 67] and statistical methods [38, 84] can be made.

For both domains problems arise when contextual information is to be included in the analysis and the various information sources have to be synchronized. In soccer for example, contextual clues like replays and distinguishing camera movement don't appear at the exact moment of the event, therefore the timing has to be estimated.

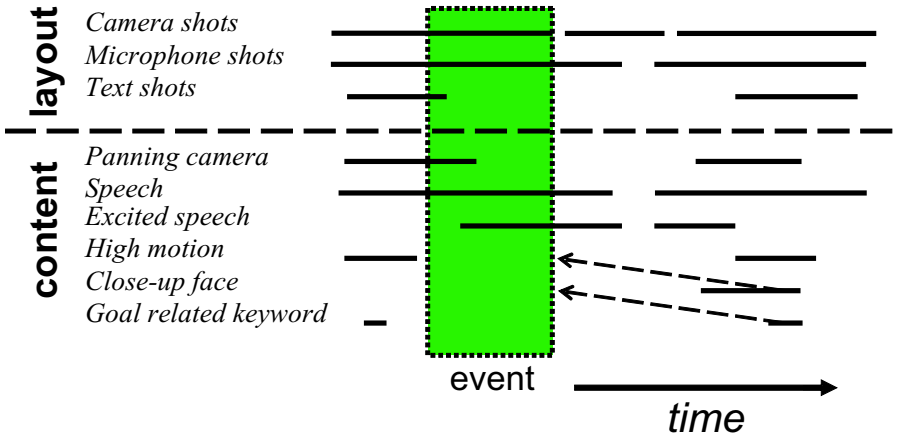


Figure 3.1: Detector based segmentation of a multimodal soccer video document into its layout and content elements with a goal event (box) and contextual relations (dashed arrows).

In news on the other hand, there is a clear relation between the visibility moment of overlayed text and the introduction of a speaker, i.e. it is unlikely that the overlay will appear at the end of the camera shot that views the speaker. Hence, their synchronization should be relative to each other. To tackle the problems of proper synchronization and inclusion of contextual clues for multimodal video analysis we propose the statistical TIME framework.

3.3 Multimedia Event Classification Framework

We view a video document from the perspective of its author [142]. Based on a predefined semantic intention, an author combines certain multimedia layout and content elements to express his message. For analysis purposes this authoring process should be reversed. Hence, we start with reconstruction of layout and content elements. To that end, discrete detectors, indicating the presence or absence of specific layout and content elements, are often the most convenient means to describe the layout and content. This has the added advantage that detectors can be developed independently of one another. To combine the resulting detector segmentations into a common framework, some means of synchronization is required. To illustrate, consider Fig. 3.1. In this example a soccer video document is represented by various time dependent detector segmentations, defined on different asynchronous layout and content elements. At a certain moment a goal occurs. Clues for the occurrence of this event are found in the detector segmentations that have a value within a specific position of the time-window of the event, e.g. excited speech of the commentator. But also in contextual detector segmentations that have a value before, e.g. a camera panning towards the goal area, or after the actual occurrence of the event, e.g. the occurrence of the keyword *score* in the time stamped closed caption. Clearly, in terms of the theoretical framework, it doesn't matter exactly what the detector segmenta-

tions are. What is important is that we need means to express the different visual, auditory, and textual detector segmentations into one fixed representation without loss of their original layout scheme.

Hence, for automatic classification of a semantic event, ω, we need to grasp a video document into a common pattern representation. In this section we first consider how to represent such a pattern, x, using multimodal detector segmentations and their relations, then we proceed with statistical pattern recognition techniques that exploit this representation for classification using varying complexity.

3.3.1 Pattern Representation

Applying layout and content detectors to a video document results in various segmentations, we define:

Definition 3.3.1 (TIME Segmentation) *Decomposition of a video document into one or more series of time intervals, τ, based on a set of multimodal detectors.*

To model synchronization and context, we need means to express relations between these time intervals. Allen showed that thirteen relationships are sufficient to model the relationship between any two intervals. To be specific, the relations are: *precedes*, *meets*, *overlaps*, *starts*, *during*, *finishes*, *equals*, and their inverses, identified by adding *_i* to the relation name [8]. For practical application of the Allen time intervals two problems occur. First, in video analysis exact alignment of start- or endpoints seldom occurs due to noise. Second, two time intervals will always have a relation even if they are far apart in time. To solve the first problem a fuzzy interpretation was proposed in [6]. The authors introduce a margin, T_1, to account for imprecise boundary segmentations, explaining the fuzzy nature. The second problem only occurs for the relations *precedes* and *precedes_i*, as for these the two time intervals are disjunct. Thus, we introduce a range parameter, T_2, which assigns to two intervals the type *NoRelation* if they are too far apart in time. Hence, we define:

Definition 3.3.2 (TIME Relations) *The set of fourteen fuzzy relations that can hold between any two elements from two segmentations, τ_1 and τ_2, based on the margin T_1 and the range parameter T_2.*

Obviously the new relations still assure that between two intervals one and only one type of relation exists. The difference between standard Allen relations and TIME relations is visualized in Fig. 3.2.

Since TIME relations depend on two intervals, we choose one interval as a reference interval and compare this interval with all other intervals. Continuing the example, when we choose a camera shot as a reference interval, the goal can be modeled by a swift camera pan that *starts* the current camera shot, excited speech that *overlaps_i* the camera shot, and a goal related keyword in the closed caption that *precedes_i* the camera shot within a range of 6 seconds. This can be explained because of the time lag between actual occurrence of the event and its mentioning in the closed caption. Although a panning camera, excited speech, and a goal related keyword are possible important cues for a goal event, it is their combination with specific TIME relations

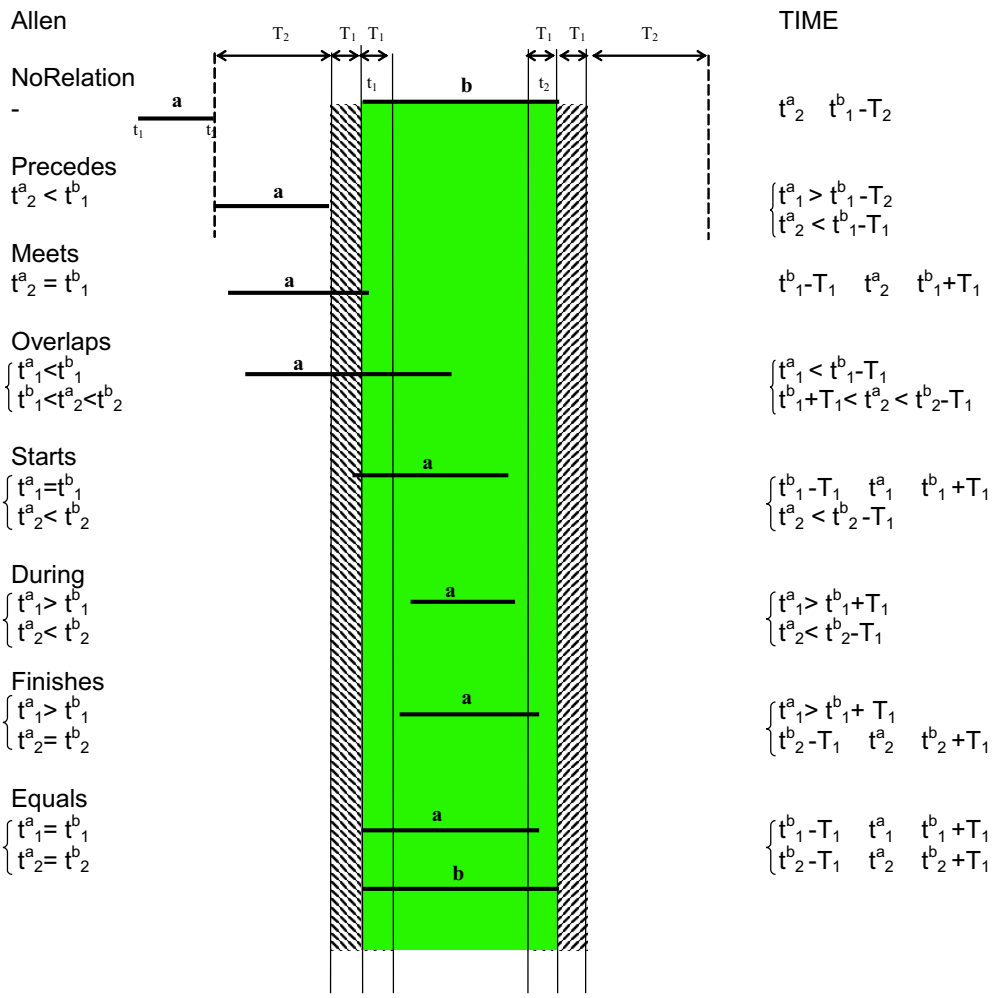


Figure 3.2: Overview of the differences between exact Allen relations and TIME relations, extended from [6].

that makes it key information with respect to the semantics. Also note that the interval based TIME relations have a clear advantage over point based representations, since the relative ordering of segmentations is preserved, and the relations don't suffer from variable lengths between various segmentations. Moreover, by combining TIME segmentations and TIME relations it becomes possible to represent events, context, and synchronization into one common framework. Hence, we define:

Definition 3.3.3 (TIME Representation) *Model of a multimedia pattern x based on the reference interval τ_{ref}, and represented as a set of n TIME relations, with d TIME segmentations.*

In theory, the number of TIME relations, n, is bounded by the number of TIME segmentations, d. Since, every TIME segmentation can be expressed as a maximum of fourteen TIME relations with the fixed reference interval, the maximum number of

TIME relations in any TIME representation is equal to $14(d-1)$. In practice, however, a subset can be chosen, either by feature selection techniques [68], experiments, or domain knowledge.

With the TIME representation we are able to combine layout and content elements into a common framework. Moreover, it allows for proper modeling of synchronization and inclusion of context as they can both be expressed as time intervals.

3.3.2 Pattern Classification

To learn the relation between a semantic event ω, and corresponding pattern x, we exploit the powerful properties of statistical classifiers. In standard pattern recognition, a pattern is represented by features. In the TIME framework a pattern is represented by related detector segmentations.

The statistical classification process is composed of two phases: training and testing. In the first phase, the optimal pattern configuration of relations is learned from the training data. In the second phase, the statistical classifier assigns the most probable event to a pattern based on the detected segmentations and their TIME relations. To prevent overtraining of the classifier, patterns in the testing phase should be drawn from an independent data set.

In literature a varied gamut of statistical classifiers is proposed, see [68] for an excellent overview. For our purpose, classification of semantic events in video documents, a classifier should adhere to the following principles:

- *Binary representation*: since TIME relations are binary by default, the statistical classifier should be able to handle a binary pattern representation;
- *No independence assumption*: since there is a clear dependency between clues found in different modalities, a statistical classifier should not be based on an independence assumption;
- *Learn from few examples*: since the events of importance in a video can be limited, the statistical classifier should be able to learn from few examples;

Three statistical classifiers with varying complexity, adhering to the predefined principles, will be discussed. We start with the C4.5 decision tree [116], then we proceed with the Maximum Entropy framework [16, 73], and finally we discuss classification using a Support Vector Machine [158].

C4.5 Decision Tree

The C4.5 decision tree learns from a training set the individual importance of each TIME relation by computing the gain ratio [116]. Based on this ratio a binary tree is constructed where a leaf indicates a class, and a decision node chooses between two subtrees based on the presence of some TIME relation. The more important a TIME relation is for the classification task at hand, the closer it is located near the root of the tree. Because the relation selection algorithm continues until the entire training set is completely covered, some pruning is necessary to prevent overtraining.

Decision trees are considered suboptimal for most applications [68]. However, they form a nice benchmark for comparison with more complex classifiers and have the added advantage that they are easy to interpret.

Maximum Entropy

Whereas a decision tree exploits individual TIME relations in a hierarchical manner, the Maximum Entropy (MaxEnt) framework exploits the TIME relations simultaneously. In MaxEnt, first a model of the training set is created, by computing the expected value, E_{train}, of each TIME relation using the observed probabilities $\tilde{p}(x, \omega)$ of pattern and event pairs, [16]. To use this model for classification of unseen patterns, we require that the constraints for the training set are in accordance with the constraints of the test set. Hence, we also need the expected value of the TIME relations in the test set, E_{test} [16]. The complete model of training and test set is visualized in Fig. 3.3. We are left with the problem of finding the optimal reconstructed model, p^*, that finds the most likely event ω given an input pattern x, and that adheres to the imposed constraints. From all the possible models, the maximum entropy philosophy dictates that we select the one with the maximum entropy. It is shown in [16] that there is always a unique model $p^*(\omega|x)$ with maximum entropy, and that $p^*(\omega|x)$ has a form equivalent to:

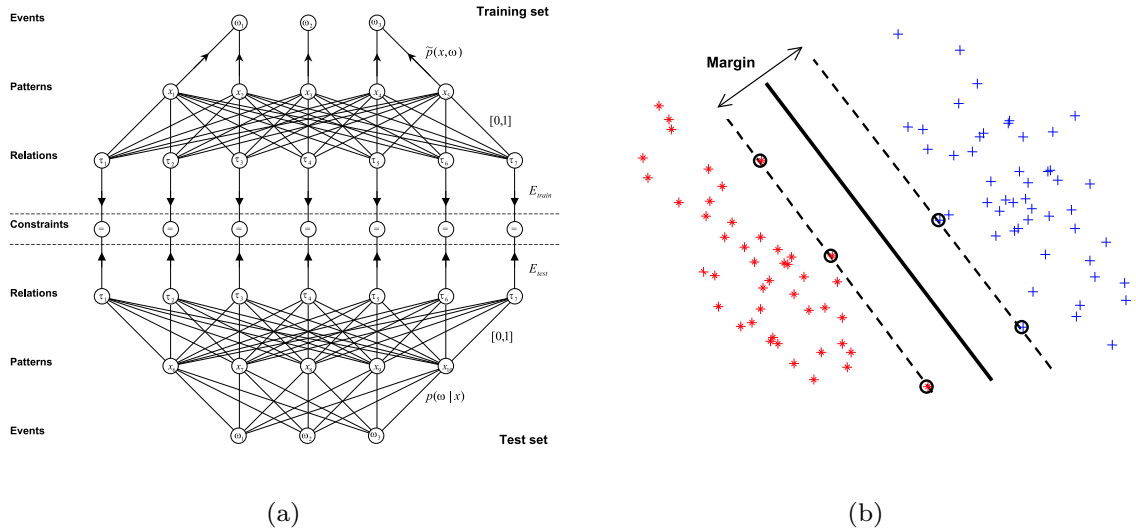
$$p^*(\omega|x) = \frac{1}{Z} \prod_{j=1}^{n} \alpha_j^{\tau_j(x,\omega)} \tag{3.1}$$

where α_j is the weight for TIME relation τ_j and Z is a normalizing constant, used to ensure that a probability distribution results. The values for α_j are computed by the *Generalized Iterative Scaling* (GIS) [34] algorithm. Since GIS relies on both E_{train} and E_{test} for calculation of α_j, an approximation proposed by [77] is used that relies only on E_{train}. This allows to construct a classifier that depends completely on the training set. The automatic weight computation is an interesting property of the MaxEnt classifier, since it is very difficult to accurately weigh the importance of individual detectors and TIME relations beforehand.

Support Vector Machine

The Support Vector Machine (SVM) classifier follows another approach. Each pattern x is represented in a n-dimensional space, spanned by the TIME relations. Within this relation space an optimal hyperplane is searched that separates the relation space into two different categories, ω, where the categories are represented by $+1$ and -1 respectively. The hyperplane has the following form: $\omega|(\mathbf{w} \cdot x + b)| \geq 1$, where $\mathbf{w}$ is a weight vector, and b is a threshold. A hyperplane is considered optimal when the distance to the closest training examples is maximum for both categories. This distance is called the margin, see the example in Fig. 3.3.

The problem of finding the optimal hyperplane is a quadratic programming prob-



(a) (b)

Figure 3.3: (a) Simplified visual representation of the Maximum Entropy framework. Constraints, imposed by the relations, for the training set should be in accordance with those for the test set. From all possible models the one with maximum entropy is chosen. (b) Visual representation of the Support Vector Machine framework. Here a two-dimensional relation space consisting of two categories is visualized. The solid bold line is chosen as optimal hyperplane because of the largest possible margin. The circled data points closest to the optimal hyperplane are called the support vectors.

lem of the following form [158]:

$$\min_{\mathbf{w},\xi} \left\{ \frac{1}{2}\mathbf{w} \cdot \mathbf{w} + C\Big(\sum_{i=1}^{l} \xi_i \Big) \right\} \tag{3.2}$$

Under the following constraints:

$$\omega|(\mathbf{w} \cdot x_i + b)| \geq 1 - \xi_i, \quad i = 1, 2, \ldots, l \tag{3.3}$$

Where C is a parameter that allows to balance training error and model complexity, l is the number of patterns in the training set, and ξ_i are slack variables that are introduced when the data is not perfectly separable. These slack variables are useful when analyzing multimedia, since results of individual detectors typically include a number of false positives and negatives.

3.4 Multimodal Video Analysis

We consider two domains for analysis, namely soccer and news. These domains were chosen because they allow to evaluate both the importance of context and proper synchronization.

Important events in a soccer game are scarce and occur more or less random. Examples of such events are goals, penalties, yellow cards, red cards, and substitutions. We define these events as follows:

- *Goal*: the entire camera shot showing the actual goal;
- *Penalty*: beginning of the camera shot showing the foul until the end of the camera shot showing the penalty;
- *Yellow card*: beginning of the camera shot showing the foul until the end of the camera shot that shows the referee with the yellow card;
- *Red card*: beginning of the camera shot showing the foul until the end of the camera shot that shows the referee with the red card;
- *Substitution*: beginning of the camera shot showing the player who goes out, until the end of the camera shot showing the player who comes in;

These events are important for the game and therefore the author adds contextual clues to make the viewer aware of the events. For accurate detection of events, this context should be included in the analysis.

In contrast to soccer, a news broadcast is far more structured. Each episode, the author carefully edits the layout and content elements, strictly adhering to the predefined format of events in the news show. Most important events in a news broadcast are the news stories. However, due to large variability in content, they are hard to model. Therefore, we focus on events that are more uniform in content and are useful for analysis of news structure. Examples of such events are reporting anchors, monologues, split-view interviews, and weather reports. We define these events as follows:

- *Reporting anchor*: the entire camera shot showing a news anchor talking to the camera;
- *Monologue*: the entire camera shot showing a single person, not a reporting anchor or weather reporter, talking for a while;
- *Split-view interview*: the entire camera shot showing both a news anchor and an on-site reporter in dialogue;
- *Weather report*: the entire camera shot showing a weather reporter talking about the weather forecast;

For analysis, the careful editing of the events should be taken into account by means of proper synchronization.

In this section we will elaborate on the TIME segmentations and TIME relations used for both soccer and news analysis. Some of the detectors, used for the segmentation, are domain specific. It allows to integrate domain knowledge, but as these are learned and not strict they are more robust than domain knowledge captured in

Table 3.1: TIME representation for soccer analysis. T_2 indicates the contextual range used by the precedes and precedes_i relations.

TIME segmentation	TIME relations	T_2 (s)
Camera work	*during*	
Person	*during*	
Close-up	*precedes_i*	0 - 40
Goal keyword	*precedes_i*	0 - 6
Card keyword	*precedes_i*	0 - 6
Substitution keyword	*precedes_i*	0 - 6
Excitement	*All relations*	0 - 1
Info block statistics	*precedes_i*	20 - 80
Person block statistics	*precedes_i*	20 - 50
Referee block statistics	*precedes_i*	20 - 50
Coach block statistics	*precedes_i*	20 - 50
Goal block statistics	*precedes_i*	20 - 50
Card block statistics	*precedes_i*	20 - 50
Substitution block statistics	*during*	
Shot length	*during*	

rules. Other detectors were chosen based on reported robustness and training experiments. The parameters for individual detectors were found by experimentation using the training set. Combining all TIME segmentations with all TIME relations results in an exhaustive use of relations, we therefore use a subset to prevent a combinatory explosion. The subset was tuned on the training set and exploits domain knowledge. For all events, all mentioned TIME segmentations and TIME relations are used, i.e. we used the same TIME representation for all events from the same domain. For both domains, we use a fixed value of 0.5 seconds for the margin T_1. We will now first discuss the soccer representation, we then proceed with the news representation.

3.4.1 Soccer Representation

The teletext (European closed caption) provides a textual description of what is said by the commentator during a match. This information source was analyzed for presence of informative keywords, like *yellow, red, card, goal, 1-0, 1-2*, and so on. In total 30 informative stemmed keywords were defined for the various events.

On the visual modality we applied several detectors. The type of camera work [12] was computed for each camera shot, together with the shot length. A face detector [120] was applied for detection of persons. The same detector formed the basis for a close-up detector. Close-ups are detected by relating the size of detected faces to the total frame size. Often, an author shows a close-up of a player after an event of importance. One of the most informative pieces of information in a soccer broadcast

are the visual overlay blocks that give information about the game. We subdivided each detected overlay block as either info, person, referee, coach, goal, card, or substitution block [140], and added some additional statistics. For example the duration of visibility of the overlay block, as we observed that substitution and info blocks are displayed longer on average. Note that all detector results are transformed into binary output before they are included in the analysis.

From the auditory modality the excitement of the commentator is a valuable resource. For the proper functioning of an excitement detector, we require that it is insensitive to crowd cheer. This can be achieved by using a high threshold on the average energy of a fixed window, and by requiring that an excited segment has a minimum duration of 4 seconds.

We take the result of automatic shot segmentation as a reference interval. An overview of the TIME representation for the soccer domain is summarized in Table 3.1.

3.4.2 News Representation

The news events we want to classify are dominated by talking people. Most detectors that we propose are based on this observation. In the auditory modality we look for speech segments. This is simply achieved by using the previously discussed excitement detector with a lower threshold.

In the visual modality we detected faces [120] and several derived statistics, like position, number, and camera distance used. We also detected the dominant camera work used during the shot, since the events we try to classify are typically shot using a static camera. For each shot we furthermore computed the average motion, number of flashes, length, and whether it was preceded or succeeded by an effect. Text localization [12] was applied to detect regions of overlayed text. We differentiated between presence of a single region and parallel regions, e.g. one in the top of the image frame and on the bottom.

For each detected text region we recognized the text and tried to match it, using fuzzy string matching, with the city name where the news studio is located. The presence of closed caption segments was used as an additional indicator for speech. Moreover, they were scanned for presence of weather related keywords like *sunny, snow, degree, west* and so on.

Again we take the result of automatic shot segmentation as a reference interval. The TIME representation for the news domain is summarized in Table 3.2. When comparing both Table 3.1 and 3.2, one can clearly see that Table 3.1 includes more context, whereas Table 3.2 is more concerned with synchronization. In the next section we will evaluate the automatic indexing of events in soccer and news video, based on the presented pattern representation.

3.5 Results

For the evaluation of the TIME framework we used soccer and news broadcasts from Dutch national TV. We recorded 8 live soccer broadcasts, about 12 hours in total. The

Table 3.2: TIME representation for news analysis. T_2 indicates the contextual range used by the precedes and precedes_i relations.

TIME segmentation	TIME relations	T_2 (s)
Camera work	*during*	
Effect	*precedes*, *precedes_i*	0 - 4
Block length	*during*	
Camera distance	*during*	
Face left	*during*	
Face right	*during*	
Face center	*during*	
Number of faces	*during*	
Number of flashes	*during*	
Kinetic Energy	*during*	
Speech	*All relations*	0 - 1
Closed caption	*All relations*	0 - 1
Overlayed text	*All relations*	0 - 1
Parallel overlayed text	*All relations*	0 - 1
Studio keyword	*during*	
Weather keyword	*during*	

videos were digitized in 704 × 576 resolution MPEG-2 format. For the news domain we recorded 24 broadcasts, again about 12 hours in total, in 352 × 288 resolution MPEG-1 format. The audio was sampled at 16 KHz with 16 bits per sample for both domains. The time stamped teletext was recorded with a teletext receiver. For soccer analysis we used a representative training set of 3 hours and a test set of 9 hours. For news, a training and test set of 6 hours each was used. In this section we will first present the evaluation criteria used for evaluating the TIME framework, then we present the classification results obtained. After presenting two prototype systems, we end with a discussion on the results.

3.5.1 Evaluation Criteria

The standard measure for performance of a statistical classifier is the error rate. However, this is unsuitable in our case, since the amount of relevant events are outnumbered by irrelevant pieces of footage. We therefore use the precision and recall measure adapted from information retrieval. Let $|R|$ be the number of relevant camera shots, i.e. camera shots containing the specific event one is looking for. Let $|A|$ denote the answer set, i.e. the number of camera shots that are retrieved by the system. Let $|R \cap A|$ be the number of camera shots in the intersection of the sets R and

Table 3.3: Evaluation results of the different classifiers for soccer events, where duration is the total duration of all segments that are retrieved.

	Ground truth		**C4.5**		**MaxEnt**		**SVM**	
	Total	Duration	Relevant	Duration	Relevant	Duration	Relevant	Duration
Goal	12	3^m07^s	2	2^m40^s	10	10^m14^s	11	11^m52^s
Yellow Card	24	10^m35^s	13	14^m28^s	22	26^m12^s	22	12^m31^s
Substitution	29	8^m09^s	25	15^m27^s	25	7^m36^s	25	7^m23^s
$\sum$	65	21^m51^s	40	32^m35^s	57	44^m02^s	58	31^m46^s

A. Then, precision is the fraction of retrieved camera shots (A) which are relevant:

$$Precision = \frac{|R \cap A|}{|A|} \tag{3.4}$$

and recall is the fraction of the relevant camera shots (R) which have been retrieved:

$$Recall = \frac{|R \cap A|}{|R|} \tag{3.5}$$

This measure gives an indication of correctly classified events, falsely classified events, and missed events. For the evaluation of news classification, results will be plotted in a recall-precision curve.

For the evaluation of soccer we used a different approach. Since events in a soccer match can cross camera shot boundaries, we merge adjacent camera shots with similar labels. As a consequence, we loose our arithmetic unit. Therefore, precision and recall can no longer be computed. As an alternative for precision we relate the total duration of the segments that are retrieved to the total duration of the relevant segments. Moreover, since it is unacceptable from a users perspective that scarce soccer events are missed, we strive to find as many events as possible in favor of an increase in false positives. Finally, because it is difficult to exactly define the start and end of an event in soccer video, we introduce a tolerance value T_3 (in seconds) with respect to the boundaries of detection results. We used a T_3 of 7 s. for all soccer events. A merged segment is considered relevant if one of its boundaries plus or minus T_3 crosses that of a labeled segment in the ground truth.

Besides a comparison of individual classifiers, we also compare the influence of TIME on the final result. Since the benefit of using TIME for domains relying on context is obvious, we only show this result for the news domain.

3.5.2 Event Classification

For evaluation of TIME on the soccer domain, we manually labeled all the camera shots as either belonging to one of four categories: yellow card, goal, substitution, or

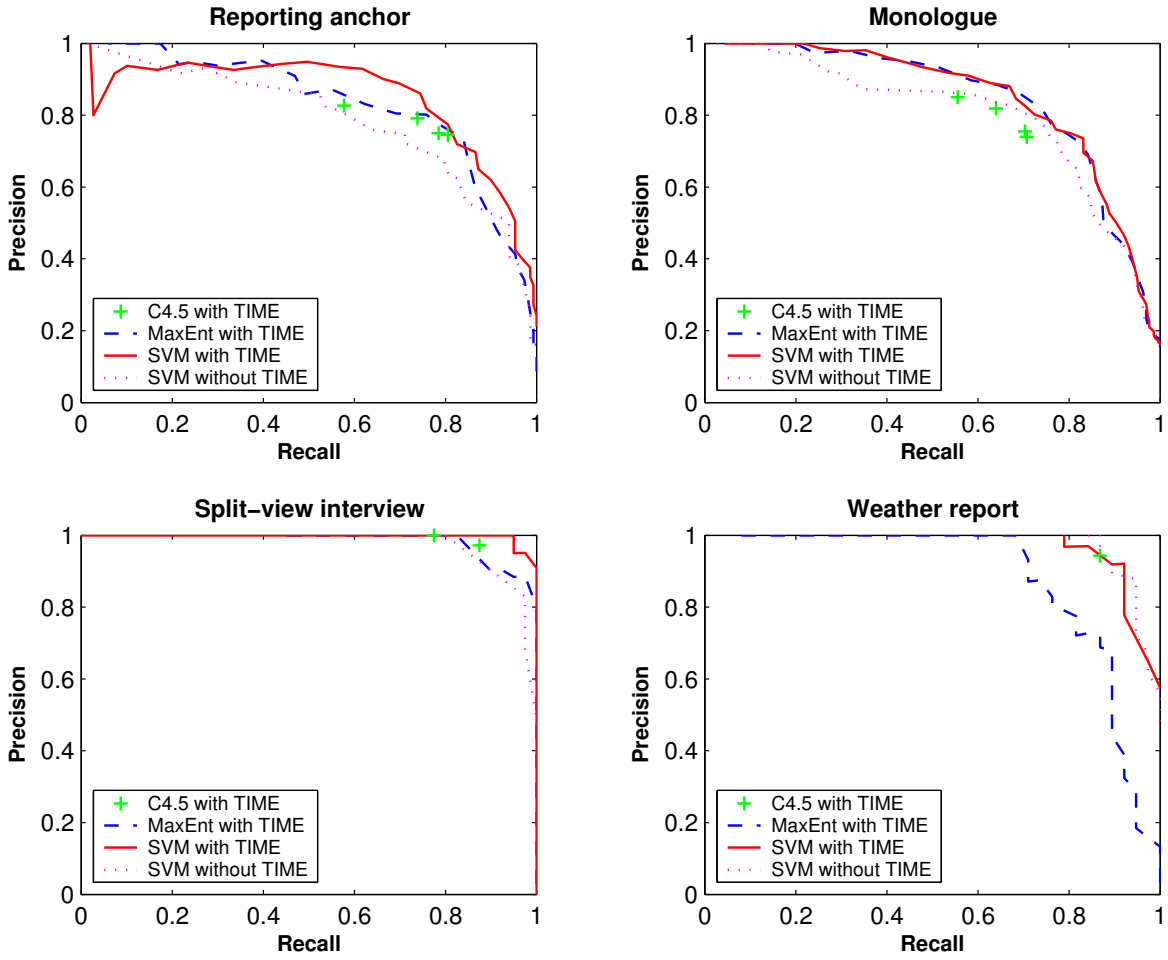


Figure 3.4: Recall-precision curves for four semantic events in news broadcasts.

unknown. Red card and penalty were excluded from analysis since there was only one instance of each in the data set. For all three remaining events a C4.5, MaxEnt, and SVM classifier† was trained. Results on the test set are visualized in Table 3.3.

When analyzing the results, we clearly see that the C4.5 classifier performs worst. Although it does a good job on detection of substitutions, it is significantly worse for both yellow cards and goals when compared to the more complex MaxEnt and SVM classifiers. When we compare results of MaxEnt and SVM, we observe that almost all events are found independent of the classifier used. The amount of video data that a user has to watch before finding these events is about two times longer when a MaxEnt classifier is used, and about one and a half times longer when an SVM is used, compared to the best case scenario. This is a considerable reduction of watching time when compared to the total duration, 9 hours, of all video documents

†For classification the following open source toolboxes were used:
S. Ruggieri. *YaDT - Yet another Decision Tree builder.*
J. Baldridge, T. Morton and G. Bierner. *OpenNLP Maxent.*
C.-C. Chang and C.-J. Lin. *LIBSVM: a library for support vector machines.*

in the test set. With the SVM we were able to detect one extra goal, compared to MaxEnt. Analysis of retrieved segments learned that results of Maximum Entropy and SVM are almost similar. Except for goal events, where nine events were retrieved by both, the remaining classified goals were different for each classifier.

For the news domain we used the same classification approach as for soccer. But now focussing on 4 events, namely: reporting anchor, monologue, split-view interview, and weather report. Again for each event a C4.5, MaxEnt, and SVM classifier was trained. Moreover, we also compared the added value of TIME by inclusion of one run with the SVM classifier where all TIME relations were replaced by *during* relations.

Results of news classification are visualized by means of recall-precision curves in Fig. 3.4. For the MaxEnt classifier we varied the threshold on the likelihood for each camera shot computed by (3.1). For SVM we varied the threshold on the margin computed by (3.2) for each camera shot. For C4.5 this is impossible because of its binary nature, we therefore plotted results of 5 pruning values. When comparing classification results of the different classifiers we observe that SVM outperforms all other classifiers, and that C4.5 achieves comparable classification results when compared with a MaxEnt classifier. MaxEnt performs better on monologues, C4.5 performs better on weather reports, and is even comparable to SVM for this event. The experimental results of SVM with and without TIME clearly show that there is a significant gain in classification results when using the TIME framework. Only for classification of weather report events, an SVM classifier without TIME can achieve comparable results as an SVM with TIME. For all other classes, it is outperformed by the SVM with TIME.

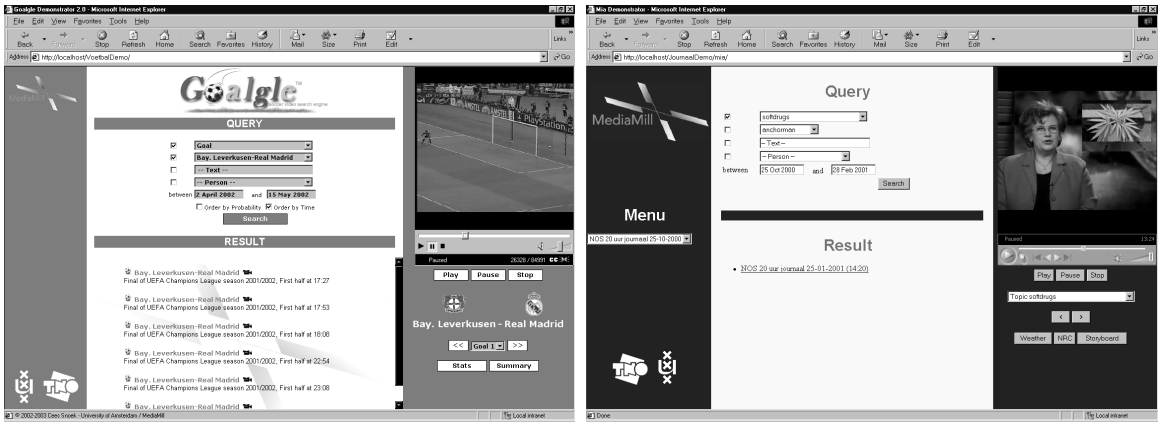
3.5.3 Implementation

Based on the current classification result we have developed the *Goalgle* soccer video search engine, and added functionality to the *News RePortal* system, see Fig. 3.5. In its current form, the web based prototypes allow to query a selection of broadcasts on keywords, persons and events. Ultimately this should result in a personalized automatic summary, that can be presented on a wide range of pervasive devices.

3.5.4 Discussion

When we take a closer look to the individual results of the different classifiers, it is striking that C4.5 can achieve a good result on some events, e.g. substitution and weather report, while performing bad on others, e.g. goal and monologue. This can, however, be explained by the fact that the events where C4.5 scores well, can be detected based on a limited set of TIME relations. For substitution events in soccer an overlay during the event is a very strong indicator, whereas a weather related keyword in the teletext is very indicative for weather reports. When an event is composed of several complex TIME relations, like goal and monologues, the relatively simple C4.5 classifier performs worse than both complex MaxEnt and SVM classifiers.

To gain insight in the meaning of complex relations in the two domains, we consider the GIS algorithm from Section 3.3.2, which allows to compute the importance or



(a) (b)

Figure 3.5: Screen dumps of (a) the Goalgle soccer video search engine, and (b) the News RePortal system.

relative weight of the different relations used. The weights computed by GIS indicate that for the soccer events goal and yellow card specific keywords in the closed captions, excitement with during and overlaps relations, a close-up afterwards, and the presence of an overlay nearby are important relations. For the news events reporting anchor and monologue, a close-up face on the left side during the shot, a low average motion during the shot, and overlayed text during the shot were of equal importance. For reporting anchors speech that starts the camera shot was important, whereas various relations with overlayed text were important for monologues. The weights for the speech relation for monologues weren't high enough to consider it very important, which is quite surprising. This can be explained by the fact that non-Dutch speakers are transcribed by means of overlayed text in the Dutch news, hence the detection of such overlayed text is more distinguishing than speech for monologues. For split-view interview events, two faces during the camera shot, meets and equals relations with overlayed text showing the location of the two speakers, overlapping and during speech relations, and the identification of a city keyword in the overlay text were important. For weather reports, besides keywords in the teletext, a long shot camera distance during the camera shot, and overlayed text with start and finish relations are of importance.

When combining the weights, MaxEnt sometimes fails to profit from multiple information sources. This is best observed in the recall-precision curve for weather reports. Overall, the SVM classifier achieves comparable or better results than MaxEnt. When we analyze false positives for both classifiers, we observe that these are caused because some of the important relations are shared between different events. For soccer this mostly occurs when another event is indeed happening in the video, e.g. a hard foul or a scoring chance. For news this especially occurs for classification of reporting anchors and monologues. Often a monologue is classified as anchor and

vice versa. We also found that close-ups of people in report footage with voice-overs, and reporting anchor's that were filmed from less usual camera positions were often falsely classified. False negatives are mainly caused by the fact that a detector failed. By increasing the number of detectors and relations in our model we might be able to reduce these false positives and false negatives. Another option is to use a cascade of classifiers, so instead of classifying each event individually, first classify events on which you can do a good job, e.g. split-view interviews, and apply another classifier on the negative results of the first classifier, and so on. This should yield better indexing results.

3.6 Conclusion

To bridge the semantic gap for multimedia event classification, a new framework is required that allows for proper modeling of context and synchronization of the heterogeneous information sources involved. We have presented the Time Interval Multimedia Event (TIME) framework that accommodates these issues, by means of a time interval based pattern representation. Moreover, the framework facilitates robust classification using various statistical classifiers.

To demonstrate the effectiveness of TIME it was evaluated on two domains, namely soccer and news. The former was chosen because of its dependency on context. The latter because of its dependence on synchronization. We have compared three different statistical classifiers, with varying complexity, and show that there exists a clear relation between narrowness of the semantic gap and the needed complexity of a classifier. When there exists a simple mapping between a limited set of relations and the semantic concept we are looking for, a simple decision tree will give comparable results as a more complex SVM. When the semantic gap is wider, detection will profit from combined use of multimodal detector relations and a more complex classifier, like the SVM. Moreover, we show that the TIME framework, including synchronization and context, outperforms the 'standard' multimodal analysis approaches common in video indexing literature.

In the future we aim to explore the usage of complex classifier combinations and architectures. Moreover, by inclusion of more textual resources we expect to be able to give a richer description of events in video, ultimately bridging the semantic gap for a large set of events.

Chapter 4

Learning Rich Semantics from Produced Video by Style Analysis

In this Chapter, we propose a generic and flexible framework for produced video indexing that is capable to learn rich semantic concepts from multimodal sources based on style analysis. Four properties that are indicative for style are identified, namely layout, content, capture, and context. By combining a fixed core of layout, content, and capture detectors together with varying context detectors into a classifier ensemble, the framework facilitates robust classification of several rich semantic concepts in produced video. Results on 120 hours of video data from the 2003 TRECVID benchmark show that it is the combination of style elements that yields the best results for produced video indexing. In addition, we demonstrate that the accuracy of the proposed framework for classification of several rich semantic concepts in broadcast news is state-of-the-art.

4.1 Introduction

Advancement in optical fiber technology and growing availability of low-cost digital multimedia recording devices enables worldwide capture, delivery, and exchange of large amounts of video documents. This overwhelming amount of digital video data will trigger the need for automatic indexing tools that can provide on-the-fly content-based annotation, allowing for effective and efficient browsing, filtering, and retrieval of specific video segments. The progress in content-based multimedia analysis, however, has not kept pace with this technology push.

Automatic techniques for video indexing suffer from the fact that it is hard to infer semantics based on features extracted from the data. This *semantic gap* [136] has hampered the development of a generic index solution. To study the problem, two major classes of video documents should be distinguished, namely non-produced and produced video. For non-produced video, e.g. security or home video, the content is more or less accidental. In contrast, video created in a production environment, e.g. a feature film or broadcast news, has an author or director to guide all facets of the production process. He or she imposes a certain style to express a semantic intention in the form of concepts. Where style is merely expressed by the choice of techniques and their interrelationship. As considers semantic indexing, the content-based multimedia research community has ignored the stylized nature of produced video.

Initial work on semantic indexing of multimedia focused on content-based analysis only. This approach is fruitful for concepts that are easy to distinguish because of their large similarity in (visual) content, e.g. tigers and soccer games. For concepts that have more variability in their content, e.g. buildings, sporting events, and dialogues, analysis methods based on content only are too fragile. Some concepts, however, share many similarities in their production style. Both sporting events and dialogues for example, are often recorded from a fixed camera distance. To distinguish this class of *rich semantic* concepts we therefore need the notion of style, see Fig. 4.1.

We perceive of the author's style as a combination of four elements. In addition to content, we identified in [142] layout as an important aspect for analysis of the author's style. As noted by [21, 24], capture of the data into a multimedia medium is also an important stylistic element. Furthermore, in a produced video a concept does not occur in isolation, but is further defined by its local context [98]. Thus where bridging the semantic gap for non-produced video is not within reach, one can potentially bridge it for produced video. The key observation to help overcome the semantic gap in produced video is that an author in many ways stylizes rich semantic concepts that appear in a video document. Thus, produced video analysis methods should exploit style to infer rich semantics.

An author thinks in concepts, and aims to stimulate all senses of the audience when expressing a semantic intention. Thus, the author combines the visual, auditory, and textual modalities in the video document. Hence, analysis methods should exploit the multimodal properties of video documents to its full potential when aiming for detection of semantics based on style analysis.

Various multimodal approaches for produced video indexing exist, see [142] for

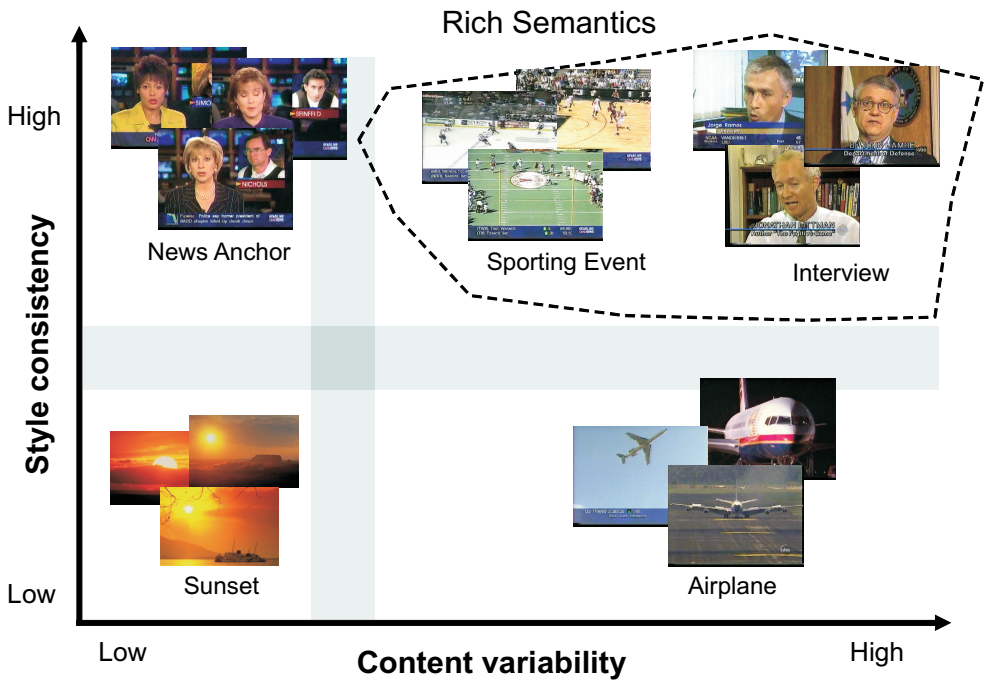


Figure 4.1: Relation between content variability and style consistency. Rich semantics have both a high variability in content and a high consistency in style.

an overview. Methods for extraction of semantics are often based on highly specific features and knowledge based classification rules [142]. Generic applicability and robustness of these methods, therefore, is limited because of their model dependency. Recently, the focus in multimodal video indexing has shifted to machine learning based approaches. This has resulted in increased robustness for classification of concepts related to setting, objects, people, and events, e.g. [4, 10, 60, 98, 141]. However, none of these methods uses style.

In this Chapter we propose a generic, flexible, and robust framework for produced video indexing based on style analysis. Our framework is generic because we learn various rich semantic concepts from a fixed set of layout, content, and capture detectors. We achieve flexibility within the framework by varying the set of context detectors, depending on the rich semantic concept one is looking for. Finally, the framework guarantees robustness by integrating all style elements into a statistical classifier ensemble.

We participated in the 2003 NIST TRECVID video retrieval benchmark [134], to demonstrate the applicability of our framework. The data set totaled 120 hours of produced news episodes from ABC and CNN. In total 17 semantic concepts were defined by TRECVID to be detected in this data set. Most concepts can be classified as content-based concepts related to setting, e.g. vegetation, objects, e.g. airplane, and people, e.g. news subject face. We focus our evaluation on classification of rich semantic news concepts that use various style elements, namely: *news subject monologue*, *non-studio setting*, *sporting event*, and *weather news*.

The organization of the remainder of this Chapter is as follows. First, we discuss in more detail related work. Then, we proceed with the introduction of our framework for produced video indexing in Section 4.3. We discuss an implementation of the framework for the news genre in Section 4.4. An extensive evaluation, demonstrating the applicability of our framework, is presented in Section 4.5.

4.2 Related Work

Initial work on produced video indexing started with the analysis of visual layout and content. A good example of this exploratory work is [176]. The authors focus on parsing and indexing of video. Based on extracted motion features they are able to classify concepts such as crowds and talking heads. Because of the exploring nature of this work, experiments are carried out on a small-scale video data set only. In addition, the multimodal nature of produced video is ignored.

Large scale multimodal produced video indexing has been pioneered by the Informedia project at Carnegie Mellon University [161]. Their approach focused on adapting techniques developed for other domains, like speech recognition, face detection, and natural language processing, into a video indexing and retrieval environment. This has resulted in a news video analysis toolbox that exploits content in a knowledge-based fashion. A drawback of their approach is the news model dependency, and therefore lack of robustness. Although the current system is shifting towards the usage of more advanced learning schemes [60], layout, capture, and context are largely ignored.

Naphade *et al.* [98] were among the early adopters of advanced pattern recognition techniques for semantic classification of produced video. In [98] they propose to model semantic concepts through probabilistic detectors, for example airplane, skydiving, and bird detectors. The authors refer to these concept detectors as multijects. Integration of multijects into a network representation, referred to as Multinet, allows inferring contextual semantics, e.g. outdoor based on detection of vegetation and sky. By combining the individual probabilities of all multijects into a Multinet, using factor graphs [98], the framework is applicable to all sorts of multimedia data and a variety of semantic indexes. However, the experiments only consider visual concepts related to the setting of the multimedia data, e.g. rocky terrain, water-body, and forestry. This indicates that this method is mostly suited for non-rich semantics.

An extended and truly multimodal version of [98] was presented in [4, 10], now as part of the IBM Research TRECVID contribution. Here the Multinet is one of the final classifiers in a pipeline of analysis steps that exploits various machine learning and multimodal integration schemes. The pipeline starts with a set of standard and semantically poor image, audio, and textual features. Based on these features the pipeline then generates several unimodal statistical models for a lexicon of 64 semantic concepts. For integration of modalities and models at the concept level, Ensemble Fusion, amongst others, is applied. This fusion scheme includes normalization of confidence scores, several combiner functions, and parameter optimization, see also [155]. All multimodal concepts then serve as the input for the Multinet that builds a context

of concepts for final semantic classification. This approach has demonstrated good results on the concept detection task of the NIST TRECVID benchmark, resulting in the highest mean average precision for this task in 2002 and 2003. Despite this success, we identify some limitations in the current pipeline approach. First, at its core the system exploits a small set of semantically poor content features. This set still focuses on visual concepts, and context. It is therefore not surprising that one of the concepts for which the authors had poor performance was female speech. Modality integration by combining classifier models at the concept level is another limitation of the pipeline approach. As it neglects the important issue of synchronizing the layouts of different modalities. In addition, the approach only considers context with concepts which occur together, disregarding their sequential order. Finally, the approach ignores capture. Hence, the current pipeline approach is not optimal for detection of rich semantic concepts in produced video that have a large variability in (visual) content, depend on synchronization between modalities, exploit temporal context, and involve specific capture properties.

A framework for synchronization of multiple modalities and inclusion of temporal context was proposed in [141]. Viewing the result of individual detectors as time intervals, allows for combination of layout and content into a common representation. The proposed representation exploits interval relationships and facilitates classification of semantic events in soccer and news using several pattern recognition methods. A drawback of the presented framework is that it ignores the capture and the context.

Combining the above, by explicitly modeling multimodal layout, content, capture, and context into a common style-based analysis framework, we are able to detect the rich semantics, as intended by the author of produced video, more accurately.

4.3 Produced Video Indexing Framework

Produced video indexing can be regarded as a reversed authoring process [142]. To arrive at a generic framework for produced video indexing, we therefore first consider video document production, in particular the role of style elements. Then we proceed with style-based produced video analysis. We end this section with a classifier combination scheme that facilitates detection of rich semantics in produced video.

4.3.1 Video Document Production Model

A produced video document is the work of an author or director who conceives an initial idea and finally produces a result, semantically reflecting this idea as good as possible. To communicate a semantic intention by means of a video, an author has an arsenal of techniques to choose from [21,24]. The choice for a specific set of techniques is restricted only by the imagination of the author and the *genre* of the video to be produced. The genre feature film presents a specific set of techniques to choose from, while another set is available for news broadcast. In practice, the author will not make all possible technical decisions in isolation. For specific tasks, the author relies on a production team of specialists. The creative choices made during the creation

process are commonly referred to as the author's style [21].

In the production team, we distinguish different creative roles. The blueprint of any produced video is the scenario provided by the *scenario writer*, e.g. a script of a feature film, or a story board of a news broadcast. The scenario contains choices for the assembly of concepts into a plot and story line of the produced video. Concepts do not occur in isolation, thus context is an important instrument for a scenario writer to define the semantics of the concept. We distinguish between spatial and temporal context, where spatial context refers to simultaneous co-occurrence of concepts and temporal context refers to sequential co-occurrence of concepts. Context poses a semantic structure on the video. Thus, for the role of the scenario writer, we define:

Definition 4.3.1 (Context) *Set of style elements, $\mathcal{S}$, that define the spatial and temporal semantic structure of a produced video document.*

Guided by the scenario, the *production design* defines the content of a video document by arranging people, objects, and setting [142]. For feature films, this includes choices for cast, costumes, setting, and so on. For news broadcasts, choices include the number of anchors, decoration of the studio, and type of weather map. We define:

Definition 4.3.2 (Content) *Set of style elements, $\mathcal{C}$, that define the people, objects, and setting appearing in a produced video document.*

The *recording unit* guides the recording of the video content. In feature films, the cinematographer and sound unit take care of camera framing, lightning, and balance and combination of microphones. In news broadcasts, one recording team works in the studio and another one on location. Both are taking care of specific recording circumstances. The recording unit uses sensors, like cameras and microphones, to capture the content into a multimedia format. Furthermore, they use devices that influence the capture, like lightning and color filters. We define:

Definition 4.3.3 (Capture) *Set of style elements, $\mathcal{T}$, that define the transfer of an observed scene into the sensory format of a produced video document.*

The *editor* is responsible for assembly and synchronization of individual pieces, after the recording is finished. For feature films, the editor takes care in synchronizing dialogues and adding music. For news, the editor assures that the voice over of the reporter is in line with the visible content. The layout results after the editor is finished with the video document, and is the combination of sensor shots, transition edits, and special effects [142]. We define:

Definition 4.3.4 (Layout) *Set of style elements, $\mathcal{L}$, that define the sensor shots, transition edits, and special effects of a produced video document.*

The author is responsible for the complex interplay of all style elements to communicate a semantic intention, within the predefined genre. Instances of individual pieces are less important, since they are interchangeable. For feature films, it does not really matter whether a desert scene is shot in the Kalahari Desert or Death Valley, and for news broadcasts, it is not important whether the anchor is sitting

behind a desk or stands in front of an infographics screen. In addition, the choice for specific recording circumstances, e.g. using a wide-angle lens in combination with a yellow color filter, does not influence the choice for a specific editing technique. This makes the individual style elements in the video production process more or less independent.

By combining the style elements for the four independent components in a specific way, an author is guiding the spectator to interpret the produced video in correspondence with the author's semantic intention. We define:

Definition 4.3.5 (Produced Video Authoring) *Process that defines how a semantic intention,* Ω*, is authored into a produced video document,* Π*, according to a genre dependent set of four independent style elements,* $\{\mathcal{L}, \mathcal{C}, \mathcal{T}, \mathcal{S}\}$*.*

Thus, we consider the author's style to consist of four independent style elements and their genre dependent combination.

4.3.2 Style Analysis

Analysis of produced video should focus on the authoring-driven production process that is responsible for the creation of the video document. Since a produced video is often available only in a raw data format, we need to identify as many of the style choices made during production as possible using detectors. We group these style detectors into four independent sets, based on the independent roles identified for video production: i.e. scenario writing, production design, recording, and editing. We cannot analyze roles directly from the data. Therefore, a style detector can at best approximate the result of each creative role involved in video production. Summarizing the above, a produced video is analyzed using four groups of style detectors: layout detectors, content detectors, capture detectors, and context detectors.

The set of layout detectors is limited by the number of modalities involved. When an editor chooses to use a special effect, this has no consequences for the sensor shot used. Thus for layout, detectors for various elements act independently of each other. We define:

Definition 4.3.6 (Layout Detectors) *Set of independent style detectors* $\hat{\mathcal{L}}$ *that yield an approximation of* $\mathcal{L}$*.*

In contrast to layout, the set of possible content detectors is unlimited in theory, see [142] for an overview. Although the combination of content elements is important for the semantics, choices made by the production design for specific instances are independent of each other, e.g. the choice for a certain actor does not influence the choice for a specific location. Hence, we consider detectors for content elements independent of each other also. We define:

Definition 4.3.7 (Content Detectors) *Set of independent style detectors* $\hat{\mathcal{C}}$ *that yield an approximation of* $\mathcal{C}$*.*

Like layout detectors, the number of possibilities for capture detectors are bounded, but now by the degrees of freedom of the recording sensors and devices. The fact that

the recording unit applies a specific camera movement, does not limit the choice for a certain color filter. Hence, the individual capture detectors are again independent of each other. We define:

Definition 4.3.8 (Capture Detectors) *Set of independent style detectors $\hat{\mathcal{T}}$ that yield an approximation of $\mathcal{T}$.*

Context can enhance or limit the number of possible semantic interpretations of a video segment. Moreover, as a scenario writer applies spatial and temporal context separately, we consider detectors independent of each other in the analysis.

Definition 4.3.9 (Context Detectors) *Set of independent style detectors $\hat{\mathcal{S}}$ that yield an approximation of $\mathcal{S}$.*

A detector-based analysis of style elements in produced video results in the mapping $\Pi \rightarrow \left\{\hat{\mathcal{L}}, \hat{\mathcal{C}}, \hat{\mathcal{T}}, \hat{\mathcal{S}}\right\}$. We need a common representation to combine the detector results of all style elements. This involves synchronization, since elements from the various modalities are not necessarily aligned. Synchronization has largely been ignored in literature, and is typically solved by aligning all detection results to a camera shot layout, although better schemes exist [141]. For style-based analysis, we define:

Definition 4.3.10 (Style Vector) *A vector $\vec{s}_i$ that contains the synchronized result of the detector ensemble $\left\{\hat{\mathcal{L}}, \hat{\mathcal{C}}, \hat{\mathcal{T}}, \hat{\mathcal{S}}\right\}$, where individual components are independent.*

where i indicates the segmentation used. The style vector, resulting from the synchronization and concatenation of individual components, forms the basis for learning rich semantics from produced video. We define:

Definition 4.3.11 (Produced Video Analysis) *Process that defines how a rich semantic class, ω, is learned from a produced video document, Π, according to a set of style vectors.*

4.3.3 Semantic Classifier

We perceive detection of rich semantics as a pattern recognition problem. We aim to detect a rich semantic class ω based on an ensemble of independent detectors represented in a style vector $\vec{s}_i$ using the probability $p(\omega|\vec{s}_i)$. This requires a classifier combination scheme. A classifier combination scheme combines results of several independent classifiers or detectors that solve the same task. However, there is no reason to assume that the same technique can not be used to combine classifiers that do not solve the same task per se, but are related semantically, i.e. share the same author intention. Except for trivial cases, detectors are imperfect and generate both false positive and false negative results. Hence, in terms of statistical pattern recognition, we consider each individual detector to act as a weak classifier. A classifier ensemble benefits from the synergy of a combined use of weak learners, resulting in improved performance. This is especially the case when the various classifiers are largely independent [68]. As we have designed $\vec{s}_i$ as an ensemble of independent detectors,

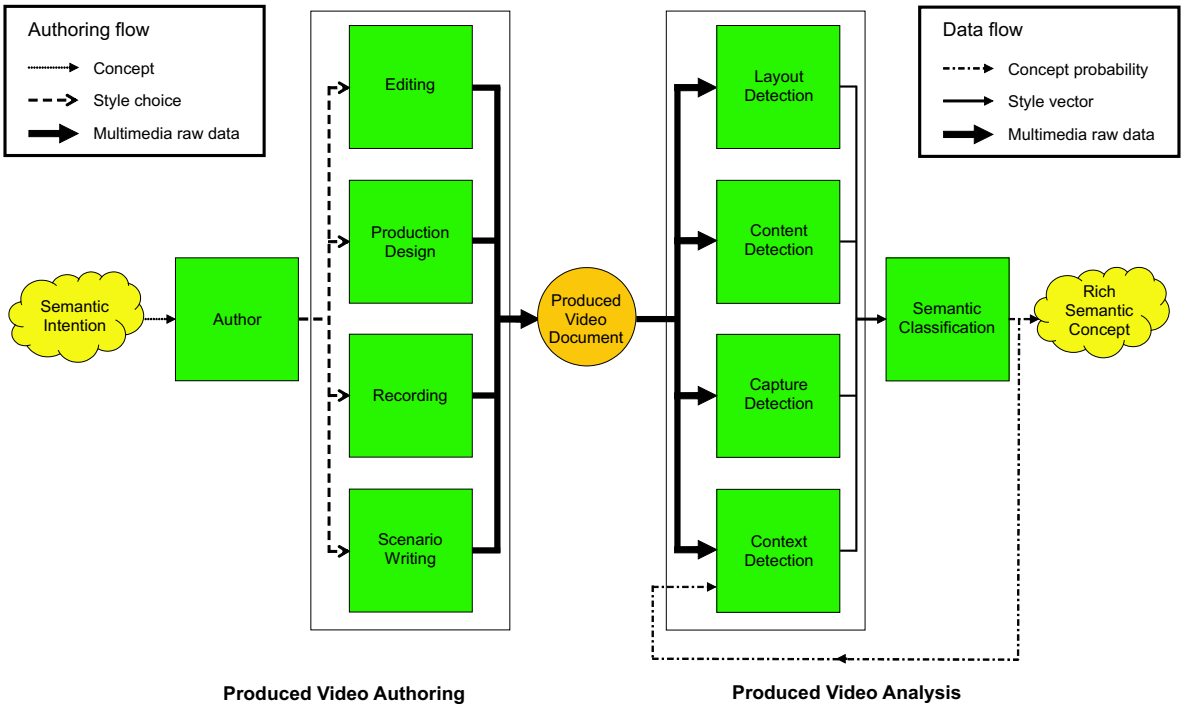


Figure 4.2: General framework for style-driven produced video indexing.

a classifier combination scheme is a natural choice for learning rich semantics from produced video.

The classifier combination scheme yields a style model that is applicable to any data set. However, discriminatory power of the style model increases by restricting the data set for which a model is developed. One can achieve restriction by limiting the data set to a specific genre, author, or both. We define:

Definition 4.3.12 (Style Model) *Model resulting after applying a classifier combination scheme to a set of style vectors.*

In literature various classifier combination schemes exist, e.g. bagging [25], boosting [128], and stacking [166]. Bagging and boosting resample the training set to obtain an ensemble, or series, of independent classifiers. They differ in the way they combine the individual results. Both schemes focus on the data and exploit independence by combining classifiers that are trained on different samples of the training set. In contrast, stacking focuses on the classifiers. This classifier combination scheme, uses the output labels of individual classifiers as input features for a *stacked* classifier, which learns how to combine the reliable classifiers in the ensemble and makes the final decision. Because a style vector is composed of independent style elements and detectors, an assurance for independence exists and there is no need for resampling. Hence, for our purpose, i.e. detection of rich semantics, stacking is a good choice.

The probabilistic output $p(\omega|\vec{s}_i)$ obtained from a stacked classifier allows to define new concept detectors, which we then add to the context. Suppose we constructed a style model for detection of fouls in soccer matches. When we apply this model to

a set of shot-segmented soccer broadcasts, it results in a probability of occurrence of fouls for each shot. We can then add this concept, together with its probability, to the context and use it for a style model that detects segments where a referee presents a red card to a player. Besides positive correlation, negative correlation is also helpful for the context. It aids in preventing false positive classification of semantically different concepts that share many style elements, e.g. goals and penalties in soccer matches. Moreover, it can be exploited for detection of rich semantics that are defined by what it is not, e.g. non-goal events. Note that by iteratively adding concepts to the context some independence is lost. As this only involves a small fraction of all detectors in the ensemble, we do not consider it a problem. The order in which context is updated can be defined by domain knowledge, experimentation, or feature selection techniques [68].

The complete framework for indexing of produced video documents in terms of rich semantics is visualized in Fig. 4.2.

4.4 An Experiment on the News Genre

We carried out a set of experiments, as part of the semantic concept detection task of the 2003 NIST TRECVID benchmark, to evaluate the viability of our produced video indexing framework.

The data for the benchmark contained about 120 hours of ABC World News Tonight and CNN Headline News from the first half of 1998. Together with 13 hours of C-SPAN programming, mainly containing public discussions, from the period 1998-2001. NIST provided the videos in MPEG-1 format.

NIST defined 17 semantic concepts. We focused on the semantically rich concepts that exploit style in many ways. Namely:

- *News subject monologue*: segment contains an event in which a single person, a news subject not a news person, speaks for a long time without interruption by another speaker. Pauses are ok if short;
- *Non-studio setting*: segment is not set in a TV broadcast studio;
- *Sporting event*: segment contains video of one or more organized sporting events;
- *Weather news*: segment reports on the weather;

Since most of these concepts exist within the news genre only, we ignore the 13 hours of C-SPAN data in our experiments.

NIST splits the corpus into an equally sized training and test set, i.e. each containing about 60 hours of produced news video. For training, we manually labeled a subset of the training set of about 24 hours, i.e. 23 ABC and 24 CNN broadcasts. We labeled examples for all four rich semantic concepts under consideration.

4.4.1 Semantic Classifier Implementation

As a stacked semantic classifier we choose the Support Vector Machine (SVM) [28, 158], which is known to be a stable classifier for various classification problems. In

addition, it has also proven to be a good choice in a multimodal video indexing setting [4, 141]. The SVM tries to find an optimal hyperplane between two classes by maximizing the margin between these two classes. It has the following form: $\omega|(\vec{w} \cdot \vec{s}_i + b)| \geq 1$, where $\vec{w}$ is a weight vector, and b is a threshold. The problem of finding the optimal hyperplane is a quadratic programming problem, that can be casted into the following form [28]:

$$\min_{\vec{w},\xi} \Big\{ \frac{1}{2}\vec{w} \cdot \vec{w} + C^{+}\Big(\sum_{i=1}^{t} \xi_i\Big) + C^{-}\Big(\sum_{i=1}^{t} \xi_i\Big) \Big\} \tag{4.1}$$

Under the following constraints:

$$\omega|(\vec{w} \cdot \vec{s}_i + b)| \geq 1 - \xi_i, \quad i = 1, 2, \ldots, t \tag{4.2}$$

Where C^{+} and C^{-} are parameters that allow to adjust the influence of the number of positive and negative examples in the training data, t is the total number of style vectors in the training set, and ξ_i are slack variables that are introduced when the data is not perfectly separable. Both data balancing and the use of slack variables are required for detection of rich semantics, since rich semantics are never evenly balanced in the data and never perfectly separable. For each concept, we perform a parameter search to optimize the settings for (4.1) in our classification scheme.

To allow for combination of style models, we convert the classification result of the SVM, i.e. the margin, to a calibrated result. Ideally one would have a posterior probability, $p(\omega|\vec{s}_i)$, that given an input style vector $\vec{s}_i$ returns a confidence value for a particular class ω. But the model dependent output of an SVM, $\gamma(\vec{s}_i)$, is not a probability. A popular and stable method for SVM output conversion was proposed in [113]. This solution exploits the empirical observation that class-conditional densities between the margins are exponential; therefore, the author suggests a sigmoid model. We apply the output of this model in our classifier architecture. This results in the following posterior probability:

$$p(\omega|\vec{s}_i) = \frac{1}{1 + \exp(\alpha\gamma(\vec{s}_i) + \beta)} \tag{4.3}$$

where the parameters α and β are maximum likelihood estimates based on the training set [113]. We rank produced video indexing results based on the probabilistic output $p(\omega|\vec{s}_i)$.

4.4.2 Style Detector Implementation

For all four style elements discussed in Section 4.3.2 detectors were developed, see Appendix A for specific implementation details. We have chosen to make the output of all style detectors discrete using an ordinal scale, as this is known to have a positive effect on SVM performance [28]. Moreover, this weakens individual detector classifiers even more, which has a positive side effect on the classifier combination scheme. To make detectors discrete we use two procedures. On the numerical output, thresholds

are applied. We map categorical output to a discrete number. We optimized all detectors and thresholds based on experiments using the training set. The basic unit of testing and performance assessment within the TRECVID benchmark is the common camera shot segmentation provided by CLIPS-IMAG [115]. We synchronize all discrete detector results, referred to as features, to the granularity of this shot segmentation.

For the layout $\mathcal{L}$ the length of a camera shot was used as a feature that characterizes tempo [3]. Presence of overlayed text, added by the editor at production time, was detected by a text localization algorithm [125]. A microphone segmentation using speech and silence detection was based on the results provided by the LIMSI speech detection system [47]. We obtain a voice over detector by combining the speech segmentation with the camera shot segmentation [146]. The total set of layout detectors is given by: $\hat{\mathcal{L}} = \{$*shot length, overlayed text, silence, voice over*$\}$.

On the content $\mathcal{C}$ a frontal face detector [130] was applied to detect people. For each analyzed frame in a shot we count the number of faces, and for each face we derive one of seven possible locations. In addition, we also measured the average amount of object motion in a camera shot [141]. Based on speaker identification [47] we have been able to identify each of the three most frequent speakers. Each camera shot is checked for the presence on the basis of speech from one of the three. For all rich semantic concepts under consideration, we learned a list of positive and negative correlated keywords using the training set. Stopwords are removed using SMART's English stoplist [123]. Based on the fraction of positive or negative keywords in the text associated with every shot, we labeled a shot as positively correlated, negatively correlated, or undecided. Text strings recognized by using Video Optical Character Recognition [125]* were checked on length and used as input for a named entity recognizer [161]. The total set of content detectors is given by: $\hat{\mathcal{C}} =\{$*faces, face location, object motion, frequent speaker, positive keywords, negative keywords, overlayed text length, video text named entity*$\}$.

From the size of detected faces [130] the camera distance used for capture $\mathcal{T}$ was computed. We distinguished between seven types of camera distance, ranging from extreme long shot to extreme close-up. When no face was detected the camera distance was set as unknown. In addition to camera distance, several types of camera work were detected [12], e.g. pan, tilt, zoom, and so on. Each camera work feature was either present or not. Finally, for capture we also computed the amount of camera motion [12], which was either high, medium, or low. The total set of capture detectors is given by: $\hat{\mathcal{T}} = \{$*camera distance, camera work, camera motion*$\}$.

The possibilities for detectors of context $\mathcal{S}$ are endless. Therefore, we restricted ourselves to spatial context only. For an approach to include temporal context we refer to [141]. Both ABC and CNN news contain many commercials. Although they may contain monologues of people promoting a product, weather related content, and even sporting events, we should not label commercials as such. Therefore, we applied a context detector that is able to detect commercials [60]. News anchors also share many characteristics with news subject monologues, it is therefore important

*For CNN the ticker tape with stock information was ignored.

that we can distinguish anchors to circumvent a false interpretation. Moreover, anchors aid in the detection of studio setting. We applied an anchor detector to stress this importance [60]. For the same reasons we developed a news reporter detector. Reporters were recognized by fuzzy matching of strings obtained from the transcript and Video Optical Character Recognition with a database of names of CNN and ABC affiliates. The set of context detectors is given by: $\hat{\mathcal{S}} = \{$*commercial, news anchor, news reporter*$\}$.

Based on a concatenation of $\left\{\hat{\mathcal{L}}, \hat{\mathcal{C}}, \hat{\mathcal{T}}, \hat{\mathcal{S}}\right\}$ into a style vector $\vec{s}_i$ we are able to train a style model. Since we aim for detection of four rich semantic concepts, and we want to exploit context, we have to define order. The order was chosen based on domain knowledge, see also Section 4.3.3. We chose sporting event as last one, because we used a limited set of specific detectors for this semantic concept. For detection of non-studio setting, both weather news and news subject monologues are useful. Hence, we chose non-studio setting as third. The order of weather news and news subject monologues is not important; we chose to detect news subject monologues first. We added all concepts iteratively to the context. We assign a rich semantic concept to a style vector, or not, based on the probability $p(\omega|\vec{s}_i)$ for each individual style model. Where we use a threshold of 0.5 on $p(\omega|\vec{s}_i)$.

4.5 Results

4.5.1 Evaluation Criteria

NIST allows all groups that participate in TRECVID to submit 10 runs of at most 2000 camera shots for each of the 17 semantic concepts. NIST evaluates all runs. For evaluation the *precision at 100* and *average precision* is used. This former value gives the fraction of correct shots within the first 100 retrieved results. Let $L^k = \{l_1, l_2, \ldots, l_k\}$ be a ranked version of the answer set A. Then precision at 100 is defined as:

$$precision\ at\ 100 = \frac{1}{100}\sum_{k=1}^{100}\lambda(l_k) \tag{4.4}$$

where indicator function $\lambda(l_k) = 1$ if l_k is an element of the result set R and 0 otherwise. The average precision is a single-valued measure that corresponds to the area under a recall-precision curve This value is the average of the precision over all relevant shots. This metric favors highly ranked relevant shots. At any given rank k let $R \cap L^k$ be the number of relevant shots in the top k of L. Then average precision is defined as:

$$average\ precision = \frac{1}{R}\sum_{k=1}^{A}\frac{R \cap L^k}{k}\lambda(l_k) \tag{4.5}$$

Within TRECVID the average precision is used as the basic metric to evaluate the conducted experiments. However, to reduce labor-intensive manual judgments of all submitted runs, a pooled ground truth, P, is used. From each submitted run a fixed

number of ranked shots is taken, these are combined into a list of unique shots. Every submission is then evaluated based on the results of assessing this merged subset, i.e. instead of using R in (4.5), P is used, where $P \subset R$.

This is a fair comparison for submitted runs, since it assures that for each submitted run at least a fixed number of shots is evaluated in the important top of the ranked list. However, for new runs, evaluation based on the pooled ground truth is unfair. Since it is very likely that within the fixed number of shots in the top of the new list a number of shots are retrieved that were not evaluated before, and hence have a negative influence on average precision. Therefore, new runs should also be judged to the same depth as the others, and unknown shots should be labeled and added to a new pooled ground truth, G, where $P \subset G \subset R$. Average precision can then by recalculated, using G, for both original submitted runs and new runs.

4.5.2 Influence of Style on Detection of Rich Semantic Concepts

To gain insight in the importance of style for produced video analysis, we trained classifiers for the various rich semantic concepts using style-based analysis. In addition, we trained classifiers for each concept using the four style elements in isolation. This resulted for each of the four rich semantic concepts in a classifier based on layout, content, capture, context, or style. For all classifier combinations, we evaluated the precision at 100. In Fig. 4.3 we plotted the number of hits as a function of the number of shots judged.

The graphs show that for all rich semantic concepts, it is the combination of style elements that yields the best results. The precision at 100 scores are respectively 0.94 for news subject monologues, 0.98 for non-studio setting, 0.85 for sporting event, and 0.99 for weather news. As expected, content is especially strong in identifying a small subset of concept instances that have low variability in their multimodal content. When this set is exhausted performance drops. This is especially prevalent for sporting events, as we only used a limited number of sport specific content detectors for this concept. For weather news an analysis based on content only achieves a good precision at 100 accuracy, but here a combined style analysis yields the best result also. Besides content, capture is an important style element. The graphs for news subject monologue, non-studio setting, and to lesser extent sporting events, support this observation. For weather news, the current set of capture detectors play no role of importance. Layout is somewhat useful in isolation when aiming for detection of news subject monologues and non-studio setting. For sporting event and weather news, layout is less useful. The current set of context detectors is too limited. Except for non-studio setting, usage of context in isolation is not able to classify rich semantics. The results support our claim that produced video analysis should exploit all style elements to infer rich semantics.

When we take a closer look on the results, we conclude that the news subject monologue detector achieves a high accuracy. In part, the individual detectors, e.g. the face detector, will contribute to this. The influence of style, however, is evident. Capture, for example, is a very informative style element for news subject monologues. Typically, an author records news subject monologues in close-up with a static camera.

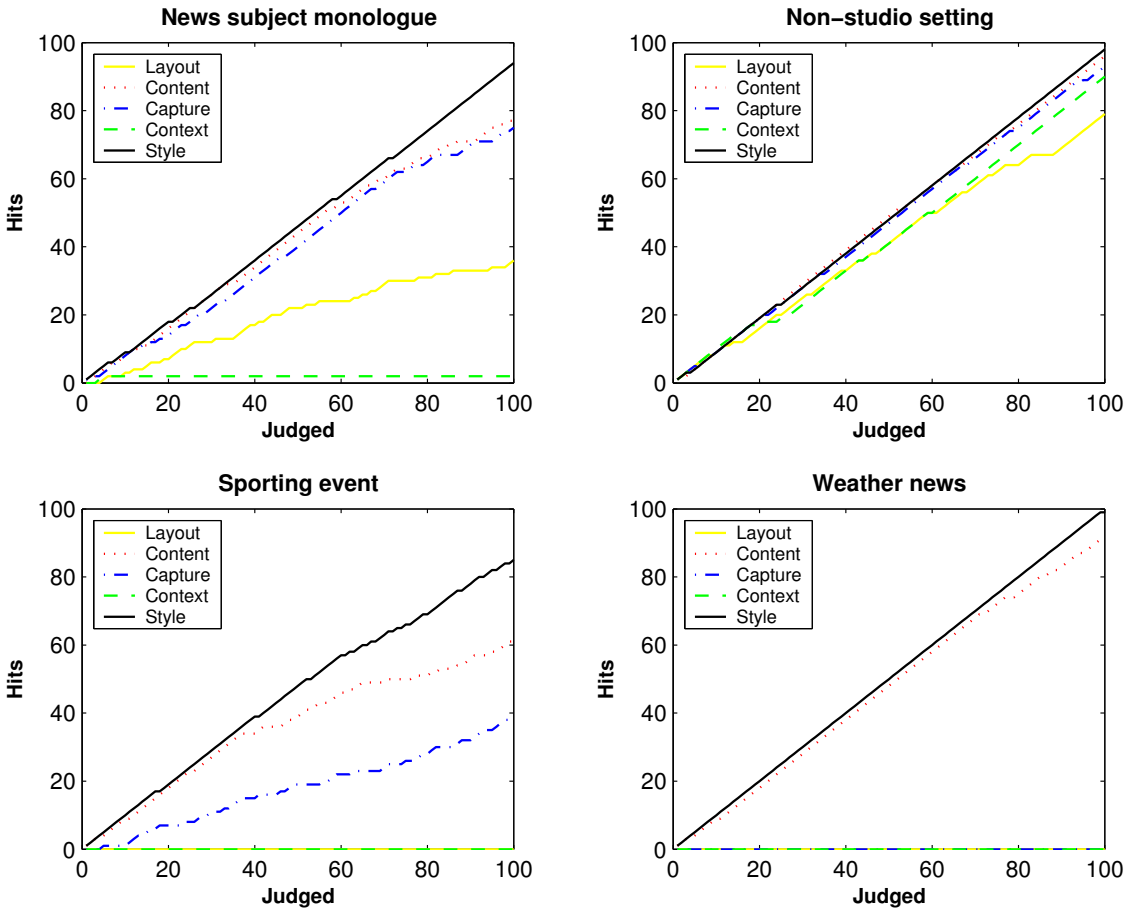


Figure 4.3: Number of hits plotted as function of the number of camera shots judged, for the first 100 results of each analyzed rich semantic concept.

Other concepts, e.g. anchors and reporters, share similarities in content and capture with news subject monologues. Context is therefore required to reduce the number of false positives. By adding layout elements, specifically presence of overlayed text, results improve even further. Another interesting observation stems from analyzing the top 100 results for each style element. Shots from ABC dominate results for content and layout, whereas shots from CNN dominate the top 100 lists for capture and context. This suggests that the authors of both news broadcasts stress different style elements for production of news subject monologues.

We can classify weather news accurately by using content detectors only. We explain this behavior by the textual keywords that we learned for this concept. Weather news has a specific and limited vocabulary; detection of this concept is therefore easy, based on textual content only. It is however, again, the combination of style elements that yields the best results. In contrast to ABC, CNN has a separate weather news report in each broadcast. This makes detection easier, since there is a large similarity in style between the various weather reports. It is therefore not surprising that shots from CNN dominate detection results. Weather news in ABC is much harder

Table 4.1: New pooling and judging statistics for news subject monologue (NSM), non-studio setting (NSS), sporting event (SE), and weather news (WN) after our evaluation.

	NSM	NSS	SE	WN
Pooled depth	100	350	150	100
Unlabeled	28	324	66	0
Judged true	28	304	30	0
Original true	266	2429	585	166
New true	294	2733	615	166

to detect.

Non-studio setting is relatively easy to detect in both ABC and CNN news by all approaches. The fact that this class of concepts is rather large accounts for this behavior. Detection is possible with all four style elements in isolation, even by context alone. This can be explained because non-studio setting is defined by what it is not. Hence, inclusion of anchors and weather news already reduces the number of possible false detections considerably. Analysis of results by combined style analysis shows that adding news subject monologues to the set of context detectors has a very positive influence on correctly detected non-studio setting concepts. Most news subject monologues are produced on location and are therefore not set in a broadcast studio.

Content detectors that are strong indicators for sporting events in our current implementation are sport specific keywords, a large amount of object motion, and absence of frontal faces. The type of camera work used for capture, also aids in correct classification of sporting events. Context and layout are not useful in isolation for detection of sporting events. Similar to weather news, CNN broadcasts sporting events in a separate report. This similarity in style makes detection of sporting events easier for CNN than ABC.

Results on sporting events also show the influence of genre on our framework. Sport is a generic genre that contains a large variety of sub-genres, like basketball, football, and golf. Although similarities in style exist between produced broadcasts of these events, specific sport sub-genres may have large differences in individual style elements. Object and camera motion characteristics for basketball for example, are more similar to football, than to golf. A generic sporting event classifier will therefore profit from a large pool of context detectors for specific sport sub-genres.

In terms of precision, content is the most dominant style element for all four rich semantic concepts. This is visible in Fig. 4.3, where content parallels style for the highest ranked results. This makes analysis based on content useful for applications that require only a limited set of correct results. However, when a large set of correct results is required, one should use all style elements. As can be observed from Fig. 4.3, it is the combination of style elements that strengthens recall, and overall performance.

Other rich semantic concepts, for which the proposed framework is likely to per-

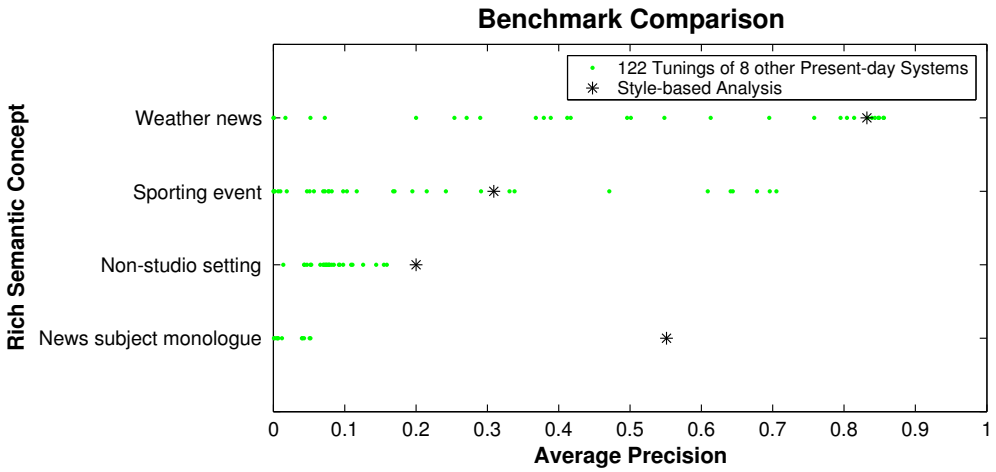


Figure 4.4: Comparison of style-based analysis results with 8 other present-day indexing systems.

form well, are specific instances of the ones discussed. We can split news subject monologue, for example, into public speech, press conference, and interview. In addition, detection of political reporters and financial news anchors should also be possible. Semantic concepts that only have similarity in content, e.g. boat, train, and building are much harder to detect, since our framework works particularly well if similarity in more than one style element exists.

4.5.3 Benchmark Comparison

In our second experiment we compared results of our style-based produced video indexing framework with 8 present-day systems participating in TRECVID 2003. A total of 122 system settings were submitted to TRECVID 2003 for the four rich semantic concepts considered, including the work of [10, 155]. Since our experiments were performed after the 2003 TRECVID benchmark, we had to assure for a fair comparison that we judged at least the same number of camera shots as TRECVID [134]. We used the procedure explained in Section 4.5.1. Results of this evaluation are summarized in Table 4.1. Based on the new pooled ground truth we evaluated average precision for our concepts, and recalculated the average precision for all other systems and their various settings.

The results are summarized in Fig. 4.4. Our framework works particularly well for news subject monologues, improving upon the other approaches by more than a factor ten, see also [146]. Clearly demonstrating the potential of style for detection of rich semantic concepts. For non-studio setting, our method is slightly better then the other systems. Although other systems obtain a twice as good average precision performance for sporting event, our framework works surprisingly well on this concept. This is especially surprising if the limited number of sport-specific detectors in the current implementation is taken into account. For example, we did not use the fact that we can distinguish sporting events based on a large uniform visual setting like

grass or ice. The average precision results for weather news are comparable among the best approaches. The benchmark results show that style-based analysis allows for generic indexing of rich semantic concepts with performance that is comparable to the state-of-the-art.

4.6 Conclusion

When producing a video document, an author uses style to express an intention. Thus, when aiming for semantic analysis of produced video, style is a necessity. For produced video indexing, we identified four different roles, and mapped these roles to four independent style elements within a generic framework. The framework starts with the definition of a set of detectors for each of the style elements considered, i.e. layout, content, capture, and context. To combine style detector results, and learn the rich semantics; the framework utilizes a classifier combination scheme. This scheme facilitates enrichment of semantics by iteratively updating the context. By combining the style detectors in an iterative classifier combination scheme, the framework allows for rich semantic indexing.

Experiments with the four style elements on four rich semantic concepts demonstrate that style-based analysis allows for generic indexing of rich semantic concepts in video. For all analyzed concepts, content is the most important style element. In terms of performance, content is specifically useful when aiming for good precision on a limited set of retrieved items. However, the combination of style elements yields the best overall performance for both precision and recall. This makes our framework a good candidate when aiming for retrieval of a large set of items.

In addition to these results, we performed an experiment on the 2003 TRECVID benchmark, in which we compare our work with competing approaches. The results show that the proposed framework obtains an accuracy favorable in detection of news subject monologues, non-studio setting, and weather news and only lagging behind to dedicated sporting event detection algorithms. We consider this another strong indicator of the approach.

The proposed framework is applicable to any archive of produced video. However as a consequence of the approach, to acquire additional discriminatory power the video documents in the archive should have similarity in genre or their author. We obtain optimal results when we learn separate style models for individual authors within a specific genre.

Apart from content and context, the set of possible style detectors is almost complete. For future research, we therefore aim to augment the lexicon of detectable rich semantic concepts. We believe this is achievable by extending the set of (visual) content detectors related to objects and setting. We are convinced that their impact on detection of rich semantic concepts in video documents will boost progress in multimedia analysis.

Chapter 5

The Semantic Value Chain: A Unifying Architecture for Generic Indexing of Multimedia Archives

To facilitate semantic access to multimedia archives, we propose the semantic value chain. As opposed to most current methods, the semantic value chain allows for semantic video indexing using a generic approach. While doing so, it unifies the most successful existing video indexing methods into a common architecture. The semantic value chain extracts semantic concepts from video based on three consecutive analysis steps. The chain starts in the content link. In this link, we follow a data-driven approach of indexing semantics. The style link is the second link. Here we tackle the indexing problem by viewing a video from the perspective of production. Finally, in the context link we view semantics in context. We learn an optimal configuration of analysis links, on a per-concept basis, to arrive at a technique taxonomy for semantic concept detectors. To show the generality of the proposed approach we develop detectors for a lexicon of 32 concepts. In addition, we evaluate the semantic value chain against the 2004 NIST TRECVID video retrieval benchmark, using a news archive of 184 hours. Top ranking performance in the semantic concept detection task indicates the merit of the semantic value chain for generic indexing of multimedia archives.

5.1 Introduction

Query-by-keyword forms the foundation for machine-based interaction of humans with text repositories. Elaborating on the success of text-based search engines, the query-by-keyword paradigm is also gaining momentum in multimedia retrieval scenarios. For multimedia archives it is hard to achieve effective access, however, when based on keywords that appear in the text only. Video archives require semantic access where all modalities can contribute to the concept.

For semantic access, multimodal indexing is inevitable. For well-defined semantic concepts sophisticated and specialized versions of such indexing methods are available, see [97,142] for an overview. In contrast to their textual counterparts, generic methods for semantic indexing in multimedia are neither generally available, scalable in their computational needs, nor robust in their performance. As a consequence, semantic access to multimedia archives is limited still. Therefore, a new semantic video indexing methodology is required when aiming for semantic access to multimedia archives.

The main problem for any semantic video indexing approach is the semantic gap between data representation and their interpretation by humans [136]. In an effort to limit the size of the semantic gap, many video indexing approaches have focused on specific semantic concepts with a small intra-class and large inter-class variability of content. Concepts like *sunsets* [137] and *news anchors* [175], have become icons for specific video indexing methodologies. Although specific methods have aided in achieving substantial progress, this road is the hard way when compared to the thousands and thousands of concepts which are needed. It is simply impossible to bridge the semantic gap by designing a tailor-made solution for each concept.

In this Chapter we propose a generic approach for semantic indexing of multimedia archives. While doing so, we do not ignore the vast amount of work performed in developing specialized concept detection methods [7, 13, 54, 60, 137, 161, 175]. If we measure success of these methods in terms of benchmark performance, Informedia [60, 161] stands out. They focus on combining techniques from computer vision, speech recognition, natural language understanding, and artificial intelligence into a video indexing and retrieval environment. This has resulted in a large set of isolated and specialized concept detectors [60]. We build our generic approach in part on their specialized concept detection methods, but we do not use them in isolation.

In contrast to specialized concept detection methods, generic semantic indexing methodologies from video are scarce. We discuss three good examples of generic semantic index approaches [10, 41, 145].

In the first one, Fan *et al.* [41] propose the *ClassView* framework. This framework combines hierarchical semantic indexing with hierarchical retrieval. At the lowest level, the framework supports indexing of shots into concepts based on a large set of low-level visual features. At the second level, Bayes' rule classifies concepts further into semantic clusters. By assigning shots to a hierarchy of concepts, the framework supports queries based on semantic and visual similarity. This allows for hierarchical retrieval. As the authors indicate, the framework will provide more meaningful results if it would support multimodal analysis. We aim for generic semantic indexing also, but we include multimodal analysis from the beginning.

In the second one [10], the authors propose a system for generic semantic indexing using a detection pipeline. The pipeline starts with feature extraction, followed by consecutive aggregations on features, multiple modalities, and concepts. Finally, the pipeline optimizes the result by rule-based post filtering. We explain the success of their system by the fact that all modules in the pipeline select the best of multiple hypotheses, and the exhaustive use of machine learning. Moreover, the authors were among the first to recognize that semantic indexing profits substantially from context. We adopt and extend their ideas related to hypothesis selection, machine learning, and the use of context for semantic indexing.

All of the above methods ignore the important influence of style in the analysis process. In addition to content and context, as used in the above references, we identify layout and capture in [145] as important factors for semantic indexing of produced video. Based on these style elements, we propose a generic framework for produced video indexing; combining four sets of style detectors in an iterative semantic classifier. Results indicate that the method obtains high accuracy for rich semantic concepts, rich meaning that style is exploited in many ways. The framework is less suited for concepts that are not stylized. In this Chapter, we generalize the idea of using style for semantic indexing.

We propose a generic approach for semantic indexing, we call the *semantic value chain*. It unifies the most successful approaches for semantic video indexing [10, 60, 145, 161] into a common architecture. The architecture is build on several specialized detectors, multimodal analysis, hypothesis selection, and machine learning. Furthermore, it covers the notions of content, style, and context. To demonstrate the effectiveness of the semantic value chain, the semantic indexing experiments are evaluated within the 2004 NIST TRECVID video retrieval benchmark [102].

The organization of this Chapter is as follows. First, we introduce the TRECVID benchmark in Section 5.2. Our system architecture for generic semantic indexing is presented in Section 5.3. We present results in Section 5.4.

5.2 TRECVID Benchmark

Evaluation of multimedia systems has always been a delicate issue. Due to copyrights and the sheer volume of data involved, multimedia archives are fragmented and mostly inaccessible. Therefore, comparison of systems has traditionally been difficult, often impossible even. To accommodate these hardships NIST started organizing the TRECVID video retrieval benchmark. The benchmark aims to promote progress in video retrieval via open, metrics-based evaluation [102]. Tasks include camera shot segmentation, story segmentation, semantic concept detection*, and several search tasks. We have participated in the semantic concept detection task of the 2004 NIST TRECVID video retrieval benchmark.

*TRECVID refers to this task as the feature extraction task, to prevent misunderstanding with feature extraction as defined in the semantic value chain we refer to it as the semantic concept detection task.

5.2.1 Multimedia Archive

The video archive of the 2004 TRECVID benchmark extends on the data set used in 2003. The archive is composed of 184 hours of ABC World News Tonight and CNN Headline News and is recorded in MPEG-1 format. The development data consists of the archive used in 2003. It contains approximately 120 hours covering the period of January until June 1998. The 2004 test data contains the remaining 64 hours, covering the period of October until December 1998. Together with the video archive, CLIPS-IMAG [115] provided a camera shot segmentation. We evaluate semantic indexing within the TRECVID benchmark, to demonstrate the effectiveness of the semantic value chain for semantic access to multimedia archives.

5.2.2 Evaluation Criteria

Participation in TRECVID is based on the submission of results for one of the concepts in the semantic concept detection task. Where a submission, or run, contains a ranked list of at most 2000 camera shots per semantic concept. For each concept, participants are allowed to submit 10 runs.

To determine the accuracy of submissions we use *average precision*, and *precision at 100*, following the standard in TRECVID evaluations. The average precision is a single-valued measure that corresponds to the area under a recall-precision curve. This value is the average precision over all relevant judged shots. Let $L^k = \{l_1, l_2, \ldots, l_k\}$ be a ranked version of the answer set A. At any given rank k let $R \cap L^k$ be the number of relevant shots in the top k of L, where R is the total number of relevant shots. Then average precision is defined as:

$$average\ precision = \frac{1}{R} \sum_{k=1}^{A} \frac{R \cap L^k}{k} \lambda(l_k) \tag{5.1}$$

where indicator function $\lambda(l_k) = 1$ if $l_k \in R$ and 0 otherwise. As the denominator k and the value of $\lambda(l_k)$ are dominant in determining average precision, it can be understood that this metric favors highly ranked relevant shots.

TRECVID uses a pooled ground truth P, to reduce labor-intensive manual judgments of all submitted runs. They take from each submitted run a fixed number of ranked shots, which is combined into a list of unique shots. Every submission is then evaluated based on the results of assessing this merged subset, i.e. instead of using R in (5.1), P is used, where $P \subset R$.

Apart from average precision, we also report the precision at depth 100 in the result set. This value gives the fraction of correct shots within the first 100 retrieved results.

5.3 Semantic Value Chain Analysis

The essence of produced video, like broadcast news, is that an author creates it. Before creation, the author starts with a semantic idea: an interplay of concepts and

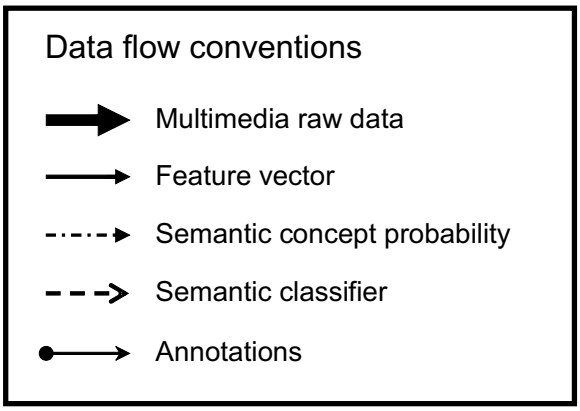


Figure 5.1: Data flow conventions as used in this Chapter. Different style of arrows indicate different data structures.

context. To stress the semantics of the message, and to guide the audience in its interpretation, the author combines various style elements. The video aims at an effective semantic communication. Hence, the core of semantic indexing is to reverse this authoring process [142]. We follow this path to arrive at a system architecture for semantic indexing in video.

Before we elaborate on the video indexing architecture, we first define a lexicon Λ_S of 32 semantic concepts. The lexicon includes all 10 concepts defined in the semantic concept detection task at hand. We choose the additional concepts based on the indices described in [142], as well as anticipated positive influence on the result of the 10 benchmark concepts. The following concepts form the semantic concept lexicon:

- $\Lambda_S = \{$*airplane take off, American football, animal, baseball, basket scored, beach, bicycle, Bill Clinton, boat, building, car, cartoon, financial news anchor, golf, graphics, ice hockey, Madeleine Albright, news anchor, news subject monologue, outdoor, overlayed text, people, people walking, physical violence, road, soccer, sporting event, stock quotes, studio setting, train, vegetation, weather news*$\}$;

The lexicon contains both general concepts, like *people*, *car*, and *beach*, as well as specific concepts such as *airplane take off* and *news subject monologue*. With the proposed system architecture, we aim to detect all 32 concepts.

The semantic value chain is composed of three links. It follows the reverse authoring process. Each link in the chain detects semantic concepts. In addition, one can exploit the output of a link in the chain as the input for the next one. The semantic value chain starts in the *content link*. In this link, we follow a data-driven approach of indexing semantics. The *style link* is the second link. Here we tackle the indexing problem by viewing a video from the perspective of production. This link aids especially in indexing of rich semantics. Finally, to enhance the indexes further, in the *context link*, we view semantics in context. One would expect that some concepts,

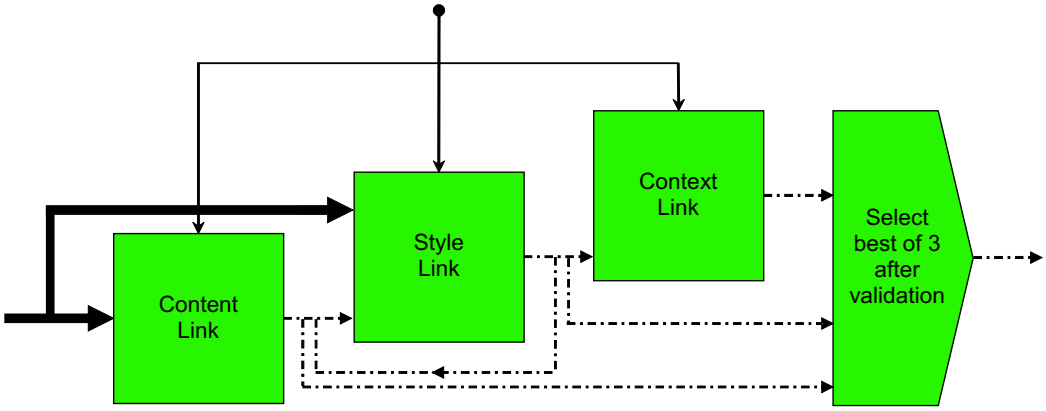


Figure 5.2: The semantic value chain for one concept, using the conventions of Fig. 5.1.

like *vegetation*, have their emphasis on content where the style (of the camera work that is) and context (of concepts like *graphics*) do not add much. In contrast, more complex events, like *people walking*, profit from incremental adaptation of the analysis to the intention of the author. The virtue of the semantic value chain is that it selects the best chain of analysis links on a per-concept basis.

The links in the semantic value chain exploit a common architecture with a standardized input-output model to allow for semantic integration. The conventions to describe the system architecture are indicated in Fig. 5.1. An overview of the semantic value chain is given in Fig. 5.2.

5.3.1 General Architecture

We perceive of semantic indexing in video as a pattern recognition problem. We first need to segment a video. We opt for camera shots, indicated by i, as the basic time frame as it is known to maximize the chance for semantic machine interpretation [35]. Given pattern x, part of a shot, the aim is to detect a semantic concept ω from shot i using probability $p_i(\omega|x_i)$. Each analysis link in the semantic value chain extracts x_i from the data, and exploits a learning module to learn $p_i(\omega|x_i)$ for all ω in the semantic lexicon Λ_S. We exploit supervised learning to learn the relation between ω and x_i. The development data of the multimedia archive, together with labeled samples, are for learning classifiers. The other data, the test data, are set aside for testing. The general architecture for supervised learning in each link is illustrated in Fig. 5.3.

Supervised learning requires labeled examples. In part, we rely on the provided ground truth of TRECVID 2003 [83, 102]. We remove the many errors from this annotation effort. It is extended manually to arrive at a reliable ground truth for all concepts in lexicon Λ_S. We split the development data a priori into a non-overlapping training set and validation set to prevent overfitting of classifiers in the semantic value chain. The training set contains 85% of the development data, the validation set contains the remaining 15%. The number of annotated examples in the training set

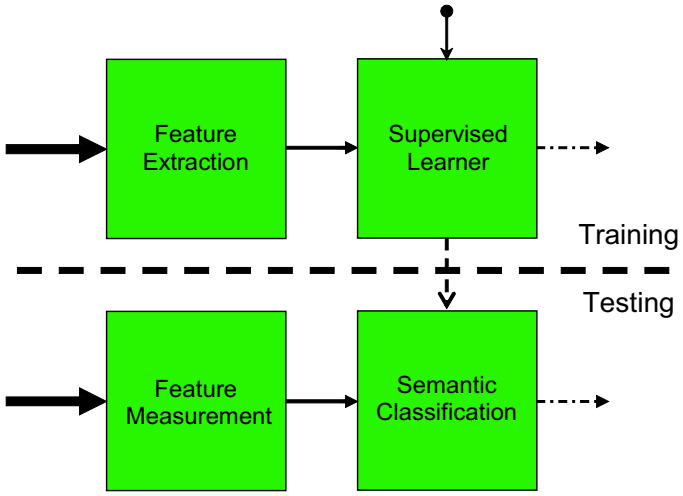


Figure 5.3: General link architecture in the semantic value chain, using the conventions of Fig. 5.1.

and the validation set for each concept are summarized in Table 5.1.

We choose from a large variety of supervised machine learning approaches to obtain $p_i(\omega|x_i)$. For our purpose, the method of choice should be capable of handling video documents. To that end, ideally it must learn from a limited number of examples, it must handle unbalanced data, and it should account for unknown or erroneously detected data. In such heavy demands, the Support Vector Machine (SVM) framework [28, 158] has proven to be a solid choice [10, 141]. The usual SVM method provides a margin in the result. We prefer Platt's conversion method [113] to achieve a posterior probability of the result. SVM classifiers thus trained for ω on development data, result in an estimate $p_i(\omega|x_i, \vec{q})$, where $\vec{q}$ are parameters of the SVM yet to be optimized.

The influence of the SVM parameters on concept detection is significant [95]. We obtain good parameter settings for a classifier, by using an iterative search on a large number of SVM parameter combinations. We measure average precision performance of all parameter combinations and select the combination that yields the best performance, $\vec{q}^*$. Here we use 3-fold cross validation [68] to prevent overfitting of parameters. The result of the parameter search over $\vec{q}$ is the improved model $p_i^*(\omega|x_i, \vec{q}^*)$. In the following we drop $\vec{q}^*$ where obvious.

This concludes the introduction of the general architecture.

5.3.2 Content Link

We view of video in the content link from the data perspective. In general, three data streams or modalities exist in video, namely the auditory modality, the textual modality, and the visual one. As speech is often the most informative part of the auditory source, we focus on visual features, and on textual features obtained from

Table 5.1: Semantic concepts and the number of labeled examples used for the training set and the validation set.

Semantic Concept	Training	Validation	Semantic Concept	Training	Validation
Weather news	267	40	Golf	73	23
Stock quotes	135	28	People	2022	367
News anchor	2032	367	American football	24	9
Overlayed text	133	16	Outdoor	3904	791
Basket scored	556	89	Car	813	193
Graphics	551	97	Bill Clinton	503	130
Baseball	382	61	News subject monologue	1994	364
Sporting event	1181	224	Animal	702	123
People walking	995	181	Road	748	182
Financial news anchor	182	32	Beach	218	56
Ice hockey	187	43	Train	111	33
Cartoon	310	67	Madeleine Albright	94	2
Studio setting	2565	428	Building	2571	442
Physical violence	1416	289	Airplane take off	463	80
Vegetation	833	146	Bicycle	144	25
Boat	286	41	Soccer	33	8

transcribed speech. After modality specific data processing, we combine features in a multimodal representation. The data flow in the content link is illustrated in Fig. 5.4.

Visual Analysis

In the visual modality, we aim for segmentation of an image frame f into regional visual concepts. Ideally, a segmentation method should result in a precise partitioning of f according to the object boundaries, referred to as strong segmentation. However, weak segmentation, where f is partitioned into internally homogenous regions within the boundaries of the object, is often the best one can hope for [136]. We obtain a weak segmentation based on a set of visual feature detectors. Prior to segmentation we remove the border of each frame, including the space occupied by a possible ticker tape. The basis of feature extraction in the visual modality is weak segmentation.

Invariance was identified in [136] as a crucial aspect of a visual feature detector, e.g. to design features which limit the influence of accidental recording circumstances. We use invariant visual features to arrive at weak segmentation, as the conditions under which semantic concepts appear in large multimedia archives may vary greatly.

The feature extraction procedure we adhere to, computes per pixel a number of invariant features in vector $\vec{u}$. This vector then serves as the input for a multi-class SVM [28] that associates each pixel to one of the regional visual concepts defined in a visual concept lexicon Λ_V, using a labeled training set. Based on Λ_S, we define the

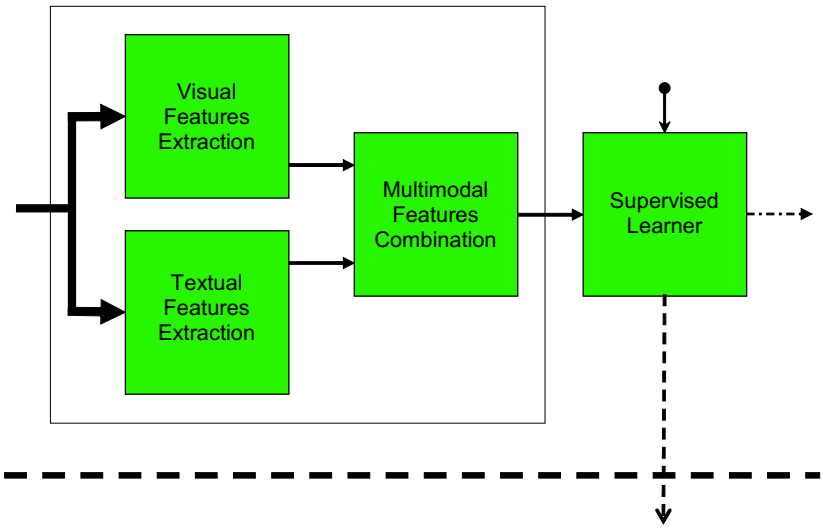


Figure 5.4: Feature extraction and classification in the content link, special case of Fig. 5.3.

following set of regional visual concepts:

- $\Lambda_V = \{$*colored clothing, concrete, fire, graphic blue, graphic purple, graphic yellow, grassland, greenery, indoor sport court, red carpet, sand, skin, sky, smoke, snow/ice, tuxedo, water body, wood*$\}$;

As we use invariant features, only a few examples per visual concept class are needed; in practice less then 10 per class. This pixel-wise classification results in the image vector $\vec{w}_f$. Where $\vec{w}_f$ is a weak segmentation of frame f in terms of regional visual concepts from Λ_V, see Fig. 5.5 for an example segmentation.

We use Gaussian color measurements [48] to obtain $\vec{u}$ for weak segmentation. We decorrelate *RGB* color values by linear transformation to the opponent color system [48]. Smoothing the values with a Gaussian filter suppresses acquisition and compression noise. The size of the Gaussian filters is varied to obtain a color representation that is compatible with variations in the target object size. Normalizing each opponent color value by its intensity suppresses global intensity variations. This results in two chromaticity values per color pixel. Furthermore, we obtain rotationally invariant features by taking Gaussian derivative filters and combining the responses into two chromatic gradients. The seven measurements in total, and each calculated over three scales, yield a 21 dimensional invariant feature vector $\vec{u}$ per pixel.

Segmenting image frames into regional visual concepts at the granularity of a pixel is computationally intensive. We estimate that the processing of the entire TRECVID data set would have taken around 250 days on the fastest sequential machine available to us. As a first reduction of the analysis load, we analyze 1 out of 15 frames only. For the remaining image processing effort we apply the Parallel-Horus software architecture [131]. This architecture, consisting of a large collection of low-level image processing primitives, allows the programmer to write sequential applications with efficient parallel execution on commonly available commodity clusters. Application of

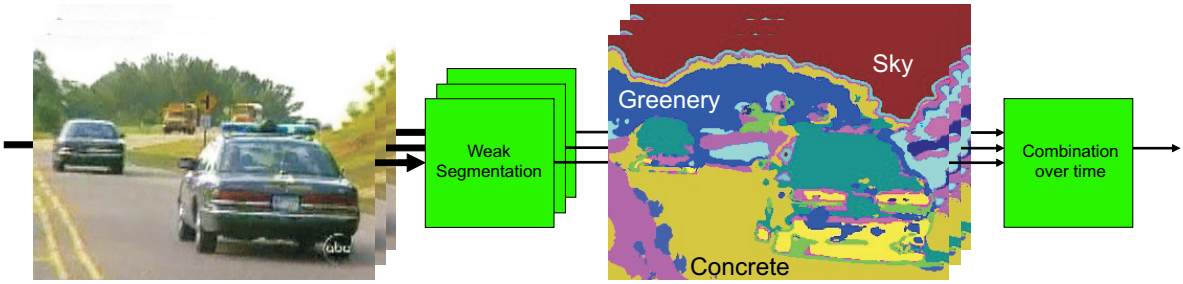


Figure 5.5: Computation of the visual features, see Fig. 5.4, is based on weak segmentation of an image frame into regional visual concepts. A combination over time is used to select one frame as representative for the shot.

Parallel-Horus, in combination with a distributed cluster consisting of 200 dual 1-Ghz Pentium-III CPUs [14], reduced the processing time to less than 60 hours [131].

The features over time are combined into one vector for the shot i. Averaging over individual frames is not a good choice, as the visual representation should remain intact. Instead, we opt for a selection of the most representative frame or visual vector. To decide which f is the most representative for i, weak segmented image $\vec{w}_f$ is the input for an SVM that computes a probability $p^*_f(\omega|\vec{w}_f)$. We select $\vec{w}_f$ that maximizes the probability for a concept from Λ_S within i, given as:

$$\vec{v}_i = \arg\max_{f \in f_i} p^*_f(\omega|\vec{w}_f) \tag{5.2}$$

The visual vector $\vec{v}_i$, containing the best weak segmentation, is the final result of the visual analysis.

Textual Analysis

In the textual modality, we aim to learn the association between uttered speech and semantic concepts. A detection system transcribes the speech into text. From the text we remove the frequently occurring stopwords. After stopword removal, we are ready to learn semantics.

To learn the relation between uttered speech and concepts, we connect words to shots. We make this connection within the temporal boundaries of a shot. We derive a lexicon of uttered words that co-occur with ω using the shot-based annotations of the development data. For each concept ω, we learn a separate lexicon, Λ^ω_T, as this uttered word lexicon is specific for that concept. We modify the procedure for Person X concepts, i.e. *Madeleine Albright* and *Bill Clinton*, to optimize results. In broadcast news, a news anchor or reporter mentions names or other indicative words just before or after a person is visible. To account for this observation, we stretch the shot boundaries with five seconds on each side for Person X concepts. For these concepts, this procedure assures that the textual feature analysis considers more textual content. For feature extraction we compare the text associated with each shot with Λ^ω_T. This comparison yields a text vector $\vec{t}_i$ for shot i, which contains the histogram of the words in association with ω.

Multimodal Analysis and Classification

The result of the content link is a multimodal vector $\vec{m}_i$ that integrates all unimodal results. We concatenate the visual vector $\vec{v}_i$ with the text vector $\vec{t}_i$, to obtain $\vec{m}_i$. After this modality fusion, $\vec{m}_i$ serves as the input for a supervised learning module. To optimize parameter settings, we use 3-fold cross validation on the training set. The content link associates probability $p_i^*(\omega|\vec{m}_i)$ with a shot i , for all ω in Λ_S.

5.3.3 Style Link

In the style link we conceive of a video from the production perspective. Based on the four roles involved in the video production process [145], this link analyzes a video by four related style detectors. Layout detectors analyze the role of the editor. Content detectors analyze the role of production design. Capture detectors analyze the role of the recording unit. Finally, context detectors analyze the role of the scenario writer, see Fig. 5.6. Note that in contrast to the content link, where we learn specific content features from a data set, content features in the style link are generic and independent of the data set.

Style Analysis

We develop detectors for all four style roles as feature extraction in the style link, see Appendix A for specific implementation details. We have chosen to convert the output of all style detectors to an ordinal scale, as this allows for easy fusion.

For the layout $\mathcal{L}$ the length of a camera shot is used as a feature, as this is known to be an informative descriptor for genre [142]. Overlayed text is another informative descriptor. Its presence is detected by a text localization algorithm [125]. To segment the auditory layout, periods of speech and silence are detected based on an automatic speech recognition system [47]. We obtain a voice over detector by combining the speech segmentation with the camera shot segmentation [145]. The set of layout features is thus given by: $\mathcal{L} = \{$*shot length, overlayed text, silence, voice over*$\}$.

As concerns the content $\mathcal{C}$, a frontal face detector [130] is applied to detect people. We count the number of faces, and for each face its location is derived [145]. Apart from faces, we also detect the presence of cars [130]. In addition, we measure the average amount of object motion in a camera shot [141]. Based on speaker identification [47] we identify each of the three most frequent speakers. The camera shot is checked for the presence on the basis of speech from one of the three [145]. The length of text strings recognized by Video Optical Character Recognition [125] is used as a feature [145]. In addition, the strings are used as input for a named entity recognizer [161]. On the transcribed text obtained by the LIMSI automatic speech recognition system, we also apply named entity recognition. The set of content features is thus given by: $\mathcal{C} =\{$*faces, face location, cars, object motion, frequent speaker, overlayed text length, video text named entity, voice named entity*$\}$.

For capture $\mathcal{T}$, we compute the camera distance from the size of detected faces [130, 145]. It is undefined when no face is detected. In addition to camera distance, several types of camera work are detected [12], e.g. pan, tilt, zoom, and so on. Finally,

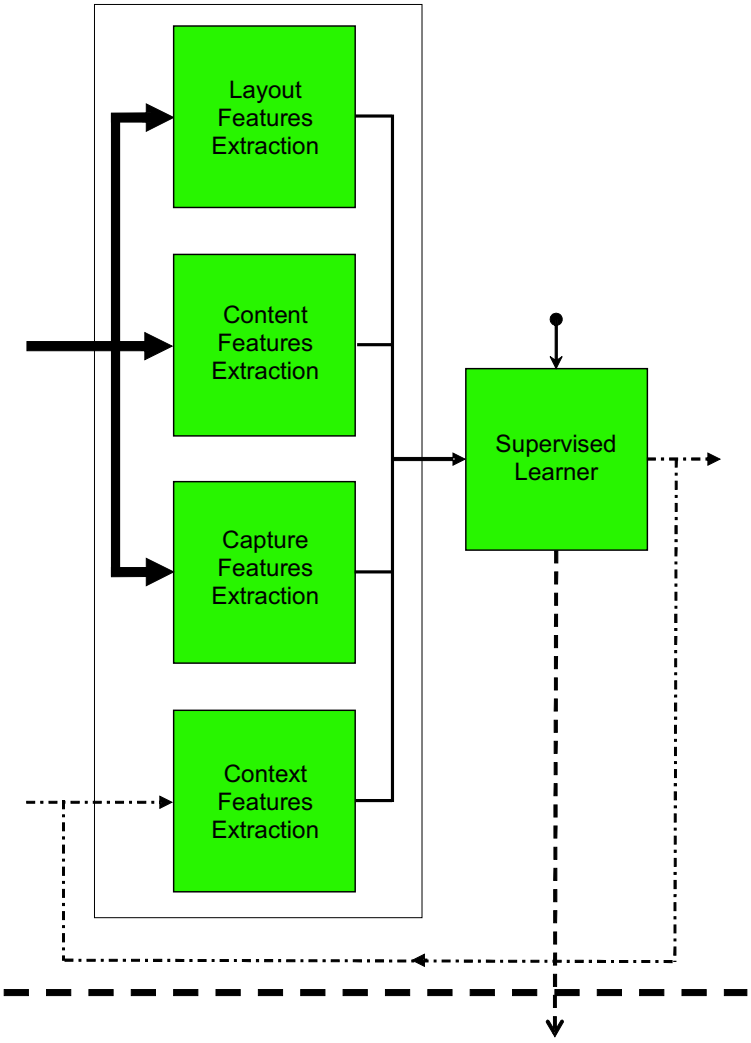


Figure 5.6: Feature extraction and classification in the style link, special case of Fig. 5.3.

for capture we also estimate the amount of camera motion [12]. The set of capture features is thus given by: $\mathcal{T} = \{camera\ distance,\ camera\ work,\ camera\ motion\}$.

The context $\mathcal{S}$ serves to enhance or reduce the correlation between semantic concepts. Detection of *vegetation* can aid in the detection of a *forest* for example. Likewise, the cooccurrence of a *space shuttle* and a *bicycle* in one shot is improbable. As the performance of semantic concept detectors is unknown and likely to vary between concepts, we exploit iteration to add them to the context. The rationale here is to add concepts that are relatively easy to detect first. They aid in detection performance by increasing the number of true positives or reducing the number of false positives. As initial concept we detect news reporters. We recognize news reporters by edit distance matching of strings obtained from the transcript and video text with a database of names of CNN and ABC affiliates [145]. The other concepts that are

added to the context stem from Λ_S. To prevent bias from domain knowledge, we use the performance on the validation set of all concepts from Λ_S in the content link as the ordering for the context. For this ordering we again refer to Table 5.1. To assign detection results for the first and least difficult concept, $\omega_1 = weather\ news$, we rank all shot results on $p_i^*(\omega_1|\vec{m}_i)$. This ranking is then exploited to categorize results for ω_1 into one of five levels. The basic set of context features is thus given by: $\mathcal{S} = \{news\ reporter,\ content\ link\ \omega_1\}$.

The concatenation of $\{\mathcal{L}, \mathcal{C}, \mathcal{T}, \mathcal{S}\}$ for shot i yields style vector $\vec{s}_i$. This vector forms the input for an iterative classifier that trains a style model for each concept in lexicon Λ_S.

Iterative Style Classification

We start from an ordering of concepts in the context, as defined above. The iteration of the classifier begins with concept ω_1. After concatenation with the other style features this yields $\vec{s}_{i,1}$ the first style vector of the first iteration. $\vec{s}_{i,1}$ contains the combined results of the content link and the style link. We classify ω_1 again based on $\vec{s}_{i,1}$. This yields the a posterior probability $p_i^*(\omega_1|\vec{s}_{i,1})$. When $p_i^*(\omega|\vec{s}_i) \geq \delta$ the concept ω_1 is considered present in the style representation. Else, it is considered absent. The threshold δ is set a priori at a fixed value of 0.5. In this process the classifier replaces the feature for concept ω_1, from the content link, by the new feature ω_1^+. The style link adds more aspects of the author influence to the results obtained with the content link. In the next iteration of the classification procedure the classifier adds $\omega_2 = stock\ quotes$ from the content link to the context. This yields $\vec{s}_{i,2}$. As explained above, the classifier replaces the ω_2 feature from the content link by the styled version ω_2^+ based on $p_i^*(\omega_2|\vec{s}_{i,2})$. This iterative process is repeated for all ω in lexicon Λ_S.

We classify all ω in Λ_S again in the style link. As the result of the content link is only one of the many features in our style vector representation in the style link, we also use 3-fold cross validation on the training set to optimize parameter settings in this link. We use the resulting probability as output for concept detection in the style link. In addition, it forms the input for the next link in our semantic value chain.

5.3.4 Context Link

The context link adds context to our interpretation of the video. Our ultimate aim is the reconstruction of the author's intent by considering detected concepts in context.

Semantic Analysis

The style link yields a probability for each shot i and all concepts ω in Λ_S. The probability indicates whether a concept is present. We fuse all these semantic features of the style link for a shot i into a context vector, $\vec{c}_i$, see Fig. 5.7.

From $\vec{c}_i$ we learn relations between concepts automatically. To that end, $\vec{c}_i$ serves as the input for a supervised learning module, which associates a contextual prob-

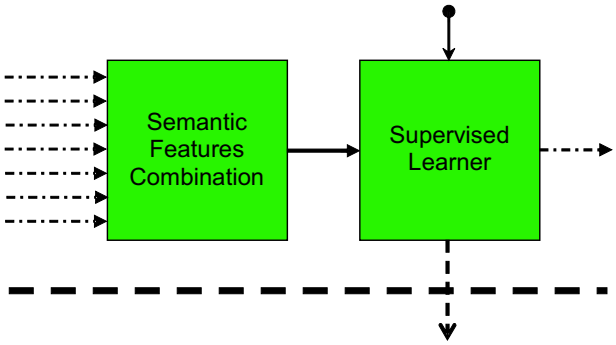


Figure 5.7: Feature extraction and classification in the context link, special case of Fig. 5.3.

ability $p_i^*(\omega|\vec{c}_i)$ to a shot i for all ω in Λ_S. To optimize parameter settings, we use 3-fold cross validation on the previously unused data from the validation set.

The output of the context link is also the output of the entire semantic value chain on video documents. On the way we have included in the semantic value chain, the results of the analysis on raw data, facts derived from production by the use of style features, and a context perspective of the author's intent by using semantic features. For each concept we obtain a probability based on content, style, and context. We select from the three possibilities the one that maximizes average precision based on validation set performance. The semantic value chain provides us with the opportunity to decide whether a one-shot link is best for the concept only concentrating on content, or a two-link classifier increasing discriminatory power by adding style elements to content, or that a concept profits most from a consecutive analysis using content, style, and context.

5.4 Results

5.4.1 Detection of 32 Semantic Concepts

We evaluated detection results for all 32 semantic concepts in each link of the semantic value chain. We report the *precision at 100*, which indicates the number of correct shots within the first 100 results, in Table 5.2.

We observe from the results that the learned best chain (printed in bold) indeed varies over the concepts. The virtue of the semantic value chain is demonstrated by the fact that for 12 concepts, the learning phase indicates it is best to concentrate on content only. For 5 concepts, the semantic value chain demonstrates that a two-step procedure is best (where in 15 cases addition of style features has a marginal positive or negative effect). For 15 concepts, the context link obtains a better result. Context aids substantially in the performance for 5 concepts.

The results demonstrate the virtue of the semantic value chain. Concepts are divided by the link after which they achieve best performance. Some concepts are just content, style does not affect them. In such cases as *American football* there is style-wise too much confusion with other sports to add new value in the chain. Shots

Table 5.2: Test set precision at 100 after the content link, the style link, and the context link from the semantic value chain for a lexicon of 32 semantic concepts. The best chain that is selected for a concept based on validation set experiments is indicated in bold.

Semantic Concept	Content Link	Style Link	Context Link
Weather news	**1.00**	1.00	1.00
Stock quotes	**0.89**	0.77	0.77
News anchor	0.98	0.98	**0.99**
Overlayed text	0.84	**0.99**	0.93
Basket scored	0.24	0.21	**0.30**
Graphics	0.92	0.90	**0.91**
Baseball	**0.54**	0.43	0.47
Sporting event	0.77	**0.98**	0.93
People walking	0.65	0.72	**0.83**
Financial news anchor	0.40	**0.70**	0.71
Ice hockey	**0.71**	0.68	0.60
Cartoon	0.71	0.69	**0.75**
Studio setting	0.95	0.96	**0.98**
Physical violence	0.17	0.25	**0.31**
Vegetation	**0.72**	0.64	0.70
Boat	0.42	0.38	**0.37**
Golf	**0.24**	0.19	0.06
People	0.73	0.78	**0.91**
American football	**0.46**	0.18	0.17
Outdoor	0.62	0.83	**0.90**
Car	0.63	0.81	**0.75**
Bill Clinton	**0.26**	0.35	0.37
News subject monologue	0.55	**1.00**	1.00
Animal	0.37	0.26	**0.26**
Road	0.43	0.53	**0.51**
Beach	0.13	0.12	**0.12**
Train	**0.07**	0.07	0.03
Madeleine Albright	**0.12**	0.05	0.04
Building	**0.53**	0.46	0.43
Airplane take off	0.10	0.08	**0.08**
Bicycle	0.09	**0.08**	0.07
Soccer	**0.01**	0.01	0.00

containing *stock quotes* suffer from a similar problem. Here false positives contain many stylistically similar results like graphical representations of survey and election results. For complex concepts, analysis based on content and style is not enough.

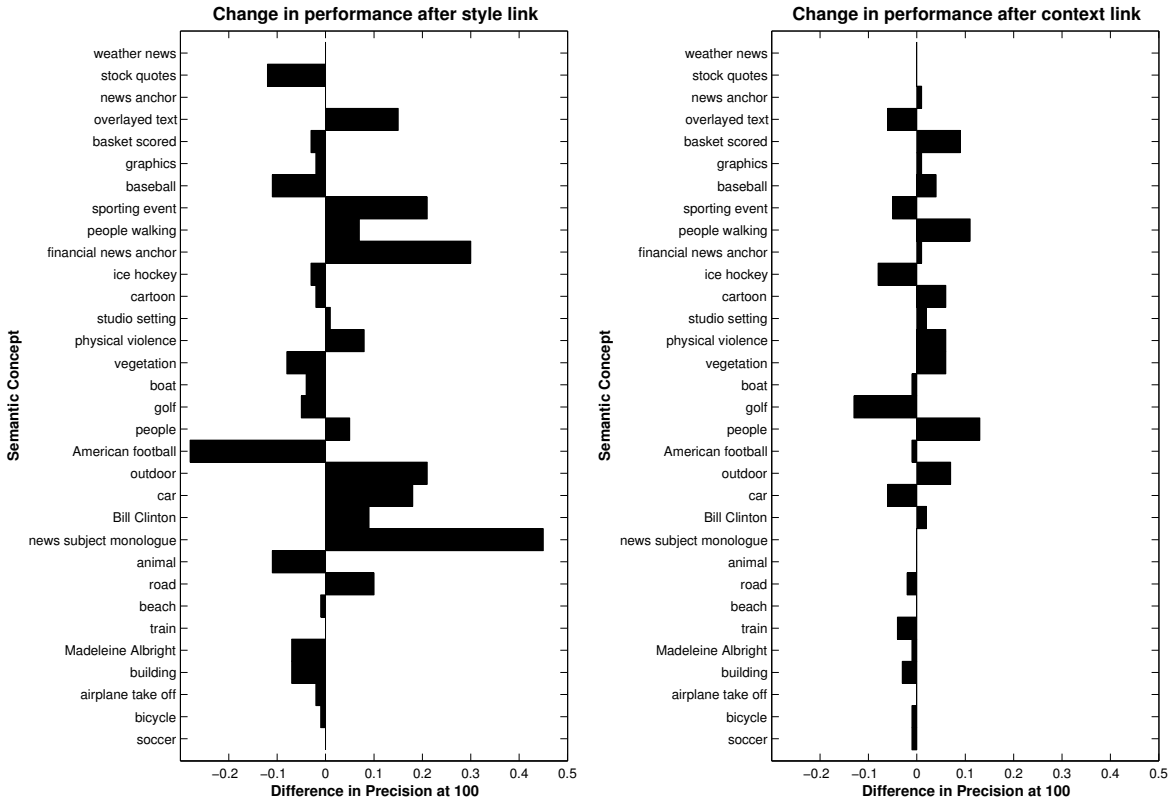


Figure 5.8: Influence of the style link (a) and the context link (b) on precision at 100 performance for a lexicon of 32 semantic concepts.

They require the use of context. The context link is especially good in detecting named events, like *people walking*, *physical violence*, and *basket scored*. The results offer us the possibility to categorize concepts according to the analysis link of the semantic value chain that yields the best performance.

The content link seems to work particularly well for semantic concepts that have a small intra-class variability of content: *weather news* and *news anchor* for example. In addition, this link aids in detection of accidental content like *building*, *vegetation*, *bicycle*, and *train*. However, for some of those concepts, e.g. *bicycle* and *train* the performance is disappointing still. Another observation is, that when one aims to distinguish sub-genres, e.g. *ice hockey*, *baseball*, and *American football*, the content link is the best choice.

After the style link we obtain an increase in performance for 12 concepts, see Fig. 5.8a. Especially when the concepts are semantically rich: e.g. *news subject monologue*, *financial news anchor*, *sporting event*, and *outdoor*, the style helps. As expected, index results in the style link improve on the content link when style is a distinguishing property of the concept and degrade the result when similarity in style

exists between different concepts.

Results after the context link in Fig. 5.8b show that performance increases for 13 concepts. The largest positive performance difference between the context link and the style link occurs for concept *people*. Concept *people* profits from sport-related concepts like *baseball*, *basket scored*, *American football*, *ice hockey*, and *sporting event*. In contrast, *golf* suffers from detection of *outdoor* and *vegetation*. When we detect *golf*, these concepts are also present frequently. The inverse, however, is not necessarily the case, i.e. when we detect *outdoor* it is not necessarily on a golf court. Based on these observations we conclude that, apart from named events, detection results of the context link are similar to those of the style link. Based on presence of semantically related concepts index results improve, but the context link is unable to capture the semantic structure between concepts and for some concepts this is leading to a drop in performance.

The above results show that the semantic value chain facilitates generic video indexing. In addition, the semantic value chain provides the foundation of a technique taxonomy for solving semantic concept detection tasks. The fact that sub-genres like *ice hockey*, *golf*, and *American football* behave similarly indicate the predictive value of the chain for other sub-genres. The same holds for semantically rich concepts like *news subject monologue*, *financial news anchor*, and *sporting event*. We showed that for named events, such as *basket scored*, *physical violence*, and *people walking*, one should apply a detector that is based on the entire semantic value chain. The significance of the semantic value chain is its generalizing power combined with the fact that addition of new information in the analysis can be considered by concept type.

5.4.2 Benchmark Comparison

We performed an experiment within the TRECVID benchmark to show the effectiveness of the semantic value chain for detection of semantic concepts among 12 present-day video indexing systems. The TRECVID 2004 procedure prescribes that 10 pre-defined concepts are evaluated. Hence, we report the official benchmark results for 10 concepts in our lexicon only. The 10 benchmark concepts are, however, representative for the entire lexicon of 32. All evaluations are based on the semantic value chain.

We compare our work with the 11 other participants in TRECVID 2004. We select from each participant the system tuning with the best performance for a concept out of a maximum of 10 tunings. For ease of explanation we do not take the optimal tunings of the semantic value chain, as reported in [143], into account. Instead, we use a similar parameter setting for all concepts. Hence, we favor other systems in this comparison. Results are visualized in Fig. 5.9 for each concept.

Relative to other video indexing systems the semantic value chain performs the best for two concepts, i.e. *people walking* and *physical violence*, and second for five concepts, i.e. *boat*, *Madeleine Albright*, *Bill Clinton*, *airplane take off*, and *road*. For two concepts we perform moderate, i.e. *basket scored* and *beach*. Here the best approaches are based on specialized concept detection methods that exploit domain

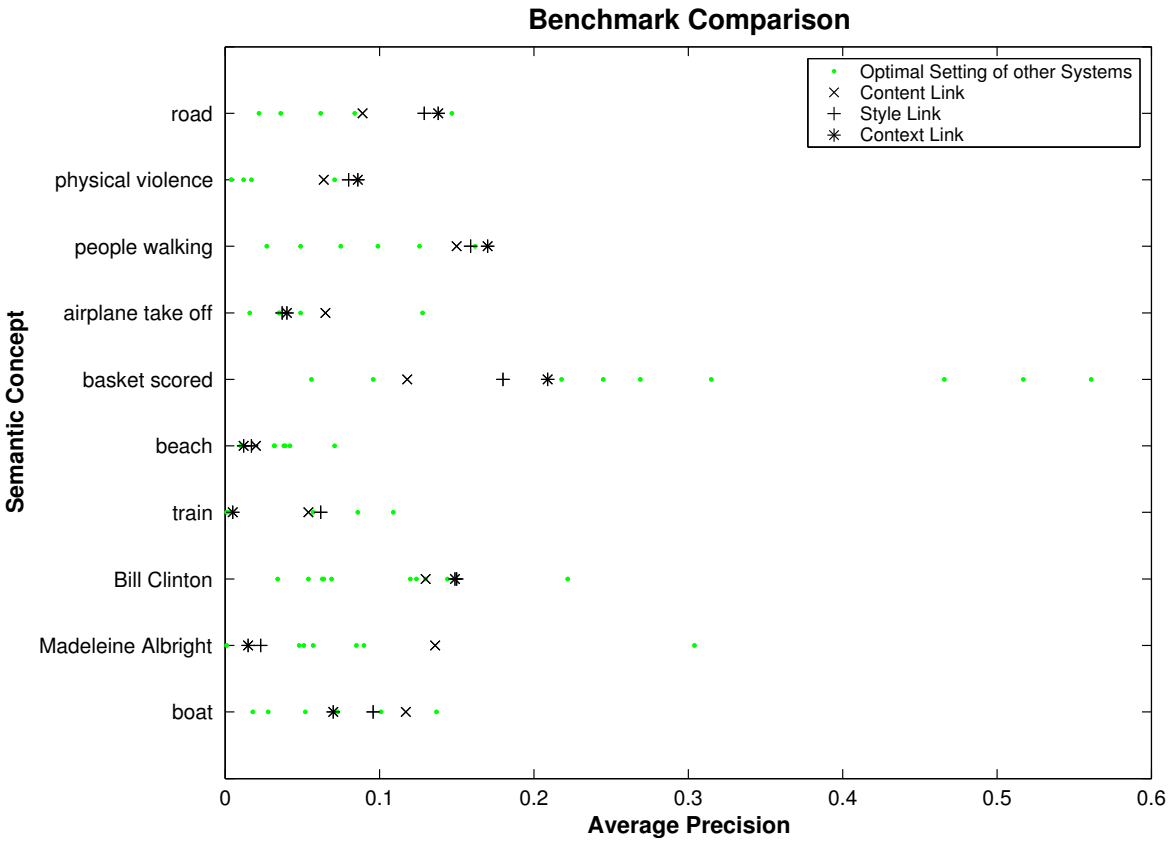


Figure 5.9: Comparison of semantic value chain results with 11 other present-day indexing systems [102].

knowledge. The big disadvantage of these methods is that they are specifically designed and implemented for one concept. They do not scale to other concepts. The benchmark results show that the semantic value chain allows for generic indexing with state-of-the-art performance.

5.4.3 Usage Scenarios

The results from the semantic value chain facilitate the development of various applications. The lexicon of 32 semantic concepts allows for querying a video archive by concept. In [147], we combined into a semantic video search engine query-by-concept, query-by-keyword, query-by-example, and interactive filtering. In addition to interactive search, the set of indexes is also applicable in a personalized retrieval setting. A feasible scenario is that users with a specific interest in sports are provided with personalized summaries when and where they need it. The sketched applications provide a semantic access to multimedia archives.

5.5 Conclusion

We propose the semantic value chain for semantic access to multimedia archives. The semantic value chain is a generic approach for video indexing. In an effort to bridge the semantic gap, it unifies the state-of-the-art in semantic video indexing into a common system architecture. The architecture is built on a variety of specialized detectors, multimodal analysis, hypothesis selection, and machine learning. The semantic value chain selects an analysis chain from the content link, the style link, and the context link. It relies on the content link only when concepts share many similarities in their multimodal content. It combines the content link with the style link when the professional habits of television making add value to the concept. Finally, it exploits a consecutive chain based on content, style, and context for concepts that require additional semantic value.

Experiments with a lexicon of 32 semantic concepts demonstrate that the semantic value chain allows for generic video indexing. In addition, the results over the various links indicate that a technique taxonomy exists for solving semantic concept detection tasks. Finally, the semantic value chain is successfully evaluated within the 2004 TRECVID benchmark. With one and the same set of system parameters two concepts, i.e. *people walking* and *physical violence*, came out best against 11 other present-day systems with average precision scores of 0.170 and 0.086 respectively. For five concepts our system scored the second best, i.e. *boat* (0.117), *Madeleine Albright* (0.136), *Bill Clinton* (0.150), *airplane take off* (0.065), and *road* (0.138). Just two performed poorly in this comparison, i.e. *basket scored* (0.209) and *beach* (0.020). The results show that the semantic value chain allows for state-of-the-art performance without the need of implementing specialized detectors. We consider this the best indicator of the approach.

A semantic value chain is as strong as its weakest link. Introduction of feature selection and knowledge representations in the various links will increase results. In its current form the context link takes the results of the style link for granted; and results are only adapted when there is enough contextual evidence from the other concepts to do so. Hence, a large set of annotated examples is an important asset. To alleviate the burden of annotation, current developments in active learning deserve more attention. In addition, we need to tackle the greatest challenge ahead: extend the lexicon of semantic concepts to a set that is competitive with human knowledge. This will have an unprecedented impact on multimedia repository usage scenarios.

For the moment, the average precision resulting from completely automatic indexing ranges from 0.020 to 0.209. What this means is that in 184 hours of standard produced video, and after training on a few hundred examples from separated data, only a small fraction of the instances in the footage are retrieved. For daily practice footage selection this may be good enough to offer already a variety of choice. These numbers are not good enough yet for high precision search but compared to the reported results of just 2 years ago [102], automated search in video archives lures at the horizon.

Chapter 6

A Lexicon-Driven Paradigm for Interactive Multimedia Retrieval

The semantic gap separates the raw multimedia data-driven features on one end from user interpretation at the other end. The gap is quite big and requires more than just text-analysis to overcome. The semantic gap dictates that only a limited lexicon of semantic concepts can be learned automatically, thus user involvement is essential. We combine in this Chapter learning of a limited lexicon of semantic concepts with similarity and interaction into a common paradigm to bridge the semantic gap. The core of the paradigm is formed by first detecting a lexicon of 32 semantic concepts. From there, we explore the combination of query-by-concept, query-by-similarity, and user interaction into an integrated video search engine. The paradigm is evaluated within the 2004 NIST TRECVID video retrieval benchmark, using a news archive of 184 hours. Benchmark results show that the lexicon-driven search paradigm is highly effective for interactive multimedia retrieval. In addition, we demonstrate that the paradigm yields top ranking performance when users have experience with the concepts in the lexicon and their anticipated performance.

6.1 Introduction

For text collections, search technology has already evolved to a mature level. Retrieval tools have found in the Internet a medium to prosper, opening new ways to do business, to do science, to be informed. All of this was realized only 15 years after the introduction of the web. The success has wet the appetite for retrieval from multimedia repositories. Now there is a problem. Multimedia archives do not release their content as easily as text does. So, in providing access to multimedia archives, current search engines [20, 52] often rely on filename and accompanying textual sources only. This approach is fruitful when a meticulous and complete description of the content is available. In many circumstances however, time and resources are missing for very detailed annotation. In addition, all-purpose completeness in describing multimedia is practically impossible. Moreover, text-only retrieval ignores the treasure of information that is available in the visual and auditory information stream. The reduction to text-only retrieval for multimedia retrieval is too simplistic.

Unfortunately, techniques for multimedia retrieval are not that effective yet in mining the semantic gold hidden in video archives. The main problem for any multimedia retrieval methodology aiming for access is the semantic gap between multimedia data representation and their interpretation by humans [136]. Where users seek high-level semantics, they are being offered low-level abstractions of the data instead. Not surprisingly, the user experience with multimedia retrieval is one of frustration. Therefore, a new paradigm of semantics is required when aiming for access to multimedia archives.

In [136], learning, similarity, and interaction are identified as key techniques to bring semantics to the user. The multimedia research community follows this track [10, 60, 119, 144, 164, 170]. However, there is no consensus on how to combine key techniques into a common paradigm. A crossroad can be observed in the current multimedia retrieval landscape, where we identify three major trails. One path follows the analogy of information retrieval. It considers multimedia retrieval as an extension of probabilistic text retrieval, e.g. [164]. The authors concentrate on the construction of generative probabilistic models for the various modalities. They conclude that combined analysis is effective only when the modalities yield reasonable retrieval scores in isolation. Another direction is followed in [170]. After text retrieval the authors boost results by combining multimodal detectors. The authors show that a query-dependent adaptation of the weights for the various detectors has a positive influence on overall multimedia retrieval performance. We consider both directions promising, but still too much depending on text retrieval. We follow the multimedia trail instead; in which text retrieval is an important signpost, but not the guiding compass. Others also explore this frontier [10, 60, 119, 144]. We advocate that the ideal multimedia retrieval system should first learn a lexicon of concepts, based on multimedia analysis, to be used for the initial search. Then, the ideal system should employ similarity and interaction to refine the search until satisfaction. The combination of learning, similarity, and interaction escapes the semantic gap.

As the semantic gap implies pre-indexing of all concepts of interest is impossible, we propose a multimedia retrieval paradigm that is build on three principles:

learning of a limited set of semantic concepts, multimedia data similarity, and user interaction. Within the proposed paradigm, we explore the combination of query-by-concept, query-by-similarity, and interactive filtering using an integrated video search engine. To demonstrate the effectiveness of our multimedia retrieval paradigm, the interactive search experiments are evaluated within the 2004 NIST TRECVID video retrieval benchmark [102].

The organization of this Chapter is as follows. First, we formulate the problem in terms of related work in Section 6.2. Our multimedia retrieval paradigm is presented in Section 6.3. We describe the experimental setup in which we evaluated our paradigm in Section 6.4. We present results in Section 6.5.

6.2 Problem Formulation and Related Work

We wish to provide users semantic access to multimedia archives. The question is how we should exploit the combination of learning, similarity, and interaction into an effective multimedia retrieval paradigm? In addition, the question arises how to measure multimedia retrieval performance. One obtains effective semantic access only if an answer is provided for both questions.

To answer the first question, we start to focus on three example methodologies that also advocate the combination of learning, similarity, and interaction for semantic access. From there we discuss the implication for the proposed multimedia retrieval paradigm.

Rautiainen *et al.* [119] proposed a semantic access method which combines three search engines into a common paradigm. A text search engine analyzes the text obtained from automatic speech recognition and allows for keyword-based retrieval. A visual search engine combines color and edge features and facilitates query-by-example. Finally, a concept search engine allows to query on a lexicon of 15 semantic concepts. The authors show that a weighted combination of search engines yields the best result for retrieval. However, as the authors indicate, overall performance was discouraging. In part, this can be explained by the inaccurate concept detectors. We adopt and generalize their thoughts on combining query-by-keyword, query-by-example, and query-by-concept for multimedia retrieval, but we add interaction in our paradigm.

Informedia [29, 60] is often found among the top performers in benchmarks like TRECVID. Their multimodal system employs a set of specialized concept detectors. It is especially strong in (interactive) search scenarios. In [29], the authors explain the success in interactive retrieval as a consequence of using storyboards, i.e. a grid of key frame results that are related to a keyword-based query. As queries for semantic concepts are hard to tackle using the textual modality only, the interface also supports filtering based on semantic concepts. The filters are based on a lexicon of 10 pre-indexed concepts with mixed performance [60]. Because of this variance in reliability the filters are applied after a keyword-based search. The disadvantage of this approach is the dependence on keywords for initial search. Because of the semantic gap, user-interaction with this restricted answer set results in limited semantic access. We

embrace the use of storyboards, querying, and filtering for interactive retrieval, but we emphasize query-by-concept in the interaction process, where possible, to limit the dependence on keywords.

A system for generic semantic indexing is proposed in [10, 96]. The system starts with feature extraction, followed by consecutive aggregations on features, multiple modalities, and concepts. Finally, the system optimizes the result by rule-based post filtering. In spite of the use of a lexicon of 64 reported concepts, interactive retrieval results with the web-driven *MARVEL* system [10] are disappointingly low compared to [29]. We attribute this observation, in part, to the lack of speed of the web-based interface. Moreover, it lacks video playback functionality. However, the largest problem is the complex query interface that offers too many possibilities and prevents users from quick retrieval of video segments of interest. We adopt and extend their ideas related to semantic indexing, but we take a different road for interactive retrieval.

To combine learning, similarity, and interaction we adopt the view of Smeulders *et al.* [136]. They define interaction as the interplay between the user, the multimedia data, and its semantic interpretation. We propose a multimedia retrieval paradigm that analyzes multimedia at both a semantic and a data level. A search engine provides users with the possibility to interact with the data. Several query interfaces allow users to retrieve results, which are visualized in the form of a storyboard. Interactive retrieval using the proposed paradigm facilitates semantic access to multimedia archives.

In response to the need for measuring multimedia retrieval performance, we note that it has always been a delicate issue. Multimedia archives are fragmented and mostly inaccessible due to copyrights and the sheer volume of data involved. As a consequence, comparison of systems has traditionally been difficult. NIST started organizing the TRECVID video retrieval benchmark to tackle the evaluation problem. The benchmark aims to promote progress in video retrieval via open, metrics-based evaluation [102]. Tasks include camera shot segmentation, story segmentation, semantic concept detection, and several search tasks. The video archive used is composed of 184 hours of ABC World News Tonight and CNN Headline News. The development data contains approximately 120 hours covering the period of January until June 1998. The test data contains the remaining 64 hours, covering the period of October until December 1998. Together with the video archive came automatic speech recognition results donated by LIMSI [47]. CLIPS-IMAG [115] provided a camera shot segmentation and corresponding key frames. The camera shots serve as the unit for retrieval. We evaluate our multimedia retrieval paradigm within the TRECVID benchmark, to demonstrate its effectiveness.

6.3 Multimedia Retrieval Paradigm

We propose a lexicon-driven paradigm to equip users with semantic access to multimedia archives. The aim is to retrieve from a multimedia archive S, which is composed of n unique shots $\{s_1, s_2, \ldots, s_n\}$, the best possible answer set in response to a user

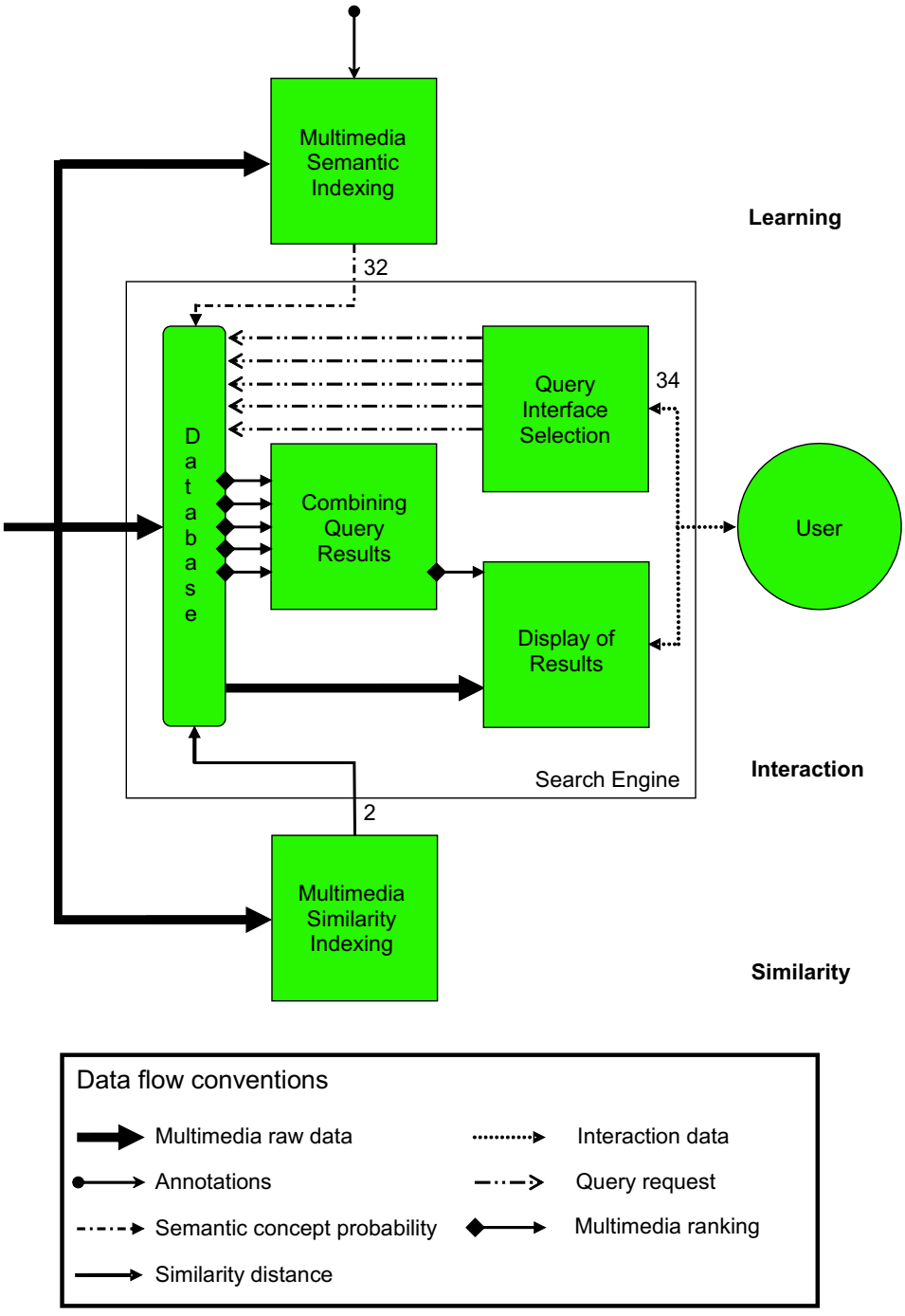


Figure 6.1: The lexicon-driven paradigm for interactive multimedia retrieval combines learning, similarity, and interaction. It learns to detect a lexicon of 32 semantic concepts. In addition, it computes 2 similarity distances. A search engine then presents 2 interfaces for query-by-similarity and 32 interfaces for query-by-concept. Based on interaction a user refines search results until an acceptable standard is reached.

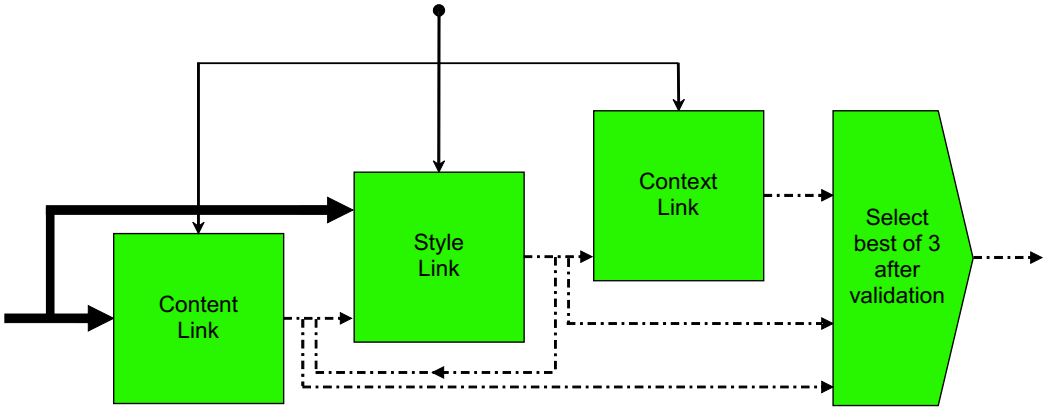


Figure 6.2: The semantic value chain for generic semantic video indexing [144], using the conventions of Fig. 6.1. Data flow for detection of one concept is visualized.

information need. To that end, the paradigm combines learning, similarity, and interaction. The paradigm exploits learning by means of a multimedia semantic indexing component, using machine learning for automatic indexing of a lexicon of 32 concepts. It associates a probability of concept presence to every shot. It stores the probabilities of all detected concepts for each shot in a database. In addition to learning, the paradigm also facilitates multimedia analysis at a similarity level. In the similarity component, 2 similarity functions are applied to index the data in the visual and textual modality. It results in 2 similarity distances for all shots, which are also stored in a database. A search engine offers users an access to the stored indexes and the video data in the form of 34 query interfaces; i.e. 2 query-by-similarity interfaces and 32 query-by-concept interfaces. The query interfaces emphasize the lexicon-driven nature of the paradigm. The search engine handles the query requests, combines the results, and displays them to an interacting user. Within the paradigm, we perceive of interaction as a combination of querying the search engine and selecting relevant results. A schematic overview of the retrieval paradigm is given in Fig. 6.1. The various components of the paradigm are now explained in more detail.

6.3.1 Multimedia Semantic Indexing

We follow our previous work [144] in the idea that the essence of produced video is its creation by an author. Style is used to stress the semantics of the message, and to guide the audience in its interpretation. In the end, video aims at an effective semantic communication. All of this taken together, the main focus of semantic indexing must be to reverse this authoring process, for which we proposed the semantic value chain.

The semantic value chain is composed of three links, see Fig. 6.2 from [144]. The output of a link in the chain forms the input for the next one. We build this architecture on machine learning of concepts for the robust detection of semantics. The semantic value chain starts in the *content link*. In this link, it follows a data-driven approach of indexing semantics. It analyzes both the visual data and textual data. In

Figure 6.3: Instances of the 32 concepts in the lexicon, as detected with the semantic value chain.

the learning phase, it applies a support vector machine to learn concept probabilities. The *style link* addresses the elements of video production, related to the style of the author, by several style-related detectors. They include shot length, frequent speakers, camera distance, faces, and motion. Again, a support vector machine classifier is applied to learn style probabilities. Finally, in the *context link*, the probabilities obtained in the style link are fused into a context vector. Then, again a support vector machine classifier is applied to learn concepts. Some concepts, like *vegetation*, have their emphasis on content thus style and context do not add much. In contrast, more complex events, like *people walking*, profit from incremental adaptation of the analysis by using concepts like *athletic game* in their context. The semantic value chain allows for generic video indexing by automatically selecting the best chain of analysis links on a per-concept basis. As indicated earlier, it is based on a lexicon of 32 semantic concepts. Instantiations of the concepts in the lexicon are visualized in Fig. 6.3. The lexicon contains both general concepts, like *building*, *boat*, and *outdoor*, as well as specific concepts such as *basket scored* and *people walking*. The semantic value chain detects all 32 concepts with varying performance [144].

6.3.2 Multimedia Similarity Indexing

In general, three data streams run in parallel in video, namely: the auditory modality, the textual modality, and the visual modality. The variety of features one can extract from the streams is enormous, see [136, 142, 162] for an overview. Once features are computed, they can be used to define a similarity function applicable within query methods like query-by-humming [50], query-by-keyword [123], and query-by-example [136]. As auditory examples are often unavailable, we provide users with possibilities for query-by-keyword and query-by-example to access a video.

To arrive at query-by-keyword we first derive words from automatic speech recognition [47]. We remove stopwords using the SMART's English stoplist [123]. We then construct a vector space by taking all transcribed words. We rely on latent semantic indexing [36] to reduce the search space to 400 dimensions. While doing so, the method takes co-occurrence of related words into account by projecting them onto the same dimension. The rationale is that this reduced space is a better representation of the search space. Users query the space reduced by latent semantic indexing using keywords that are projected to the same dimensions [167].

In the visual modality the query is by example. For all key frames in the video archive, we compute the perceptually uniform *Lab* color histogram [49] using 32 bins for each color channel. Users compare key frames with the Euclidean distance among histograms.

6.3.3 Search Engine

Video search engines are often dictated by technical possibilities rather than actual user needs [78]. Frequently this results in an overly complex search engine. To shield the user from technical complexity, we store all computed indexes in a database. Users interact with the search engine based on query interfaces. After a user issues a query it is processed and combined into a final result which is presented to the user. The elements of our search engine are now discussed in more detail.

Query Interface Selection

The basis for interactive selection of query results forms the set of 32 concepts in the lexicon. Users may rely on direct query-by-concept for concepts from this lexicon. Since the lexicon contains the concept *boat*, all information needs related to *ships* benefit from query-by-concept. This is an enormous advantage for the precision of the search. Users can also make a first selection when a query includes a super-class or a sub-class of a concept in the lexicon. For example, when searching for *vehicles* one can use the available concepts *car*, *boat*, *bicycle*, *train*, and *airplane take off* from the lexicon. Apart from querying on presence of a concept, users may also query on absence of a concept. This aids in reducing ambiguity. Consider for example a query on *sporting events* but not *ice hockey*. The lexicon of 32 concepts aids users in various ways in specifying their queries.

For search topics not covered by the concepts in the lexicon, users have to rely on similarity in the form of query-by-keyword and query-by-example. Applying query-by-keyword in isolation allows users to find very specific topics only if they are mentioned in the transcription from automatic speech recognition. Based on query-by-example, on either provided or retrieved image frames, key frames that exhibit a similar color distribution can augment results further. This is especially fruitful for repetitive key frames that contain similar visual content throughout the archive, such as previews, graphics, and commercials.

Naturally, the search engine provides users the possibility to combine query interfaces. This is helpful when a concept is too general and needs refinement. For

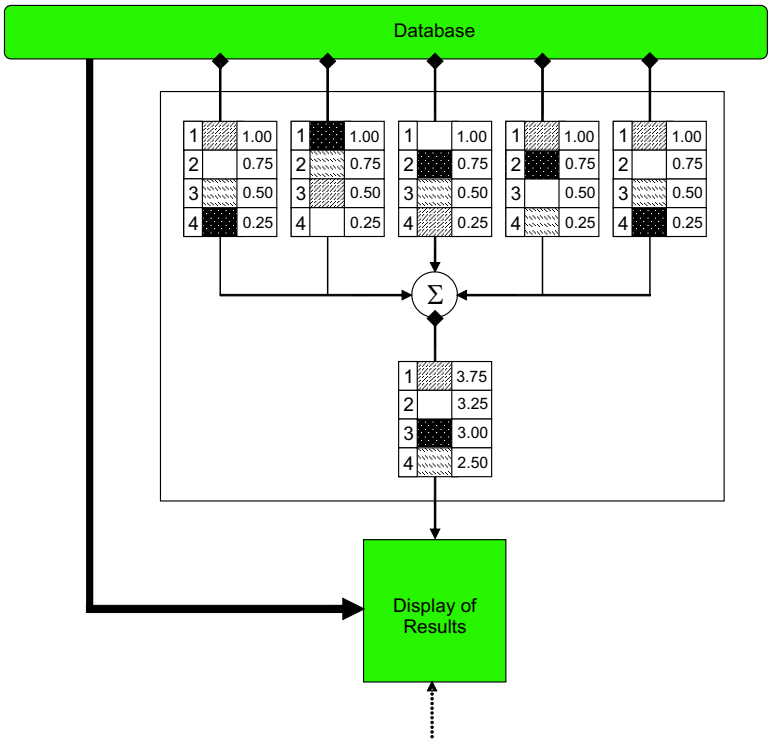


Figure 6.4: Combining query results in the video search engine is based on weighted reordering of ranked lists. In the detail from Fig. 6.1, five query requests by a user have resulted in five multimedia rankings of a data set. For simplicity the example data set contains only four shots. All ranked shots are weighted according to Eq. (6.3). The sum of all weights per shot, as defined in Eq. (6.4), is exploited to present a user a final ranked list of search results.

example when searching for Microsoft stock quotes, a user may combine query-by-concept *stock quotes* with query-by-keyword *Microsoft*. While doing so, the search engine exploits both the semantic indexes and the multimedia data indexes.

Combining Query Results

As indicated before, the search engine provides users with 34 query interfaces. Each query interface acts as a ranking operator Φ_i on the multimedia archive S, where $i \in \{1, 2, \ldots, 34\}$. The search engine stores results of each ranking operator in a ranked list ρ_i, which we denote by:

$$\rho_i = \Phi_i(S) \tag{6.1}$$

To rank S query-by-concept exploits semantic probabilities, while query-by-keyword and query-by-example use similarity distances. When users mix query interfaces, and hence several numerical scores, this introduces the question how to combine the results. Therefore, the search engine uses a decision method based on rankings to combine query results.

Various ranking combination methods exist [64]. In general, the purpose of multimedia retrieval is to obtain as many accurate results with a high ranking as possible. A ranking method based on reordering of results is therefore a good choice [64]. To reorder the ranked lists of results, we first determine the rank r_{ij} of shot s_j over the various ρ_i. Denoted by:

$$r_{ij} = \rho_i(s_j) \tag{6.2}$$

We define a weight function $w(\cdot)$ that computes the weight of s_j in ρ_i based on r_{ij}. This linear weight function gives a higher weight to shots that are retrieved in the top of ρ_i and gradually reduces to 0. This function is defined as:

$$w(r_{ij}) = \frac{n - r_{ij} + 1}{n} \tag{6.3}$$

We aggregate the results for each shot s_j by adding the contribution from each ranked list ρ_i. We then use the final ranking operator Φ^* to rank all shots from S in descending order based on this new weight. This combination method yields a final ranked list of results ρ^*, defined as:

$$\rho^* = \Phi^* \left(\left\{ \sum_{i}^{m} w(r_{ij}) \right\}_{j=1,2,\ldots,n} \right) \tag{6.4}$$

where m indicates the number of selected query interfaces. The combination procedure is visualized in Fig. 6.4.

Display of Results

We use a grid of key frames for visualization of ranked results. After inspection of this storyboard, a user selects video shots of interest and adds them to a list of relevant results. We also offer users the possibility to browse through the temporal dimension of a specific video, after selection of an active shot. If requested, playback of specific shots is also possible. To select query interfaces and combine retrieval results we rely on interaction by a user. The interface of the search engine, depicted in Fig. 6.5, allows for easy combination of query interfaces and swift visualization of retrieved results.

6.4 Experimental Setup

6.4.1 Interactive Search

We performed our experiments within the interactive search task of the 2004 TRECVID benchmark to demonstrate the significance of the proposed paradigm. The goal of the interactive search task is to satisfy a multimedia information need. Given such a need, in the form of a search topic, a user is engaged in an interactive session with a video search engine. Based on the results obtained, a user rephrases queries; aiming at retrieval of more and more accurate results. To limit the amount of user interaction

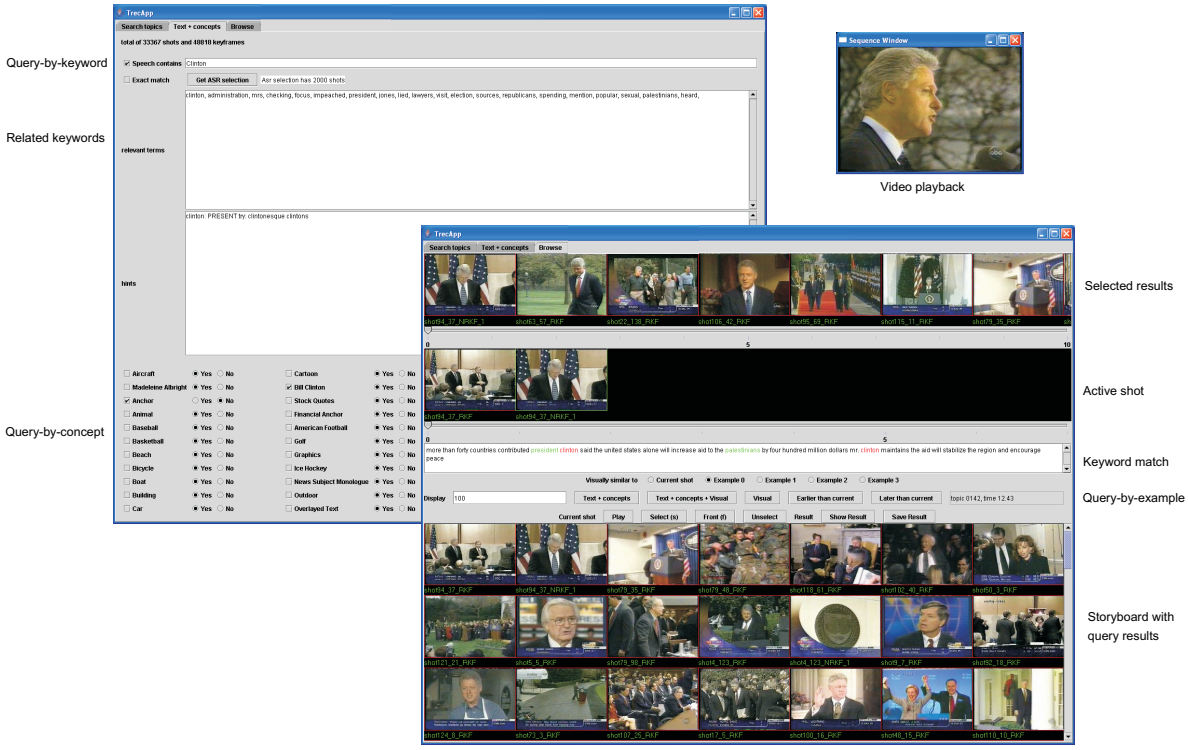


Figure 6.5: Interface of the video search engine. The system allows for interactive query-by-concept using 32 concepts. In addition, it facilitates query-by-similarity in the form of query-by-keyword, and query-by-example. Results are presented in a storyboard.

and to measure search system efficiency, all individual search topics are bounded by a 15-minute time limit. The interactive search task contains 23 search topics in total. They became known only 10 days before the deadline of submission. Hence, they were unknown at the time we developed the semantic concept detectors. We experimented with four expert users. User A had knowledge about the semantic concepts and their development set performance. The other users, B, C and D, were given a 15 minute introduction to the system only. In line with the TRECVID submission procedure, each user was allowed to submit for assessment up to a maximum of 1000 ranked results for the 23 search topics.

6.4.2 Evaluation Criteria

To determine the retrieval accuracy on individual search topics TRECVID uses *average precision*. The average precision is a single-valued measure that corresponds to the area under a recall-precision curve. This value is the average precision over all relevant judged shots. To be precise, let $L^k = \{l_1, l_2, \ldots, l_k\}$ be a ranked version of the answer set A. At any given rank k let $R \cap L^k$ be the number of relevant shots in the top k of L, where R is the total number of relevant shots. Then average precision

is defined as:

$$average\ precision = \frac{1}{R}\sum_{k=1}^{A}\frac{R \cap L^k}{k}\lambda(l_k) \tag{6.5}$$

where indicator function $\lambda(l_k) = 1$ if $l_k \in R$ and 0 otherwise. As the denominator k and the value of $\lambda(l_k)$ are dominant in determining average precision, it can be understood that this metric favors highly ranked relevant shots.

TRECVID uses a pooled ground truth P, to reduce labor-intensive manual judgments of all submitted runs. They take from each submitted run a fixed number of ranked shots, which is combined into a list of unique shots. Every submission is then evaluated based on the results of assessing this merged subset, i.e. instead of using R in (6.5), P is used, where $P \subset R$. This yields an incomplete ground truth, but a fair comparison of submissions.

As an indicator for overall search system quality TRECVID computes the mean average precision over all search topics from one run by a single user.

6.5 Results

6.5.1 Lexicon-Driven Interactive Retrieval

We plot the complete numbered list of search topics in Fig. 6.6. Together with the topics, we plot the benchmark results for 61 users with 14 present-day interactive multimedia retrieval systems. The figure includes our experiments with users A, B, C, and D respectively. Based on the results, we gain insight in the contribution of the proposed paradigm for individual search topics.

For most search topics, users of the proposed paradigm for interactive multimedia retrieval score above average. Furthermore, users of our approach obtain the highest average precision for seven search topics (Topics: 3, 14, 15, 16, 18, 20, 21). We explain the success of our interactive retrieval paradigm in part by the lexicon used. In our lexicon, there was an (accidental) overlap with the requested concepts from some search topics; for example *ice hockey*, *bicycle*, and *Bill Clinton* (Topics: 6, 16, 20), where performance is very good. Implying that there is much to be expected from a larger set of concepts in the lexicon. For other concepts, users could use available semantic concepts for filtering, e.g. *sporting event* for tennis player (Topic: 18) and *animal* for horses (Topic: 21). So in our method, abstract concepts make sense even when they are referred to indirectly. When a user finds an answer to a search topic in a repeating commercial, query-by-example is particularly useful. Search topics profiting from this observation are those related to bicycle and tennis player (Topics: 16, 18). As an exception, for search topics related to the concept *building* (Topics: 2, 22), our retrieval method performed badly compared to the best results. We explain this behavior by the fact that building was not the distinguishing concept in these topics, but rather concepts like *flood* and *fire*. When we compare results for lexicon related topics among users A, B, C, and D, it is striking that user A performs significantly better for most topics. The results imply that experience with the concepts and their

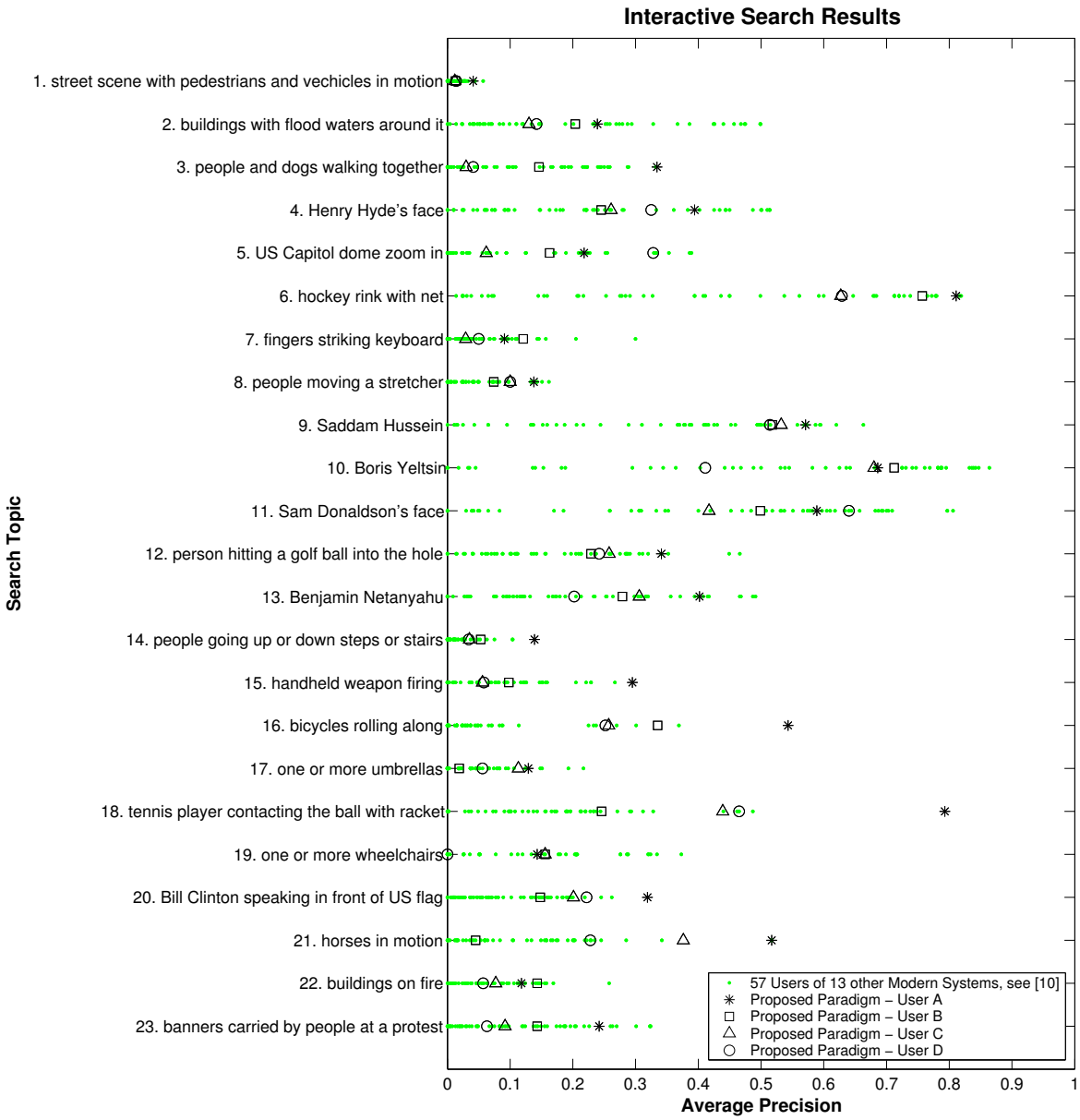


Figure 6.6: Comparison of interactive search results for 23 topics performed by 61 users of 14 present-day multimedia retrieval systems. Results for the users of the proposed paradigm are indicated with special markers.

performance pays off. This is a general observation that also holds for the other systems.

Users of the paradigm performed moderate for search topics that did not have a clear overlap with the concepts in the lexicon. Performance is for most topics however, still above average. Examples are wheelchairs (Topic: 19), umbrellas (Topic: 17), and person X related search topics that were not in the lexicon (Topics: 4, 9, 10, 11, 13). Note that the results of the various users are now much closer to each other. For

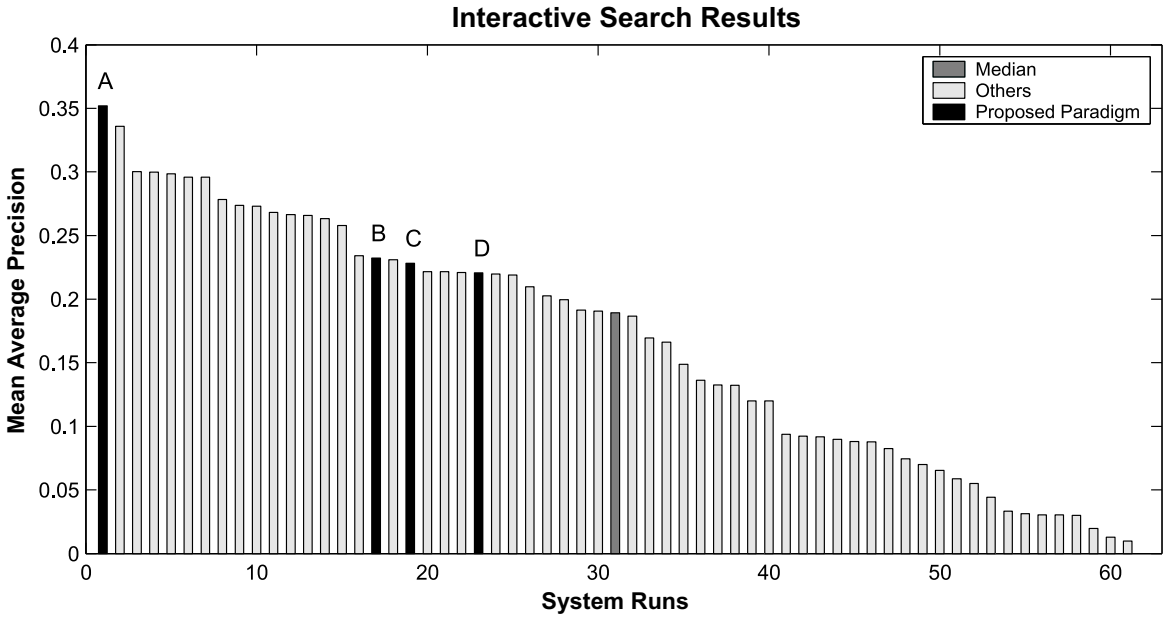


Figure 6.7: Overview of all interactive search runs submitted to TRECVID 2004, ranked according to mean average precision. Users who exploited the proposed paradigm are indicated with A, B, C, and D respectively.

these topics, we grant an important role in the obtained results to the interface of the video search engine. Because users could quickly select segments of interest, the search engine aided for search topics that users could not address with concepts from the lexicon.

6.5.2 Benchmark Comparison

To gain insight in the overall quality of our lexicon-driven interactive retrieval paradigm. We compare the results of our 4 users with 57 other users that participated in the interactive retrieval task of the 2004 TRECVID benchmark. We visualized the results for all submitted interactive search runs in Fig. 6.7.

The results show that user A of the proposed paradigm obtains the highest overall mean average precision with 0.352. This is the user with knowledge about the concepts. The other three users obtained mutually very similar results with an overall score of 0.227 on average. A score that is 20% higher than the median of all interactive search runs submitted (0.189). The benchmark results show that lexicon-driven interactive retrieval yields high accuracy; in case users have experience with the concepts and their anticipated performance, the results are state-of-the-art.

6.6 Conclusion

In this Chapter, we combine learning, similarity, and interaction into an effective paradigm for access to multimedia archives. We build the paradigm on three princi-

ples: learning of a limited set of semantic concepts, multimedia data similarity, and user interaction. The foundation of the paradigm is formed by first detecting a lexicon of 32 semantic concepts. Based on this lexicon, query-by-concept offers users a semantic entrance to multimedia repositories. In addition, users are provided with an entry in the form of similarity using query-by-keyword and query-by-example. Interaction with the various query interfaces is handled by a video search engine which provides feedback in the form of storyboard results. The paradigm combines learning, similarity, and interaction techniques to limit the influence of the semantic gap in multimedia retrieval.

Experiments with 4 users, 23 search topics, and 184 hours of news video indicate that the proposed paradigm is highly effective for interactive multimedia retrieval. For search topics that are related to concepts in the lexicon, query-by-concept is a good starting point. If a user is interested in footage that is repeated throughout the archive, query-by-example is the way to go. Query-by-keyword is effective when the (visual) content is described in the speech signal. Often, an interactive combination of query interfaces yields the best results. The results show that the lexicon of concepts aids substantially in interactive search performance. However, users need experience to value this asset on its true merits. Once users have faith in the lexicon of concepts, the paradigm reveals its true power. This is best demonstrated in a comparison among 61 users of 14 present-day retrieval systems within the interactive search task of the 2004 NIST TRECVID video retrieval benchmark. In this comparison, a user of the lexicon-driven paradigm who was experienced with the concepts and their performance on development data gained the highest overall score.

Our future work focusses on inclusion of experience into the interactive retrieval paradigm. We aim to provide interacting users of a video search engine with suggestions for query interfaces that are likely to yield the best retrieval result. We can determine the optimal query based on learned behavior on development data. Inclusion of experience, expels user frustration with multimedia retrieval technology.

At present, retrieval results with the proposed paradigm range from poor for topics like “find street scenes with pedestrians and vehicles in motion” to excellent for non-trivial topics like “find a hockey rink with net”. Fluctuating performance is unacceptable for commercial multimedia retrieval technology. However, as a support tool for professional users the proposed paradigm may already provide a treasure of semantic information.

Chapter 7

Semantic Search Engine Prototypes for Broadcast Video Archives

The traditional task of broadcast video archive owners like broadcast stations, museums, and cultural heritage institutions has been one of preservation. However, today's technology push in the form of automatic indexing tools, retrieval services, and broadband connections, requires a shift in strategy. Accessibility is the magic word for content owners. To cater for accessibility, content owners currently label broadcast video manually with semantic annotations. Due to the ever growing scale of broadcast video archives, this effort needs to be lightened by means of automatic technology. Since an archive of news videos differs not only in semantic content from an archive of soccer videos, but also in indexes and derived services the technology must be able to adapt itself to the archive of choice. In this Chapter we present such technology in the form of a general architecture for semantic video search engines. It fulfills the need for automatic indexing, components for index-derived services, and archive specific retrieval at a semantic level. To show the generality of the architecture we developed four prototype video search engines. Each prototype highlights different aspects of the general architecture. With the proposed architecture we provide broadcast video content owners with the possibility to develop tailor-made semantic video search engines.

7.1 Introduction

While broadcast television is often criticized for its lack of depth, a large number of broadcasted events have made it to the collective memory of mankind. Consider for example Neil Armstrong's first moon steps, the fall of the Berlin Wall, and more recently the devastating tsunami that hit many parts of Asia and east Africa. All of this broadcasted multimedia content, and more, is stored in numerous archives scattered all over the globe. At present, only the content owners have access to this treasure of multimedia material.

Multimedia content owners are becoming more and more aware of their valuable assets. By providing access to their archives they may open new services, expand business, and generate additional revenues. Before such scenarios become common place, however, asset owners need to realize that the value of their archives is in its index. This index requires more then a record of the broadcast date, channel, and program guide description. In Chapter 6 we demonstrated that users need semantic access to yield effective multimedia retrieval [147]. In obtaining an elaborate semantic index, multimedia content owners currently rely on manual annotation. This process is a tedious, cumbersome, error prone, and above all costly task. Hence, content owners need automatic semantic index technology to make their video archives accessible.

One of the conceivable scenarios is that content owners take responsibility for semantic indexing themselves, either by developing the technology in-house or by acquiring it externally. Once content owners are able to effectively index their assets it will result in search technology that is tailored to the broadcast video archive of choice, e.g. an archive of feature films, or ice hockey games. With corresponding services, like a compilation of scenes of a specific actor or a weekly overview of ice hockey highlights. Within this scenario we propose a general architecture for a semantic video search engine. It fulfills the need for automatic indexing, index-derived services, and archive specific retrieval at a semantic level. To show the generality of the architecture we developed four prototype semantic video search engines. Each prototype highlights different aspects of the general architecture. With the proposed architecture we provide content owners with the possibility to develop tailor-made broadcast video search engines.

The organization of this Chapter is as follows. First, we discuss related systems in Section 7.2. We then introduce in Section 7.3 the general semantic video search engine architecture. In Section 7.4 we present four prototypes, which exploit the proposed architecture. We end this Chapter with a discussion and perspective on future extensions.

7.2 Related Systems

Along the many dimensions on which we can assess video search systems we focus here on the broadcast video archive used, the granularity used, the methods exploited for indexing, the user interface, and the performance. For an in dept coverage of indexing techniques we refer to Chapter 2, for an overview of video search engine interfaces we

refer to [78].

As an extension of their powerful text-based search engine, Google recently introduced video search [52]. At present the searchable archive is limited to broadcast television shows from a restricted number of USA-based content providers. The system considers television shows as the unit of retrieval. For indexing, Google relies completely on closed caption analysis. This analysis introduces several drawbacks. First of all, there is a time lag between what is visible in the video and the transcription of the closed caption. Second, closed captions contain mistakes. Finally, it allows to retrieve semantic concepts that are mentioned in the closed caption only. Since the unit of retrieval is a television show, a query request results in a list of retrieved programs. Currently it is impossible to retrieve specific fragments from broadcast video. The web-based user interface shows a limited number of key frames from the entire program, it does not support video play back. The performance in terms of retrieval is what you would expect based on closed caption analysis.

Blinkx is another commercial video search engine [20]. Here the analysis is based solely on speech recognition. The advantage of speech recognition over closed caption analysis is the fact that the former allows for more accurate synchronization. However, due to an increase in background noise, speech recognition results may be less accurate then closed caption analysis. The web-based user interface combines low resolution Flash MX movies with a summary of the transcription of the speech that triggered the query request. After selection of clips of interest it links to the web sites of a set of restricted American and British content owners and plays the clips in a commodity video player, like Real or Windows Media. In addition to retrieval, Blinkx offers limited personalization by automatically storing references to potentially interesting footage based on predefined user preferences. Like Google, Blinkx is not evaluated on any known benchmark. However, retrieval performance appears to be better then Google.

Both speech recognition results and closed captions are combined in Dublin City University's Físchlár Digital Video System. Another strong point of Físchlár is the fact that it offers a variety of web-based user interfaces as variations of their core system architecture [135]. We focus here on Físchlár-News. This system captures and indexes television news from the Irish national broadcast station on a daily basis. In contrast to Google and Blinkx, their system adds structure to retrieval results in the form of news stories. Once users select news stories, more details are provided on a camera shot level. In addition, they can play back segments of interest using the Oracle video player. If users provide Físchlár-News with a 5-scale rating of the retrieved content, the system is able to perform personalized recommendations of future news stories. Variations of the Físchlár Digital Video System are evaluated with mixed performance within the international TRECVID benchmark, see the sidebar for TRECVID details.

Carnegie Mellon University's Informedia Project is renown for its combination of true multimedia analysis, indexing, and retrieval into a stand-alone video search system [161]. The system indexes broadcast news data from CNN on a story segment and camera shot level using a large pool of techniques, such as face detection, speech recognition, and video optical character recognition. The multimedia nature of the Informedia architecture offers its users a big advantage over architectures that are

TRECVID Benchmark

In 2001 NIST extended their successful Text Retrieval Conference (TREC) series with a track focussing on automatic segmentation, indexing, and content-based retrieval of digital video. With a steady increase in the size of the video archive analyzed, from 11 hours in 2001 up to 183 hours in 2004, and the international participants, from 12 in 2001 up to 33 in 2004, this track became an independent evaluation workshop (TRECVID) in 2003. The aim of TRECVID is to promote progress in the field of multimedia understanding by providing a large video collection, uniform evaluation procedures, and a forum for researchers interested in comparing their results. Already the benchmark is making a huge impact on the multimedia community, resulting in a large number of video retrieval systems and publications that report on the experiments performed within TRECVID. An overview of the work in several TRECVID tasks from 2001 to 2003 is covered in [31, 61, 99]. Our participation in TRECVID 2003 and 2004 is extensively reported in Chapters 4, 5, and 6.

based solely on text-based indexes. In addition, the system provides an index at a semantic level [60]. However, these are not robust enough yet for usage in isolation; they need to be combined with a keyword-based search on the transcribed speech to be effective. The system provides an extensive interface to query on all derived multimedia clues. It offers video play back via an integrated Windows Media player. The system has scored good results in the (interactive) search tasks of the TRECVID benchmark.

MARVEL [66] is the prototype system of IBM Research. It combines multimodal analysis with machine learning to index broadcast video archives at a semantic level [138]. The system serves as a test bed of IBM's TRECVID results. As a consequence, it strictly adheres to the conventions used in the TRECVID benchmark. Hence, the camera shot serves as the unit of retrieval. In addition, the broadcast video archives are those used in TRECVID. The indexed semantic concepts form the input for the web-based MARVEL interface. The search engine then allows to query video archives by semantic concepts. In addition, it offers users query-by-keyword on text obtained from speech transcription, and by visual example using a variety of image features. The indexing part of MARVEL has scored better results in the TRECVID benchmark than its interactive search part.

Summarizing the above, each system has a specific focus and accompanying architecture. We aim at the best of all worlds to arrive at a general architecture for a semantic video search engine. It allows for content owners to index various broadcast video archives, on several levels of granularity, using textual, auditory, and visual analysis. Moreover, it exploits multimodal analysis in combination with machine learning to yield an effective semantic index. Furthermore, the architecture provides content owners with flexibility to exploit a user interface that is tailor-made to the archive of choice.

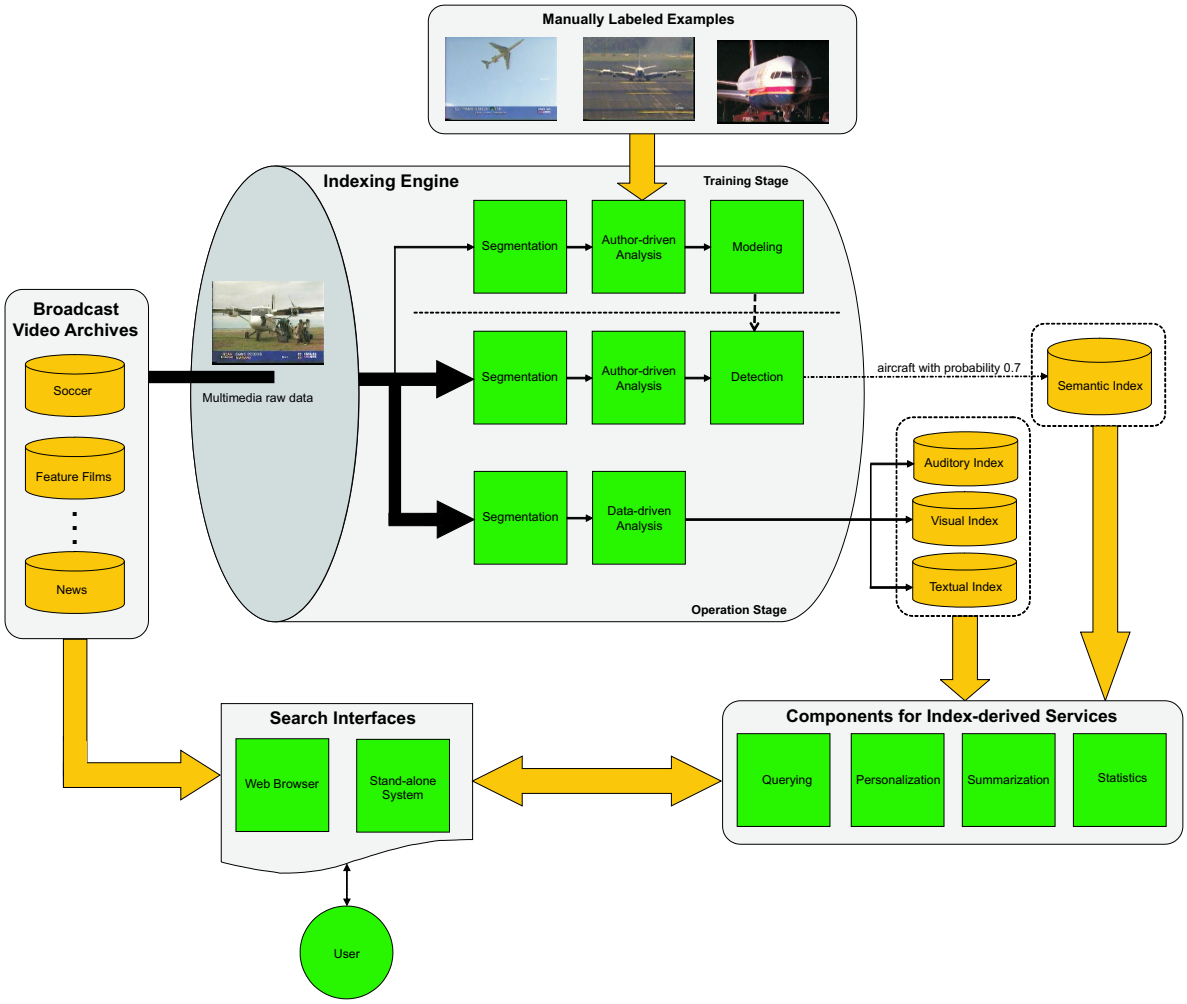


Figure 7.1: General architecture of a semantic video search engine.

7.3 Semantic Video Search Engine Architecture

In our view, a semantic video search engine consists of four basic elements. The first element is an *archive of broadcast video*. Then an *indexing engine* is required to make these assets accessible. Once an index of the broadcast video archive is available, several *components for index-derived services* can be offered to users. The fourth element is the *search interface*. It lets users interact with the video archive, the indexes, and the derived services. This is illustrated in Fig. 7.1. We will now discuss the basic elements in more detail.

The indexing engine is responsible for analysis of the raw multimedia data. As a first structuring step in the analysis we start with a segmentation of the broadcast video stream, e.g. at the granularity of camera shots or story segments. To index the segmented footage we distinguish between two types of indexes: a data index and a semantic index.

To obtain a data index, the data-driven analysis operates directly on the textual,

Authoring-driven Analysis

The central assumption in authoring-driven analysis is that any broadcast video is the result of an authoring process. An author starts with a semantic intention, an interplay of concepts and context. Based on this intention he or she produces a broadcast video document. While doing so the professional style conventions as used in broadcast television production are exploited. Ultimately this results in a digital video document that communicates the original semantic intention. When we want to extract semantics from a digital broadcast video this authoring process needs to be reversed.

For authoring-driven analysis we propose the semantic value chain, see Chapter 5. As an author uses all modalities to convey meaning, the semantic value chain starts with multimodal content analysis. First, it extracts features from the visual, textual, and auditory modality. Then the chain exploits supervised machine learning to automatically label segments with semantic concepts. In a second analysis link, the broadcast video is analyzed based on its style properties. Again using supervised machine learning for semantic labeling. Finally, semantic concepts are analyzed in context, with the potential to boost index results further. The virtue of the semantic value chain is its ability to classify semantic concepts in broadcast video in a generic fashion.

auditory, and visual stream of the video. It may use well-known techniques from the fields of natural language processing, speech recognition, and computer vision to provide a first access to the video data in the form of query-by-keyword [123], query-by-humming [50], and query-by-example [136]. The data index provides the same functionality as is currently offered by commercial video search engines.

We exploit an authoring-driven analysis to obtain an index at a semantic level. An authoring-driven analysis combines multimodal content analysis, style analysis, and context analysis, with machine learning to yield an effective semantic index, see the sidebar for details. The procedure operates in two stages: a training stage and operation stage. In the training stage a restricted set of the broadcast video material is set aside to learn various models for a lexicon of semantic concepts using machine learning. We follow a supervised learning paradigm, where we fuel the indexing engine with a set of manually annotated multimedia examples for model construction. In the operation stage, the indexing engine exploits the learned models to detect semantic concepts in previously unseen broadcast video. The semantic index lets users query broadcast video archives by concept.

After the raw multimedia is processed by the indexing engine, it stores the resulting indexes in a database. If necessary, the indexes can also be stored using an XML metadata file standard like MPEG-7 [85].

Once the data index and semantic index are stored in a database, the architecture offers several components for index-derived services that may be presented to users of a broadcast video archive. Currently these components include: querying, personalization, summarization, and statistics. The querying component depends directly on

the derived indexes. After a user issues a query, the search engine returns a ranked list of results. In case of query-by-concept these results are ranked according to the probability associated with the concept. We rely on a similarity function to rank results for the data index-derived queries. The querying component treats all users equal. Hence, queries issued by users with different interests result in the same ranked list of broadcast video footage. The personalization component offers a personal touch. This component is able to tailor query results to a user's taste. In addition, personalization based on a user profile may drive a recommender system [5]. Apart from querying and personalization, indexes may also support summarization. Finally, the indexes provide a basis for broadcast statistics. The indexes provide an effective entrance into broadcast video archives. When content owners offer the indexes on their video data over the Internet, a new set of services may arise.

Users can interact with the search engine via tailor-made search interfaces. Depending on the application the used front-end may be a web-based user interface or a stand-alone user interface.

7.4 Prototype Systems

We will now illustrate the possibilities of the general architecture by means of four prototype systems. Each system emphasizes a specific component for semantic index-derived services.

The *Goalgle* video search engine is tailored for the domain of soccer. For this specific sub-genre, a user would typically like to find game-related statistics in the form of highlight events such as goals, cards, and substitutions or search for a particular player. We digitized an archive of 12 hours of soccer games for the *Goalgle* prototype. Highlight events were indexed automatically at the camera shot level using the authoring-driven analysis described in Chapter 3. In addition, we exploit a data-driven analysis on the closed captions to facilitate query-by-keyword. The tailor-made web-based user interface of *Goalgle* is visualized in Fig. 7.2.

The web-based user interface of the *Goalgle* search engine is composed of four different panels. The query panel provides different ways of querying the system. Most interesting queries are based on finding soccer highlight events, such as goals, cards, and substitutions. One can choose to search the entire collection or search for events in a specific match. The closed captions can be searched by entering a keyword, like is used by standard text retrieval based search engines. Furthermore, one can search for video segments showing favorite players or coaches. The result panel displays a ranked list of video segments that adhere to the query entered in the query panel. Results can be ranked on probability or by their time stamp. When a result is clicked, the segment is displayed in the video panel. In the current implementation of *Goalgle* we use an integrated Windows Media player for display of the soccer video sources. When a segment is selected, a browser panel is revealed that allows to browse through the current soccer match that is displayed in the video panel. A user can jump to previous and next highlight events within this game. *Goalgle* allows users to find the highlight events they want, without the need to watch an entire soccer game.

Figure 7.2: User interface of the Goalgle soccer video search engine.

The *News RePortal* system is tailored for the domain of news. For this genre, a user would typically be interested in structured summaries that provide a quick overview of the news archive. We digitized an archive of 32 hours of Dutch broadcast news for the *News RePortal* prototype. We segmented the videos on both the topic level and the camera shot level. Topic segmentation is the method reported in [117]. Besides segmentation it also indexes segments with the most likely topic. Apart form a topic index, we also index camera shots with news-specific structuring events like news anchor, interview, and weather news using the authoring-driven analysis as reported in Chapter 3. In addition, we exploit a data-driven analysis on the closed captions to facilitate query-by-keyword. The tailor-made web-based user interface of *News RePortal* is visualized in Fig. 7.3.

The web-based user interface of the *News RePortal* system is similar to the *Goalgle* search engine. It also consists of four panels. The query panel allows to query the news video archive on news topic, e.g. Israel, or speed skating. In addition it provides query-by-concept using semantic concepts like interviews, anchors, and weather news. Like *Goalgle*, it also offers query-by-keyword on the closed caption and retrieval of specific persons. Naturally the query panel allows to combine query interfaces, in this way users can retrieve very specific information nuggets such as interviews with Shimon

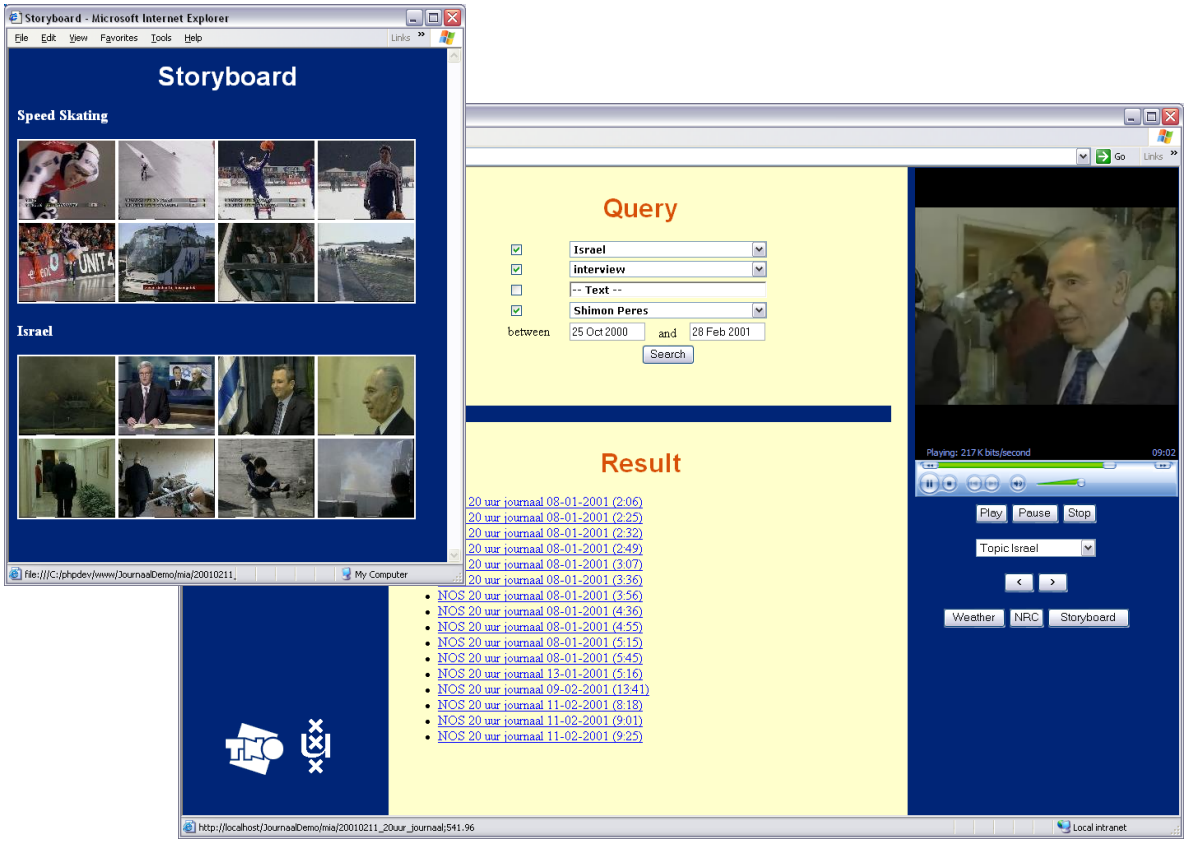


Figure 7.3: User interface of the News RePortal system, with a structured summary in the form of a storyboard.

Peres on the topic Israel. The result panel displays a ranked list of video segments that adhere to the query entered in the query panel. When a result is clicked, the segment is displayed in the video panel. We use an integrated Windows Media player for display of the news video sources. When a segment is selected, a browser panel is revealed that allows to browse through the current news episode on a per-topic basis. A strong property of the *News RePortal* system is its ability to provide users with a structured summary of the news episode in the form of a storyboard. To give a condensed but accurate summary of the various topics, we exploit an intelligent key frame selection mechanism. It filters out detected anchor shots and tries to minimize the similarity of the various selected frames. The storyboard can be exploited to jump to the various topics within this news broadcast.

The Video Personalizer (*Viper*) system is also tailored for the domain of news. In contrast to the *News RePortal*, which focuses on summarization, this prototype emphasizes personalization aspects. The news archive contains 64 hours of ABC and CNN broadcasts. All videos are segmented on the camera shot level. They are indexed using the authoring-driven analysis approach described in Chapter 5. A lexicon containing 32 semantic concepts was used for indexing. Given this lexicon, we aim to provide a person with the most relevant information given the expected use

Figure 7.4: User interface of the Viper search engine. The left panel shows the search interface, the top panel shows a spiral visualization of a search on *graphics* and the bottom panel shows a grid visualization of *ice hockey*.

and user preferences. The tailor-made web-based user interface of *Viper* is visualized in Fig. 7.4.

The web-based user interface of the *Viper* search engine is composed of six tabs. Apart from login and logout functionality *Viper* allows users to add their preferences to the system in terms of the concepts. In addition to this explicit user profiling, it also learns from user interaction. An ontology based on WordNet [42] was developed that provides a structure on the lexicon of 32 concepts and increases search possibilities by means of a category browser, i.e. it maps concepts from the query to available concepts in the lexicon if possible. We also provide a text-box based search interface for advanced users with knowledge about the available concepts. To decide whether concepts are related to a query, a pruner module computes query- and user-dependent thresholds. To present the results, *Viper* provides five visualization modes based on key frames. Four visualizations are variations of the well-known grid representation, while another uses a spiral-based representation to visualize results. Users can choose

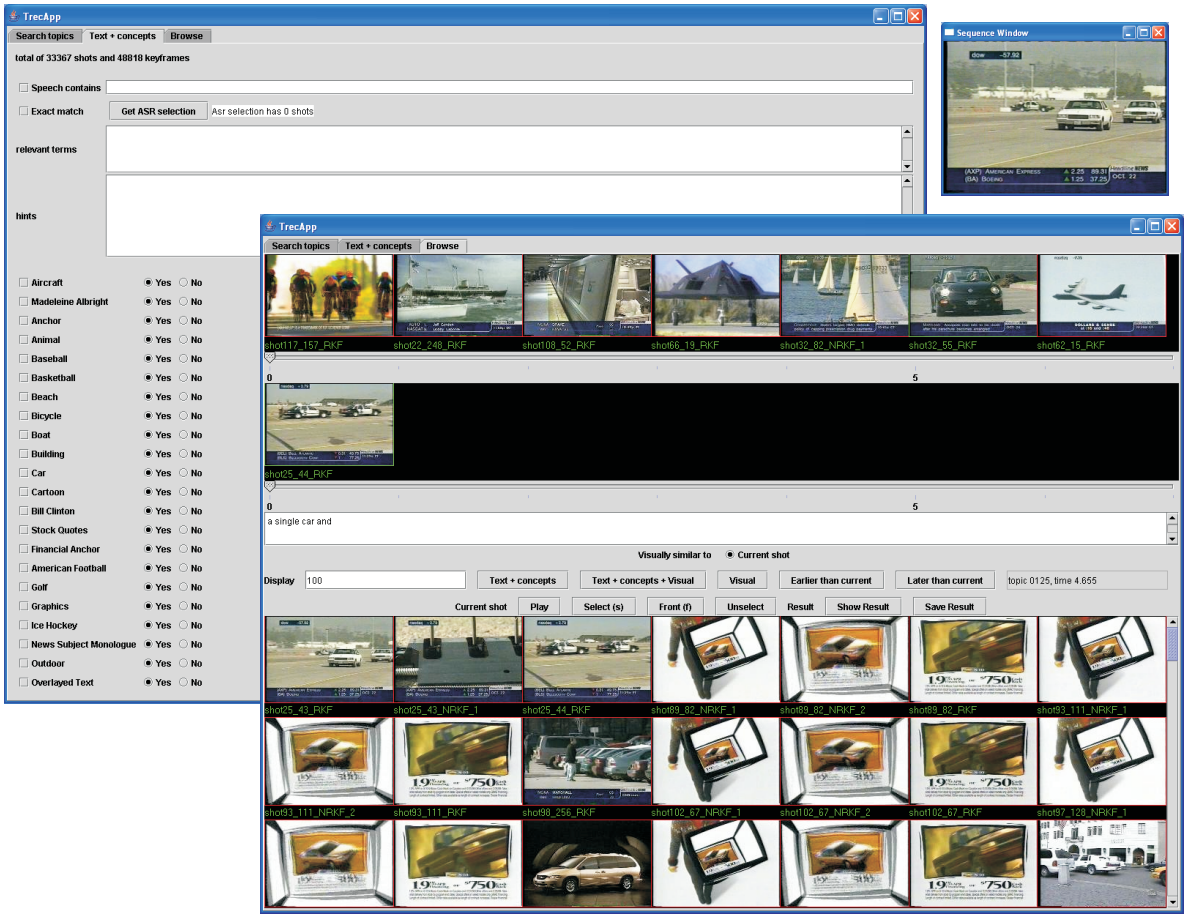


Figure 7.5: User interface of the MediaMill system. The top row of the right panel shows selected results for *vehicle*, the bottom shows results for *car*.

whatever interface they like best. After key frame selection the videos are displayed with an integrated Flash MX player. *Viper* provides different users with different query results based on the learned user preferences.

The fourth and final prototype is the *MediaMill* system. Like *Viper* this search engine exploits an archive of American broadcast news video. In contrast to the previous prototypes, this system focuses on querying. Again all videos are segmented on the camera shot level and indexed using the authoring-driven analysis approach described in Chapter 5. The lexicon of 32 semantic concepts is used for query-by-concept. Apart from the semantic index this search engine also provides a textual and visual index in the form of two similarity functions. The textual index uses Latent Semantic Indexing to allow for query-by-keyword on the speech transcript, while the visual index uses a *Lab* color histogram to facilitate query-by-example, see also Chapter 6. The stand-alone interface of the search engine is depicted in Fig. 7.5.

In contrast to the other prototypes, the *MediaMill* system uses a stand-alone interface. Where for display of the videos DirectShow is used. The reason for these choices are speed. The *MediaMill* system is developed for professional users who

swiftly need to find their information. The professional users of the system typically engage in an interactive session to retrieve the results they require. The system offers three basic query interfaces to a video archive: query-by-concept, query-by-keyword, and query-by-example. The set of concepts from the concept lexicon forms the basis for query-by-concept. For search topics not covered by concepts in the lexicon, users have to rely on a combination of query-by-keyword and query-by-example. Applying query-by-keyword in isolation allows users to find specific topics, but only when they are mentioned in the speech signal. Based on query-by-example shots that exhibit a similar color distribution can augment results further. As indicated in Chapter 6, this retrieval approach results in highly accurate semantic access to video archives.

7.5 Future Work

We present in this Chapter a general architecture for semantic video search engines, with the aim to provide content owners with tailor made video retrieval technology. It fulfills the need for automatic indexing, components for index-derived services, and archive specific retrieval at a semantic level. Based on four prototype systems we highlight different aspects of the general architecture. *Goalgle* focuses on statistics in the form of detected highlight events in soccer archives. The *News RePortal* demonstrates the summarizing capabilities of the general architecture on an archive of Dutch broadcast news. Our third prototype, *Viper*, offers users personalized delivery of American broadcast news fragments based on learned profiles. Finally, the *MediaMill* system is optimized for quickly querying broadcast news video archives. The prototypes demonstrate that the proposed search engine architecture provides application-dependent and tailor-made semantic accessibility to various broadcast video archives.

Despite the proven applicability of the general architecture, it still needs to be extended in several ways before it can offer content owners a competitive advantage over solutions by contemporary web search providers. The following elements of the general architecture need enhancements:

- *The indexing engine needs to be improved.*

  The semantic concepts form the basis for effective access to broadcast video archives. Therefore, they need to be detected with high accuracy. In addition to improved performance, the number of detectable concepts that provide the semantic index needs to be extended. Once a large lexicon of reliably detected concepts is available, more elaborate components for index derived services come within reach.

- *The personalization aspects deserve more attention.*

  We opine that from the four components for index-derived services, personalization is the most interesting. Personalization requires dedicated user profiling. More research is needed on user profiling for video search engines. When adequate profiling techniques for broadcast video search engines are available, these search engines will evolve from reactive to proactive systems, e.g. alerting users when potentially interesting footage is encountered.

- *The search interface needs extension.*

  Currently the search interfaces are limited to web-based and stand-alone front-ends. Given the possibilities and popularity of mobile devices, a user interface for mobile devices should be included in the architecture as well. Then users can access video archives when and where they want.

With these improvements the general architecture may increase access to broadcast video archives. Our future work aims for the development of an integrated semantic video search engine that offers a large set of reliable index-derived services, via several search interfaces. With the ultimate aim for content owners and content users to mine the treasure of multimedia material to their personal need.

Chapter 8

Conclusion

8.1 Summary of Contribution

This thesis makes a contribution to the field of multimedia understanding. Where our ultimate aim is to structure the digital multimedia chaos by bridging the semantic gap between computable data features on one end and the semantic interpretation of the data by a user on the other end. We distinguish between produced and non-produced multimedia or video documents. We depart from the view that a produced video document is the result of an authoring-driven production process. This authoring process serves as a metaphor for machine-driven understanding. We present a step-by-step extrapolation of this authoring metaphor for automatic multimedia understanding, see Fig. 8.1. While doing so, we cover in this thesis an extensive overview of the field, a theoretical foundation for authoring-driven multimedia understanding, state-of-the-art benchmark validation, and practical semantic video retrieval applications. Furthermore, it allows us to answer the four questions raised in Chapter 1. The authoring-driven methodology for semantic multimedia indexing is the main contribution of this thesis.

In Chapter 2 we lay the foundation for the authoring metaphor to machine understanding of multimedia. We propose a multimodal framework in which we view a video document from the perspective of its author. Within the framework we consider layout, content, and the semantic index as the significant components. Viewing a video document as the result of an authoring process, allows for seamless integration of the visual, auditory, and textual modality. In addition, the framework forms the guiding principle for identifying index types, for which automatic methods are found in literature. It unifies and categorizes these different methods. Thus it serves as a blueprint for a generic and flexible semantic video indexing system based on multimodal analysis.

The usage of multiple modalities for semantic indexing poses problems with respect to synchronization and inclusion of temporal context clues. To tackle this integration

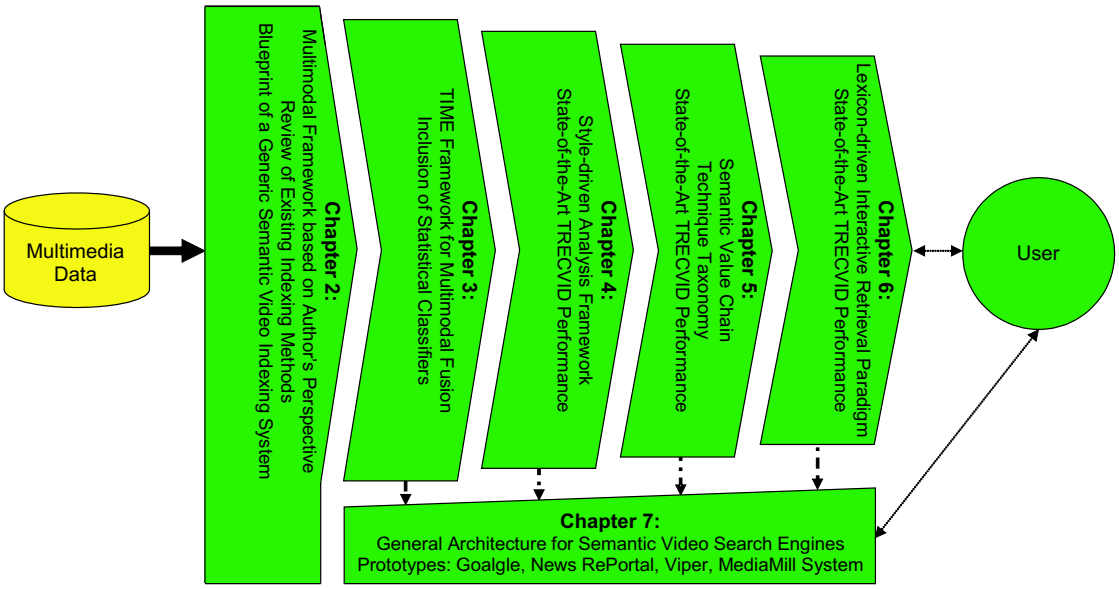


Figure 8.1: Summary of the work addressed in this thesis, including main contributions and achievements.

problem, we propose the Time Interval Multimedia Event (TIME) framework in Chapter 3. The framework explicitly handles context and synchronization. In addition, as the framework is based on statistics, it yields a robust approach for multimodal integration. We focus on the problem of combining authoring elements, in the form of content and layout segmentations, into a common analysis framework. We propose to model layout and content segmentations as time intervals to overcome the limitations of existing multimodal fusion approaches. This time interval representation allows for proper inclusion of temporal context and synchronization. Furthermore, we show that a number of statistical classifiers is applicable to the problem of semantic video indexing based on the time interval pattern representation. To demonstrate the effectiveness of TIME it was evaluated on two domains, namely soccer and news. The former was chosen because of its dependence on context. The latter because of its dependence on synchronization. We have compared three statistical classifiers, with varying complexity, and show that there exists a clear relation between narrowness of the semantic gap and the complexity of the classifier needed. Moreover, we show that the TIME framework, including synchronization and context, significantly outperforms the 'standard' multimodal analysis approaches common in video indexing literature.

Once we are able to properly fuse multimodal information sources, we are ready to add the notion of style to the repertoire of multimedia understanding techniques in Chapter 4. In addition to layout and content, we identify capture and context as important aspects of the authors style. We propose a generic and flexible framework for produced video indexing that is capable to learn rich semantic concepts from multimodal sources based on style analysis, where rich semantics means that the author exploits style in many ways. The framework allows for robust classification of several

rich semantic concepts in produced video by using a fixed core of layout, content, and capture detectors together with varying context detectors that are combined into a statistical classifier ensemble. Results on 120 hours of video data from the 2003 TRECVID benchmark show that it is the combination of style elements that yields the best results for produced video indexing. In addition, we demonstrate that the accuracy of the proposed framework for classification of several rich semantic concepts in broadcast news is state-of-the-art.

In Chapter 5 we elaborate further on the authoring-driven analysis methodology. We propose a generic approach for semantic indexing, based on the authoring metaphor, which we call the semantic value chain. To bridge the semantic gap, it unifies our work addressed in Chapters 2, 3, and 4, and recent advances in the field of multimedia understanding, into a common system architecture. The architecture is built on several specialized detectors, multimodal analysis, hypothesis selection, and machine learning. Furthermore, it covers the notions of content, style, and context. The semantic value chain extracts semantic concepts from video documents based on three consecutive analysis links, named the content link, the style link, and the context link. We learn an optimal configuration of analysis links, on a per-concept basis, to arrive at a technique taxonomy for semantic concept detectors. Experiments with a lexicon of 32 concepts demonstrate that the semantic value chain allows for generic video indexing. In addition, the semantic value chain is successfully evaluated within the 2004 TRECVID benchmark as top performer for the semantic concept detection task. The results show that the semantic value chain allows for generic indexing with state-of-the-art performance.

The semantic gap dictates that only a limited lexicon of semantic concepts can be learned automatically, thus eventually user involvement is essential. Therefore, we focus on interactive multimedia retrieval in Chapter 6. We propose a lexicon-driven retrieval paradigm for access to multimedia archives. The foundation of the paradigm is formed by the lexicon of 32 semantic concepts, as detected in Chapter 5. Based on this lexicon, query-by-concept offers users a semantic entrance to multimedia repositories. In addition, users are provided with an entry in the form of similarity using textual and visual examples. Interaction with the various query interfaces is handled by a video search engine which provides feedback in the form of storyboard results. The lexicon-driven paradigm combines learning, similarity, and interaction techniques to bridge the semantic gap in multimedia retrieval. The paradigm is evaluated within the interactive search task of the 2004 TRECVID video retrieval benchmark, using a news archive of 184 hours. Experiments show that the lexicon-driven search paradigm is highly effective for interactive multimedia retrieval. In addition, we demonstrate that the paradigm yields top ranking performance when users have experience with the concepts in the lexicon and their anticipated performance.

The technology developed in Chapters 3, 4, 5, and 6 naturally leads to the instantiation of a semantic video search engine. In Chapter 7 we present a general architecture for such a search engine, consisting of a broadcast video archive, an indexing engine, components for index-derived services, and a search interface. We highlight several aspects of the common architecture by means of four prototype systems, i.e. *Goalgle*, *News RePortal*, *Viper*, and the *MediaMill* system.

8.2 Directions for Future Research

The path for future research continues the step-by-step extrapolation of the authoring metaphor. We envision our future research efforts as a three-stage rocket that is fueled by an ontology, as illustrated in Fig. 8.2. The first stage focuses on strengthening the semantic value chain, leading us from multimedia data to a large lexicon of semantic concepts. When the limits of the semantic value chain are reached, we need to explore how concepts cluster in conceptual space using concept similarities. Finally, in the third stage we need to focus on enhancement of the interactive retrieval paradigm. Let us now make the three stages for future research more precise.

The semantic value chain can be strengthened in several ways. A first direct improvement of the semantic value chain is achieved by inclusion of better visual, auditory, and textual content features. Secondly, advancement of detectors that are capable to analyze more style elements will have its repercussions on semantic value chain performance. Thirdly, a better representation of context, using an ontology, has the potential to extend machine understanding of multimedia further. All of these advancements should be integrated with developments in machine learning.

We foresee that a semantic value chain is able to detect a large lexicon of several hundreds of concepts. Eventually, however, it will reach a boundary. Ideally, an ontology is able to deduce new concepts beyond the ones detected by a semantic value chain. Then the value of the large lexicon can be enriched further after the concepts are interconnected in an ontology. How the ontology should be engineered is an important research question, which recently drifted in a first direction [59]; taking the view that included concepts should be useful from both a visual information perspective and feasible in terms of semi-automatic detection. In the long run, the ontology should extend the lexicon of detectable semantic concepts in multimedia archives to a size that is competitive with human knowledge.

Tools that build on the ontology and the detected concepts should be able to handle uncertainty, as automatic concept detection in multimedia is never perfect. In this respect, clustering of video snippets in conceptual space may provide the answer. We need to explore how to exploit similarity based on the detected concepts in combination with its performance on training data. Potentially, this conceptual similarity indexing may increase our insight in multimedia understanding.

An ontology of semantic concepts will have a dazzling impact on interactive retrieval also. Questions that need to be answered when an ontology is available are related to the query interface. How to present the available ontology to a user? In addition, how should the conceptual similarity space be visualized to a user of multimedia archives? Apart from improved query-by-concept, there is also a need for better data-driven query-by-similarity functionality. In the visual channel for example, query-by-example needs to be extended from global image examples to regional image examples. As the suggested enhancements for both query-by-concept and query-by-similarity lead to a drastic increase in the number of possible query interfaces, we anticipate that the largest innovation in interactive retrieval will come from inclusion of expert user experience. One way to achieve this is by exploiting the concept detection performance on training data in the video search engine. Search engines can

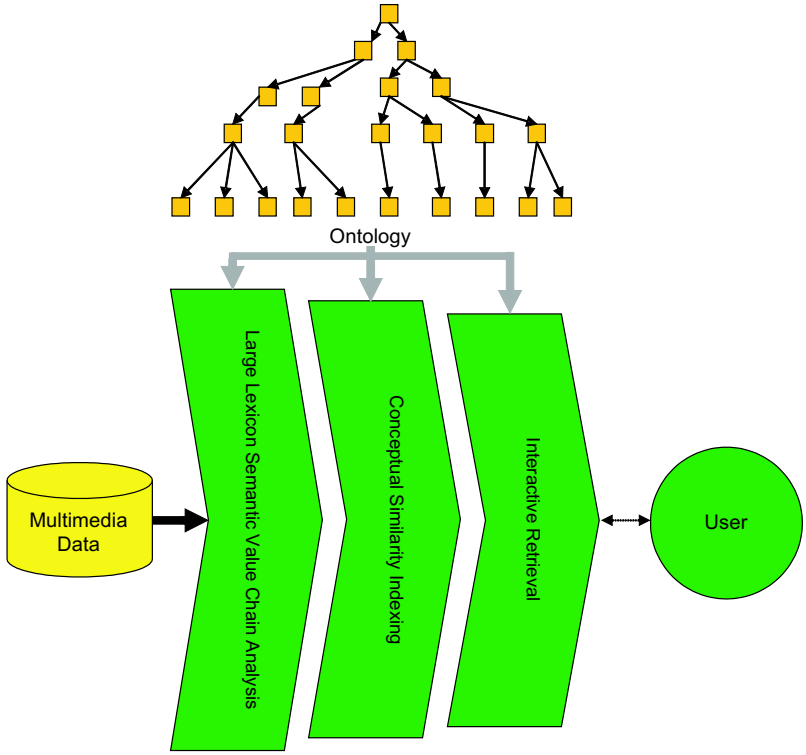


Figure 8.2: We envision our future work on machine understanding of multimedia as a three-stage rocket that is fueled by an ontology. It focuses on strengthening the semantic value chain, inclusion of conceptual similarity indexing, and enhancing interactive retrieval.

make query interface suggestions based on this learned performance. When linked to personal profiles the recommendations can be tailor-made. This will form the basis for a powerful video search engine.

8.3 General Conclusion

At the end of this thesis we are ready to ask ourselves whether we have succeeded in answering the fundamental question, i.e. *how to bridge the semantic gap for produced video?* The step-by-step extrapolation of the authoring metaphor provides us with an effective solution path.

Because different media streams yield different clues for recognition of semantics, methods aiming to bridge the semantic gap for produced video are most likely to succeed when they are multimodal instead of unimodal. Hence, for multimodal approaches the content features from the visual, auditory, and textual channels should be integrated as early as possible to yield a truly multimedia feature representation.

Such early fusion schemes require synchronization of multimedia data. To that end, its advisable to model the features as time intervals. This has the added advantage that temporal context clues can be included easily. Once multimodal features are

captured into an early fusion representation, semantic concepts should be learned by a statistical method instead of a knowledge-based approach to assure high robustness. At present, the preferred mode of operation for statistical pattern recognition from early fused multimedia representations is supervised learning with a support vector machine.

To detect concepts that have high variability in content, and high consistency in style, the early fusion representation should be based on style detectors, i.e. layout, content, capture, and context detectors. One achieves optimal results when a fixed core of style detectors is iteratively updated with more context, in the form of detected concepts. A classifier ensemble effectively combines the pool of (weak) style detectors, essentially boosting the performance to new heights.

To arrive at a generic, opposed to ad hoc, approach for semantic concept detection in produced video one should not focus on a single analysis method, i.e. content-based or style-based only. Instead, one should automatically select the best of multiple analysis combinations preferably on content, style, and context level. It increases the effort needed in the analysis substantially, but it pays off.

A generic semantic indexing approach yields a lexicon of concepts. However, automatic learning is not sufficient to cover the semantic gap completely. User interaction is required. When combined with similarity, a powerful retrieval paradigm emerges. It paves the road for video search engines offering novel index-derived services.

Methods that aim to bridge the semantic gap should demonstrate performance on publicly available data sets, using well-known evaluation protocols. When researchers join international benchmark evaluations such as TRECVID, both data sets and common performance metrics are easily obtainable. In return, it offers the multimedia understanding community insight in the approach, while at the same time promoting progress for all.

We followed the above solution path in our aim to bridge the semantic gap for produced video. For automatic analysis it resulted in the semantic value chain in Chapter 5. We reach a tentative endpoint in our endeavor to machine understanding of multimedia when we combine the semantic value chain with the paradigm for interactive multimedia retrieval, proposed in Chapter 6. To bridge the semantic gap, *a combination of automatic authoring-driven analysis and user interaction yields the most effective approach.*

However, the results should not be seen through rose-tinted spectacles. Our methods have state-of-the-art TRECVID benchmark performance, but the accuracy is still far from perfect. Despite the yearly progress in the TRECVID program, the decisive leap forward to cover the semantic gap is still to come. As a side step, we note that the authoring metaphor for multimedia understanding is evaluated on two domains, i.e. soccer and news, only.

To conclude, with the authoring metaphor we have advanced the field of multimedia understanding with an effective methodology to narrow the semantic gap substantially. We are confident that an extended exploration along the trail we blazed, in the form of future research, will structure the ubiquitous multimedia chaos further.

Appendix A

Style Detectors

In Chapters 4 and 5 we introduced and exploited the notion of style in multimedia analysis. In this Appendix we discuss the implementation of the various style detectors. Each style detector uses an existing software implementation as a basis. The output of a base detector is then aggregated and synchronized to a camera shot. Based on the features computed for the entire shot, we categorize the style features for each shot. Together these three components define a style detector. All style detectors follow the basic architecture as visualized in Fig. A.1.

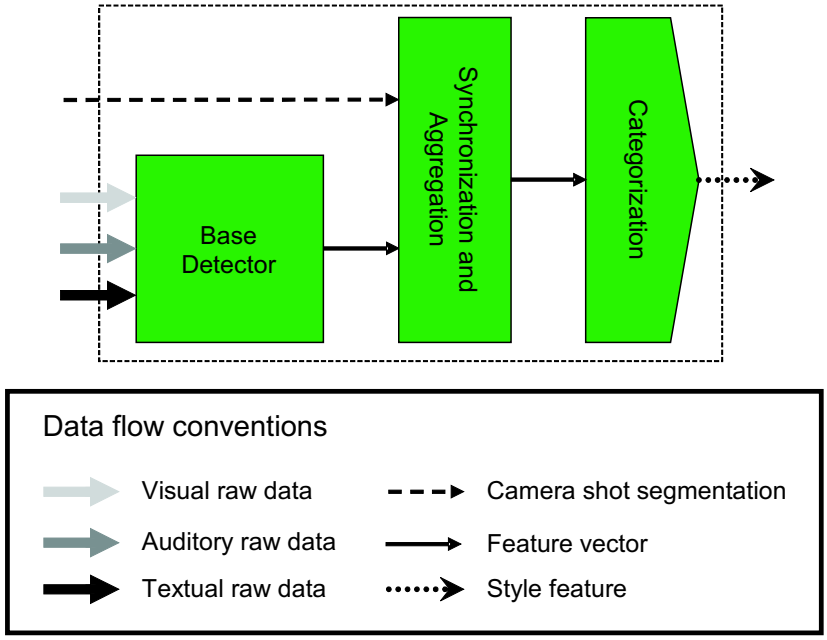


Figure A.1: Basic architecture and data flow within a style detector.

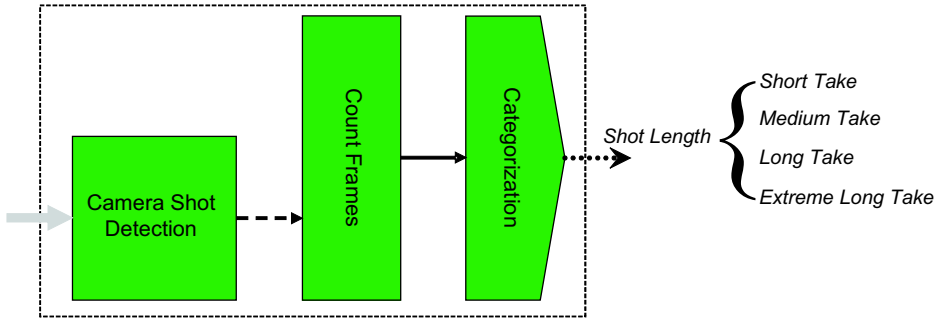


Figure A.2: Shot length detector, using the conventions of Fig. A.1.

A.1 Layout Detectors

Shot Length

An author uses variation in shot length to affect the overall rhythm of a produced video. We determine shot length based on a camera shot segmentation obtained from a camera shot detector [115]. For each shot the number of frames defines the shot length. We categorize the shot length as *short take* if a shot contains less than 70 frames. We categorize it as *medium take* if it contains 70 to 300 frames. A shot is categorized as *long take* if it contains 300 to 600 frames. In all other cases it is classified as an *extreme long take*. Note that the thresholds are chosen based on a frame rate of 29.97 frames per second. The shot length detector scheme is visualized in Fig. A.2.

Overlayed Text

Overlayed text is added by the author at production time to provide the viewer with additional descriptive information, e.g. annotation of people in broadcast news. Its presence is an important indicator for layout style. We apply a video optical character recognition system [125] to localize and extract overlayed text in a video frame. As

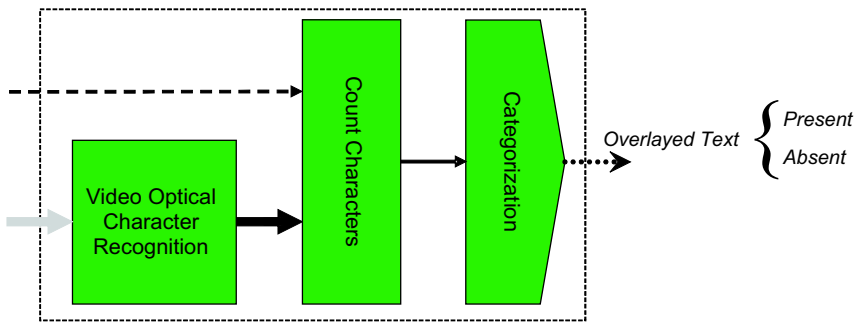


Figure A.3: Overlayed text detector.

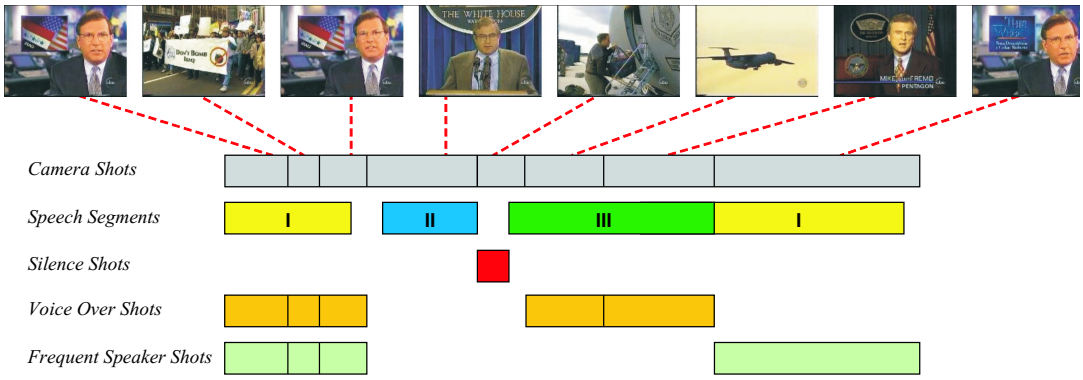


Figure A.4: Segmentation of a news video document into camera shots, speech segments with speaker identifier, silence shots, voice over camera shots, and frequent speaker camera shots.

the output of such a system may contain errors, localization of a text region is not sufficient. To increase robustness the system recognizes the text in the localized regions. Then, we count the number of characters in recognized text strings. We assume overlayed text is *present* in a shot only if one frame within the shot contains a string of at least 5 characters, else we consider it *absent*. The overlayed text detector scheme is visualized in Fig. A.3.

Silence

An author uses silence to mark transitions in the auditory layout. We detect non-speech, or silence, based on automatic speech recognition results [47]. We first count the time (in frames) between transcribed words. We consider a segment a silence if the time difference between successive words exceeds 70 frames. This results for each video in a silence segmentation. We need to combine the silence segmentation with a camera shot segmentation to obtain a decision at camera shot level. For this purpose we exploit the TIME relations proposed in Chapter 3. We ignore the *NoRelation*, *precedes* and *precedes_i* relations, as these are interesting for temporal context only.

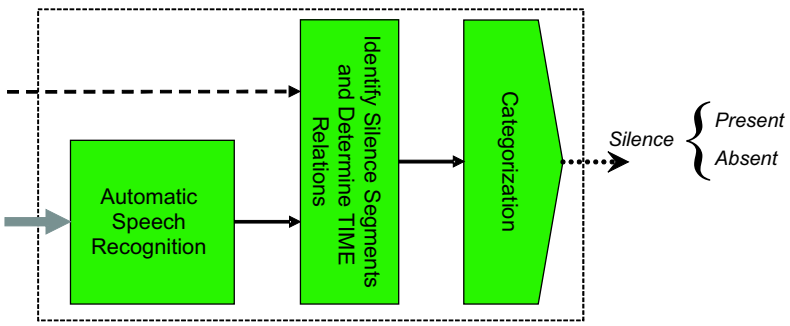


Figure A.5: Silence detector.

The shot segmentation is considered the reference segmentation. We use a value of 10 frames for the margin T_1. If a TIME relation between the shot segmentation and the silence segmentation exists, we compute the number of frames the two segmentations have in common. If this intersection exceeds 40% of the total number of frames in a reference shot we consider a silence period *present* in the shot, else *absent*, see Fig. A.4 for an example. The silence detector scheme is visualized in Fig. A.5.

Voice Over

An author uses a voice over when the content of the video is not self-descriptive and requires additional information, e.g. in sport broadcasts or documentaries. Voice over detection is also based on the automatic speech recognition results from [47]. We compare the speech segmentation with the shot segmentation. First, we count the number of cuts in the corresponding time interval of the camera shot segmentation Note that to account for imperfect segmentation, a margin of 25 frames is extracted from each end of a speech segment before counting cuts. We consider a speech segment a voice over segment when it contains more than 1 cut, this is illustrated in Fig. A.4. To map the voice over segments to camera shots we use the same TIME relations as above. But, for T_1 we now use a value of 25 frames. If a TIME relation between a camera shot and a voice over segment exists we consider a voice over *present* in the shot, else *absent*. The voice over detector scheme is visualized in Fig. A.6.

A.2 Content Detectors

Faces

Human beings are a prominent content element in produced video. To detect presence of people we apply the face detector of [130]. For each analyzed frame in a camera shot we count the number of faces present. We consider multiple faces *present* in the shot if at least two faces are detected simultaneously in 20% of the frames, else *absent*. The faces detector scheme is visualized in Fig. A.7.

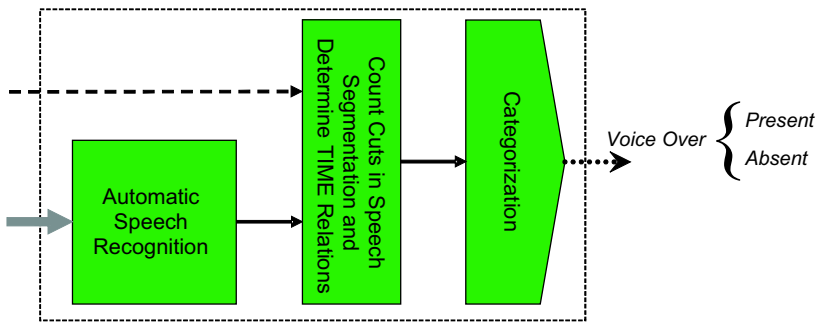


Figure A.6: Voice over detector.

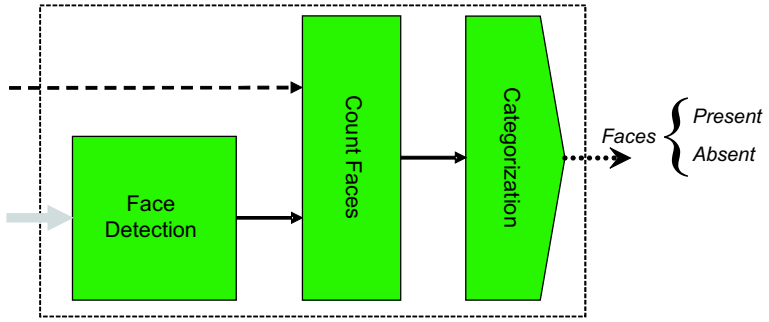


Figure A.7: Faces detector.

Face Location

Since people are important in produced video an author takes great care in filming them, i.e. to make sure they are in the right position. For the location of a detected face we divide an image frame into four equally sized regions: *bottomleft*, *topleft*, *bottomright*, and *topright*. If a face falls completely within one of these four regions the feature for that region is set. If a face covers parts of the bottomleft and topleft part of the image we set the *left* location feature. The *right* location feature works in a similar fashion. If a face can not be fitted into one of these locations the *center* location feature is set, this is illustrated in Fig. A.8. Note that we do not distinguish between top and bottom and that the larger the face the more likely its location is classified as center. This results in a total of seven face location features per detected face in a frame, initially all set to *absent*. We sum the value of all features for all detected faces in a camera shot. To aggregate the frame based face features into a camera shot, we require that a feature is true for 20% of the analyzed frames in a camera shot. If this is the case the feature is set as *present*. The face location detector

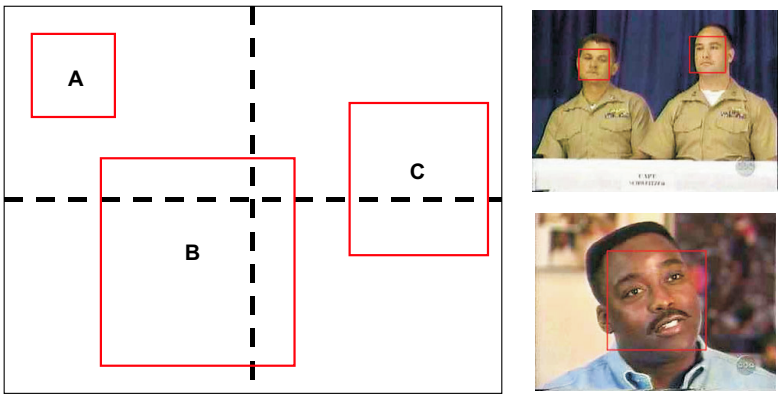


Figure A.8: Left: an image frame with three detected faces, face A is located topleft, face B is located center, and face C is located right. Right: two example image frames with detected faces.

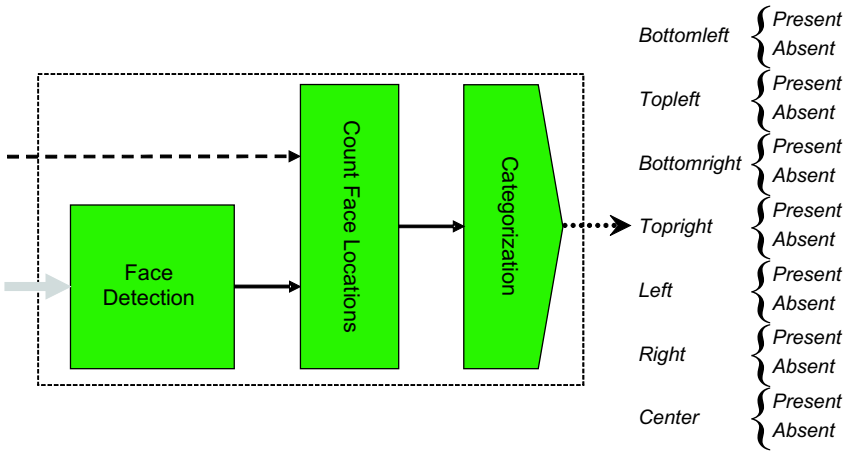


Figure A.9: Face location detector.

scheme is visualized in Fig. A.9.

Frequent Speaker

Another clue for presence of people is their speech. We use the speaker identification results from an automatic speech recognition system [47]. The system provides an identifier for each recognized speaker per analyzed video. Because the identifiers are unique for a single video document only, the recognized speakers do not scale to an entire archive. Moreover, because performance of speaker identification degrades when a large number of speakers appear in a video we do not trust blindly on the results. To accommodate for both issues we determine the three most frequent speakers per video document. Again, we refer to Fig. A.4 for an example. First, we identify the three most frequent speakers. All speech segments that are uttered by one of these frequent speakers are then mapped to camera shots using TIME relations. As before, if a relation exists between these two segmentations we consider a frequent speaker *present* in the shot, else *absent*. The frequent speaker detector scheme is visualized in Fig. A.10.

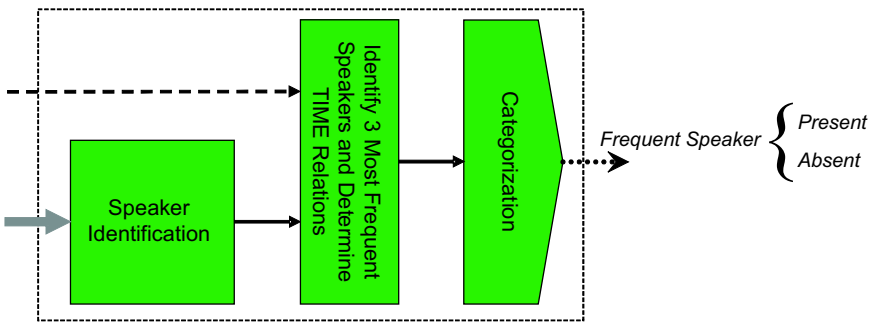


Figure A.10: Frequent speaker detector.

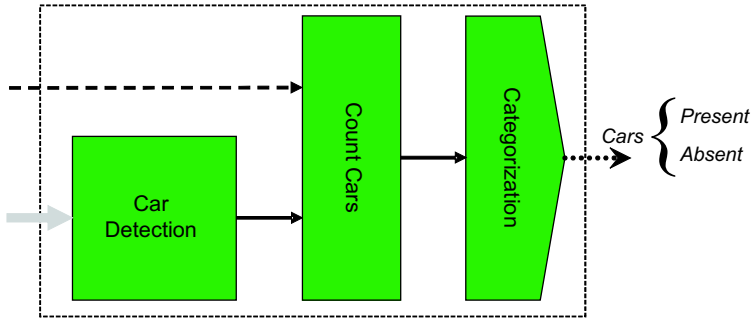


Figure A.11: Car detector.

Cars

Numerous objects appear in the content of produced video. Unfortunately, most object detectors are limited to detection of specific instances only. In our experiments we apply a car detector [130] on individual frames. This detector associates with each detected car in a frame a confidence value from 100 to 400. We consider a car detected in a frame if it has a confidence ≥ 170. We consider it *present* in the entire shot if one analyzed frame within a shot contains a detected car, else *absent*. The cars detector scheme is visualized in Fig. A.11.

Object Motion

Specific object detectors help when you know what to look for. If not, presence of object motion is the best one can hope for. We estimate the amount of motion in a camera shot by spatiotemporal image analysis [74]. We apply a Hanning filter on the x and y projection of a camera shot. This results in a background estimation of the projection. Then we use the projection and the filtered projection to compute the signal energy. We distinguish between three classes of motion based on the signal energy. If the signal energy in a shot has a value below 2 we consider it to be representative for *low* object motion. If the signal energy ranges from 2 to 80 we

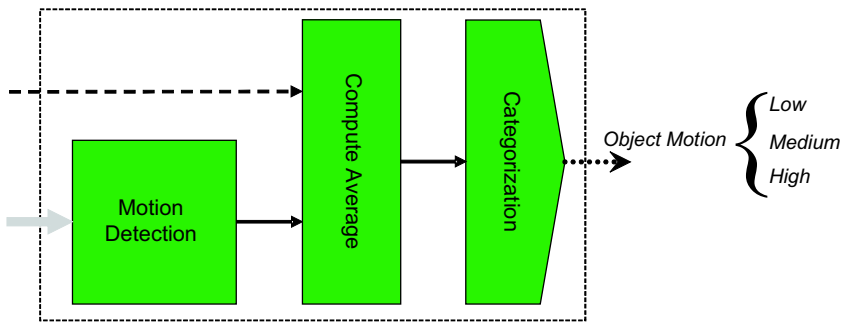


Figure A.12: Object motion detector.

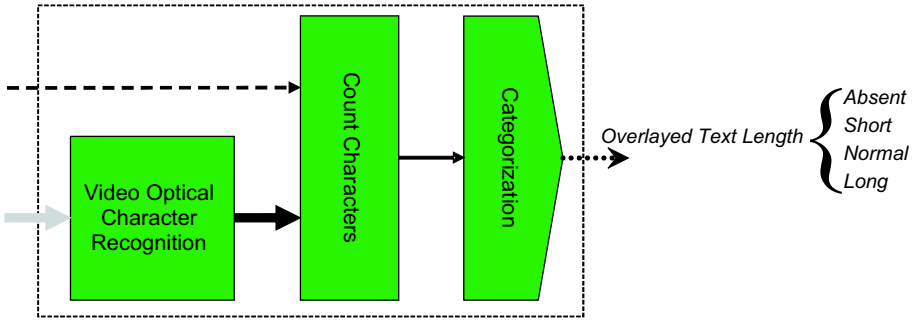


Figure A.13: Overlayed text length detector.

consider it *medium* object motion. In all other cases we consider the shot to contain a *high* amount of motion. The object motion detector scheme is visualized in Fig. A.12.

Overlayed Text Length

Whereas presence of overlayed text is an important indicator for layout style, the length of the recognized overlayed text string tells us something about its intended usage. Names for example are usually short. In contrast, product disclaimers in commercials are usually long. We apply video optical character recognition [125] to recognize overlayed text. We categorize the *overlayed text length* based on the number of recognized characters. We consider the string to be *absent* if less than 5 characters are recognized. This reduces the influence of false positives. We classify the recognized string as *short* if it contains 5 to 20 characters. It is classified as *normal* if it contains 20 to 40 characters. In all other cases it is classified as *long*. The overlayed text length detector scheme is visualized in Fig. A.13.

Video Text Named Entity

Besides the length of recognized overlayed text strings, it is interesting to know the type of annotation, e.g. is it a name of someone who is interviewed or a city scene of some known location. To obtain this information we rely on named entity recognition. Named entity recognition is known from the field of computational linguistics. Given a word, a named entity recognizer classifies it into one of eight categories: person, location, organization, date, time, percentage, monetary value, or none of the above. For our experiments we use a named entity recognizer based on [19], which is described in [170]. Every string recognized by video optical character recognition is input for the named entity recognizer. We distinguish four classes of named entities: none, person, location, and others. Every recognized string is checked for presence of one or more of the named entity types. To aggregate the string based classification to shot level, every string that falls within the boundary of a shot is analyzed for presence of named entities. This results in four features who are initially set to *absent*. If one of the strings in the shot contain one of the four named entities, the respective feature is

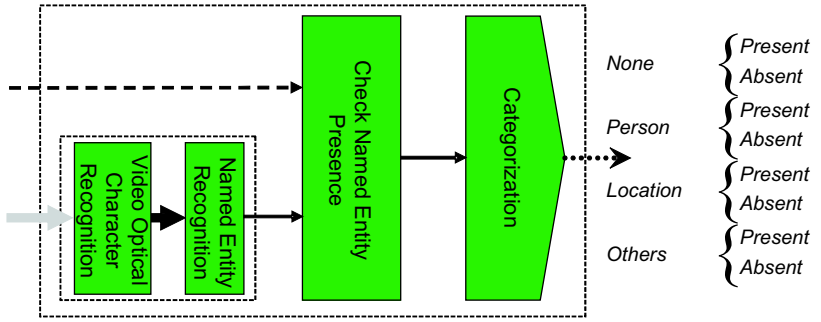


Figure A.14: Video text named entity detector.

set to *present* for the shot. The video text named entity detector scheme is visualized in Fig. A.14.

Voice Named Entity

Similar to video text named entity, we also apply named entity recognition on the transcribed speech obtained from [47]. In contrast to video text named entity, we now classify each word as one of the eight named entities. We consider more named entity types because the textual output of automatic speech recognition is more reliable then the results from video optical character recognition. For each camera shot we define eight features, initially set to *absent*. When one of the words that is uttered during a shot belongs to one of the eight named entity types, the respective feature is set to *present*. The voice named entity detector scheme is visualized in Fig. A.15.

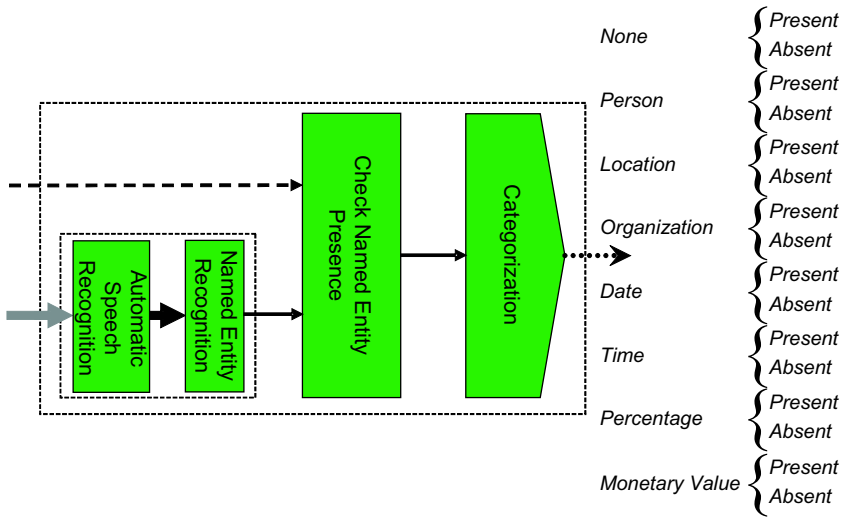


Figure A.15: Voice named entity detector.

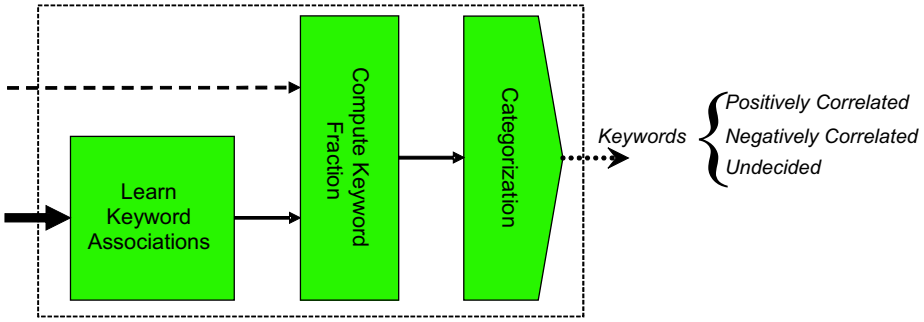


Figure A.16: Positive and negative keyword detection.

Positive and Negative Keywords

The transcribed speech is also analyzed to learn positively correlated and negatively correlated keywords. First we relate all uttered words to a camera shot segmentation. We then remove frequently occurring stopwords using SMART's English stoplist [123]. Given a set of annotated shots containing a certain concept. We learn a list of words that are uttered during these shots. The rationale is that these words probably have a positive relation with the concept under consideration. In a similar fashion we learn a list of words that have a negative relation, by taking all shots that do not have the annotation with the concepts. For unseen data we also relate the uttered words to a camera shot segmentation and remove the stopwords. The remaining words per shot are then compared to the positive word list and the negative word list. Based on the fraction of either positive or negative words in the shot we label a shot as *positively correlated*, *negatively correlated*, or *undecided*. The positive and negative keywords detector scheme is visualized in Fig. A.16.

A.3 Capture Detectors

Camera Distance

As an estimate for the camera distance we use a frame/face ratio proposed in [139]. The ratio relates the width of detected faces to the width of the frame. We compute the face width from faces detected with a face detector [130]. Based on the frame/face ratio we distinguish seven camera distance features: *extreme long shot, long shot, medium long shot, medium shot, medium close up, close up,* and *extreme close up*. For every detected face we determine the camera distance. To obtain a decision at shot level we aggregate all camera distances per analyzed frame. If a camera distance is present in 20% of the analyzed frames we consider this distance *present* in the shot also. When no face is detected in a single frame of a camera shot the camera distance is set to *absent* for all features. In this case, we consider the camera distance unknown. The camera distance detector scheme is visualized in Fig. A.17.

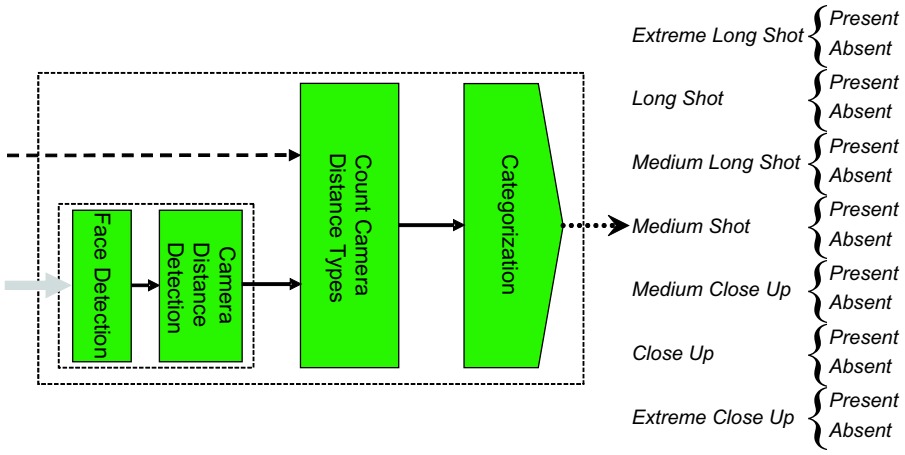


Figure A.17: Camera distance detector.

Camera Work

Several kinds of camera work exist, each creating its own specific effect, see Fig. A.18. For computation of camera work we use an algorithm based on the one reported in [74, 151]. Within a camera shot, the algorithm classifies all frames as belonging to one of six types of camera work: *static, pan, tilt, pan and tilt, zoom,* and *unknown.* To aggregate the frame-based classification to a decision at shot level, we first determine the fraction of frames in the shot that are assigned to one of the six types of camera work. Initially all types of camera work are set as *absent.* The static camera feature for the entire shot is set as *present*, if 90% of the frames in the shot are labeled as static. Each of the other five types of camera work is set as *present* if 10% of the frames in the shot are labeled with the respective operation. The rationale here is that, in general, a camera doesn't move for the entire duration of the shot. Hence, a small fraction of camera work is enough evidence to detect its presence. The camera

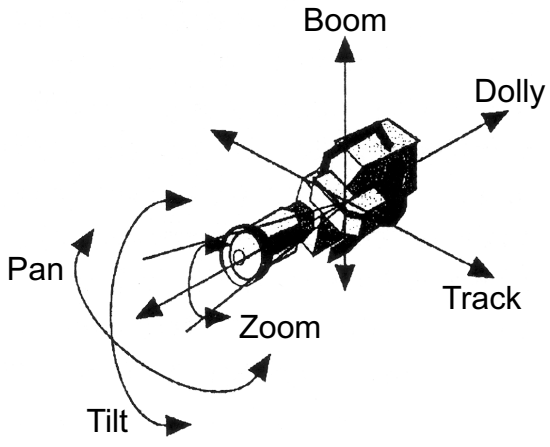


Figure A.18: Several types of camera work, adapted from [151].

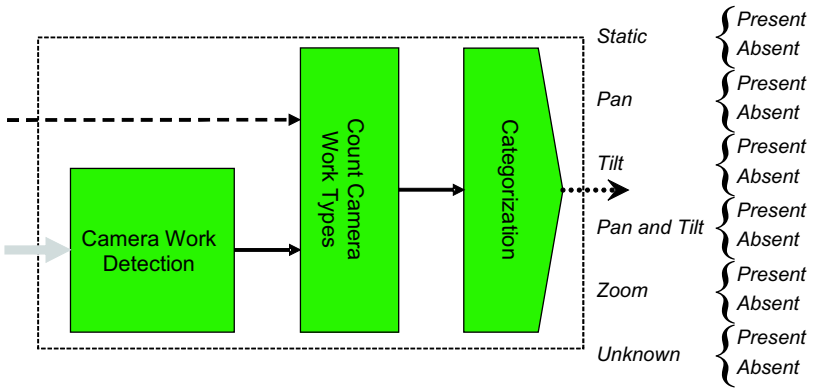


Figure A.19: Camera work detector.

work detector scheme is visualized in Fig. A.19.

Camera Motion

Besides the detection of type of camera work, the above algorithm also indicates the amount of motion that is attached to the camera operation used. We use this camera motion as a feature. For each shot the average amount of camera motion is checked. If the value is below 0.1 we consider camera motion *low* in the shot. If the camera motion ranges from 0.1 to 10 we set the camera motion feature to *medium*. In all other cases the camera motion is set to *high*. The camera motion detector scheme is visualized in Fig. A.20.

A.4 Context Detectors

Commercial

Commercials are added by an author in between broadcasts of programs. In general, viewers interpret commercials as interruptions. As a result we assume that most peo-

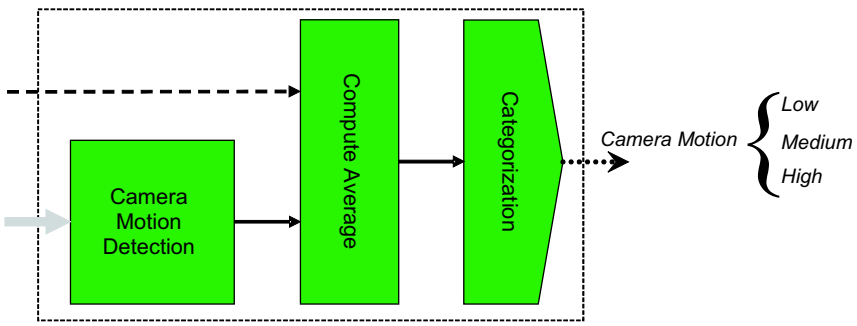


Figure A.20: Camera motion detector.

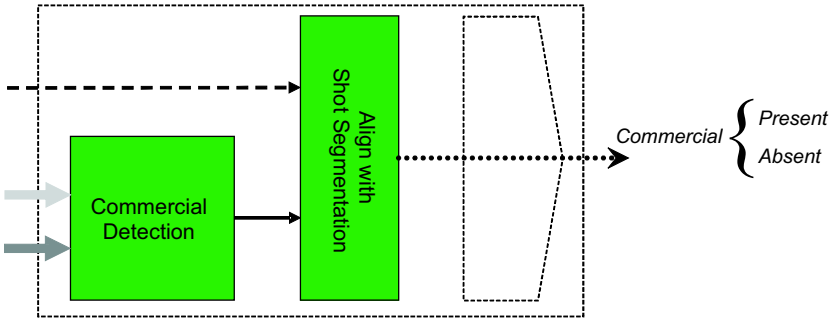


Figure A.21: Commercial detector.

ple are not interested in them. We detect commercials to prevent a false classification of news related content in commercials. We apply the commercial detector proposed in [60]. It labels key frames as a commercial or not. The output is aligned with the camera shot segmentation. The commercial detector scheme is visualized in Fig. A.21.

News Anchor

In broadcast news an author adds a news anchor to summarize the news and to connect news stories. The visual content of shots containing anchors is not very interesting. However, because anchors speak on a large number of topics there textual content may trigger a lot of false positive classifications of other concepts. To prevent this misclassification we apply a news anchor detector [60]. It labels key frames as a news anchor or not. The output is aligned with the camera shot segmentation. The news anchor detector scheme is visualized in Fig. A.22.

News Reporter

Similar to news anchors, news reporters also occur frequently in broadcast news. Again, the visual content of shots containing news reporters is usually of limited interest. To prevent misclassification of shots containing news reporters we apply

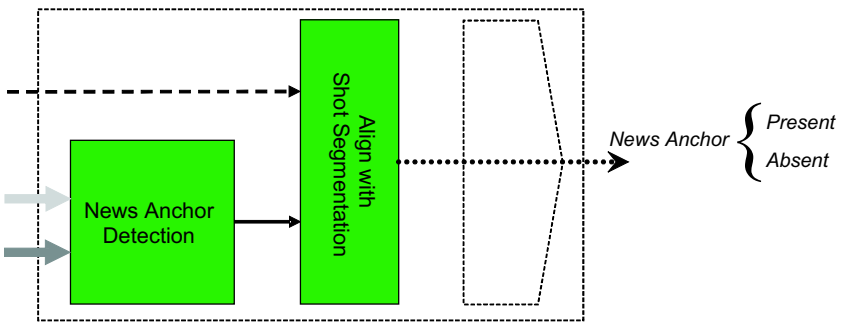


Figure A.22: News anchor detector.

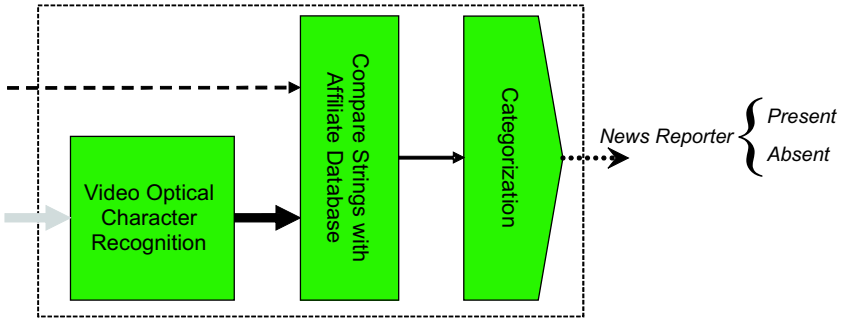


Figure A.23: News reporter detector.

a news reporter detector. It compares each recognized overlayed text string with a database of learned CNN and ABC affiliates. Since results of video optical character recognition contain errors we use the edit distance for comparison. If a match is found during a shot, we label it as a news reporter shot. The news reporter detector scheme is visualized in Fig. A.23.

Bibliography

[1] S. Abney. Part-of-speech tagging and partial parsing. In S. Young and G. Bloothooft, editors, *Corpus-Based Methods in Language and Speech Processing*, pages 118–136. Kluwer Academic Publishers, Dordrecht, 1997.

[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*, 4(4):172–186, 1996.

[3] B. Adams, C. Dorai, and S. Venkatesh. Toward automatic extraction of expressive elements from motion pictures: Tempo. *IEEE Transactions on Multimedia*, 4(4):472–481, 2002.

[4] W. H. Adams, G. Iyengar, C.-Y. Lin, M.R. Naphade, C. Neti, H.J. Nock, and J.R. Smith. Semantic indexing of multimedia content using visual, audio, and text cues. *EURASIP Journal on Applied Signal Processing*, 2003(2):170–185, 2003.

[5] G. Adomavicius and A. Tuzhilin. Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. *IEEE Transactions on Knowledge and Data Engineering*, 17(6):734–749, 2005.

[6] M. Aiello, C. Monz, L. Todoran, and M. Worring. Document understanding for a broad class of documents. *International Journal on Document Analysis and Recognition*, 5(1):1–16, 2002.

[7] 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*, 14(2):137–151, 2001.

[8] J.F. Allen. Maintaining knowledge about temporal intervals. *Communications of the ACM*, 26(11):832–843, 1983.

[9] Y. Altunbasak, P.E. Eren, and A.M. Tekalp. Region-based parametric motion segmentation using color information. *Graphical models and image processing*, 60(1):13–23, 1998.

[10] A. Amir, M. Berg, S.-F. Chang, W. Hsu, G. Iyengar, C.-Y. Lin, M.R. Naphade, A.P. Natsev, C. Neti, H.J. Nock, J.R. Smith, B.L. Tseng, Y. Wu, and D. Zhang. IBM research TRECVID-2003 video retrieval system. In *Proceedings of the TRECVID Workshop*, NIST Special Publication, Gaithersburg, USA, 2003.

[11] J. Assfalg, M. Bertini, A. Del Bimbo, W. Nunziati, and P. Pala. Soccer highlights detection and recognition using HMMs. In *IEEE International Conference on Multimedia & Expo*, Lausanne, Switzerland, 2002.

[12] J. Baan, A. van Ballegooij, J.M. Geusebroek, D. Hiemstra, J. den Hartog, J. List, C. Snoek, I. Patras, S. Raaijmakers, L. Todoran, J. Vendrig, A. de Vries, T. Westerveld, and M. Worring. Lazy users and automatic video retrieval tools in (the) lowlands. In E.M. Voorhees and D.K. Harman, editors, *Proceedings of the 10th Text REtrieval Conference*, volume 500-250 of *NIST Special Publication*, Gaithersburg, USA, 2001.

[13] N. Babaguchi, Y. Kawai, and T. Kitahashi. Event based indexing of broadcasted sports video by intermodal collaboration. *IEEE Transactions on Multimedia*, 4(1):68–75, 2002.

[14] H.E. Bal et al. The distributed ASCI supercomputer project. *Operating Systems Review*, 34(4):76–96, 2000.

[15] 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*, 19(7):711–720, 1997.

[16] A. Berger, S. Della Pietra, and V. Della Pietra. A maximum entropy approach to natural language processing. *Computational Linguistics*, 22(1):39–71, 1996.

[17] M. Bertini, A. Del Bimbo, and P. Pala. Content-based indexing and retrieval of TV news. *Pattern Recognition Letters*, 22(5):503–516, 2001.

[18] M. Bertini, A. Del Bimbo, and P. Pala. Indexing for reuse of TV news shots. *Pattern Recognition*, 35(3):581–591, 2002.

[19] D. Bikel, R. Schwartz, and R.M. Weischedel. An algorithm that learns what's in a name. *Machine Learning*, 34(1-3):211–231, 1999.

[20] Blinkx Video Search, April 2005. `http://www.blinkx.tv/`.

[21] J.M. Boggs and D.W. Petrie. *The Art of Watching Films*. Mayfield Publishing Company, Mountain View, USA, 5th edition, 2000.

[22] R.M. Bolle, B.-L. Yeo, and M.M. Yeung. Video query: Research directions. *IBM Journal of Research and Development*, 42(2):233–252, 1998.

[23] A. Bonzanini, R. Leonardi, and P. Migliorati. Event recognition in sport programs using low-level motion indices. In *IEEE International Conference on Multimedia & Expo*, pages 1208–1211, Tokyo, Japan, 2001.

[24] D. Bordwell and K. Thompson. *Film Art: An Introduction*. McGraw-Hill, New York, USA, 5th edition, 1997.

[25] L. Breiman. Bagging predictors. *Machine Learning*, 24(2):123–140, 1996.

[26] M. Brown, J. Foote, G. Jones, K. Sparck-Jones, and S. Young. Automatic content-based retrieval of broadcast news. In *ACM Multimedia*, San Francisco, USA, 1995.

[27] R. Brunelli, O. Mich, and C.M. Modena. A survey on the automatic indexing of video data. *Journal of Visual Communication and Image Representation*, 10(2):78–112, 1999.

[28] C.-C. Chang and C.-J. Lin. *LIBSVM: a library for support vector machines*, 2001. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm/`.

[29] M. Christel, C. Huang, N. Moraveji, and N. Papernick. Exploiting multiple modalities for interactive video retrieval. In *IEEE International Conference on Acoustic, Speech and Signal Processing*, volume 3, pages 1032–1035, Montreal, Canada, 2004.

[30] M. Christel, A. Olligschlaeger, and C. Huang. Interactive maps for a digital video library. *IEEE Multimedia*, 7(1):60–67, 2000.

[31] T.-S. Chua, S.-F. Chang, L. Chaisorn, and W. Hsu. Story boundary detection in large broadcast news video archives - techniques, experience and trends. In *ACM Multimedia*, New York, USA, 2004.

[32] C. Colombo, A. Del Bimbo, and P. Pala. Semantics in visual information retrieval. *IEEE Multimedia*, 6(3):38–53, 1999.

[33] Convera, December 2001. `http://www.convera.com`.

[34] J.N. Darroch and D. Ratcliff. Generalized iterative scaling for log-linear models. *The Annals of Mathematical Statistics*, 43(5):1470–1480, 1972.

[35] G. Davenport, T.G. Aguierre Smith, and N. Pincever. Cinematic principles for multimedia. *IEEE Computer Graphics & Applications*, 11(4):67–74, 1991.

[36] 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*, 41(6):391–407, 1990.

[37] N. Dimitrova, L. Agnihotri, and G. Wei. Video classification based on HMM using text and faces. In *European Signal Processing Conference*, Tampere, Finland, 2000.

[38] 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*, pages 2997–3000, Phoenix, USA, 1999.

[39] A. Ekin, A.M. Tekalp, and R. Mehrotra. Automatic soccer video analysis and summarization. *IEEE Transactions on Image Processing*, 12(7):796–807, 2003.

[40] 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*, pages 2445–2448, Istanbul, Turkey, 2000.

[41] J. Fan, A.K. Elmagarmid, X. Zhu, W.G. Aref, and L. Wu. *ClassView*: hierarchical video shot classification, indexing, and accessing. *IEEE Transactions on Multimedia*, 6(1):70–86, 2004.

[42] C. Fellbaum, editor. *WordNet: an electronic lexical database*. The MIT Press, Cambridge, USA, 1998.

[43] S.M. Fisch and R.T. Truglio, editors. *"G" is for Growing: Thirty Years of Research on Children and Sesame Street*. Lawrence Erlbaum Associates, Mahwah, USA, 2001.

[44] S. Fischer, R. Lienhart, and W. Effelsberg. Automatic recognition of film genres. In *ACM Multimedia*, pages 295–304, San Francisco, USA, 1995.

[45] M.M. Fleck, D.A. Forsyth, and C. Bregler. Finding naked people. In *European Conference on Computer Vision*, volume 2, pages 593–602, Cambridge, UK, 1996.

[46] B. Furht, S.W. Smoliar, and H.-J. Zhang. *Video and Image Processing in Multimedia Systems*. Kluwer Academic Publishers, Norwell, USA, 2th edition, 1996.

[47] J.L. Gauvain, L. Lamel, and G. Adda. The LIMSI broadcast news transcription system. *Speech Communication*, 37(1–2):89–108, 2002.

[48] J.M. Geusebroek, R. van den Boomgaard, A.W.M. Smeulders, and H. Geerts. Color invariance. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(12):1338–1350, 2001.

[49] Th. Gevers and A.W.M. Smeulders. Content-based image retrieval: An overview. In G. Medioni and S.B. Kang, editors, *Emerging Topics in Computer Vision*. Prentice Hall, 2004.

[50] A. Ghias, J. Logan, D. Chamberlin, and B.C. Smith. Query by humming – musical information retrieval in an audio database. In *ACM Multimedia*, San Francisco, USA, 1995.

[51] Y. Gong, L.T. Sin, and C.H. Chuan. Automatic parsing of TV soccer programs. In *IEEE International Conference on Multimedia Computing and Systems*, pages 167–174, 1995.

[52] Google Video Search, April 2005. `http://video.google.com/`.

[53] 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.

[54] 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*, 10(6):857–868, 2000.

[55] 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.

[56] M. Han, W. Hua, W. Xu, and Y. Gong. An integrated baseball digest system using maximum entropy method. In *ACM Multimedia*, Juan-les-Pins, France, 2002.

[57] 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.

[58] 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.

[59] A.G. Hauptmann. Towards a large scale concept ontology for broadcast video. In *International Conference on Image and Video Retrieval*, volume 3115 of *LNCS*, pages 674–675. Springer-Verlag, 2004.

[60] A.G. Hauptmann, R.V. Baron, M.-Y. Chen, M. Christel, P. Duygulu, C. Huang, R. Jin, W.-H. Lin, T. Ng, N. Moraveji, N. Papernick, C.G.M. Snoek, G. Tzanetakis, J. Yang, R. Yan, and H.D. Wactlar. Informedia at TRECVID 2003: Analyzing and searching broadcast news video. In *Proceedings of the TRECVID Workshop*, NIST Special Publication, Gaithersburg, USA, 2003.

[61] A.G. Hauptmann and M.G. Christel. Successful approaches in the TREC video retrieval evaluations. In *ACM Multimedia*, New York, USA, 2004.

[62] 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.

[63] A.G. Hauptmann and M.J. Witbrock. Story segmentation and detection of commercials in broadcast news video. In *ADL-98 Advances in Digital Libraries*, pages 168–179, Santa Barbara, USA, 1998.

[64] T.K. Ho, J.J. Hull, and S.N. Srihari. Decision combination in multiple classifier systems. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 16(1):66–75, 1994.

[65] 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.

[66] IBM Research. MARVEL MPEG-7 video search engine, April 2005. `http://mp7.watson.ibm.com/marvel/`.

[67] I. Ide, K. Yamamoto, and H. Tanaka. Automatic video indexing based on shot classification. In *First International Conference on Advanced Multimedia Content Processing*, volume 1554 of *LNCS*, pages 87–102, Osaka, Japan, 1999. Springer-Verlag.

[68] A.K. Jain, R.P.W. Duin, and J. Mao. Statistical pattern recognition: A review. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(1):4–37, 2000.

[69] R. Jain and A. Hampapur. Metadata in video databases. *ACM SIGMOD*, 23(4):27–33, 1994.

[70] P.J. Jang and A.G. Hauptmann. Learning to recognize speech by watching television. *IEEE Intelligent Systems*, 14(5):51–58, 1999.

[71] 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*, pages 366–369, Thessaloniki, Greece, 2001.

[72] 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.

[73] E.T. Jaynes. Information theory and statistical mechanics. *The Physical Review*, 106(4):620–630, 1957.

[74] P. Joly and H.-K. Kim. Efficient automatic analysis of camera work and microsegmentation of video using spatiotemporal images. *Signal Processing: Image Communication*, 8(4):295–307, 1996.

[75] 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*, volume 3972, pages 332–343, 2000.

[76] 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.

[77] R. Lau, R. Rosenfeld, and S. Roukos. Adaptive language modelling using the maximum entropy approach. In *ARPA Human Language Technologies Workshop*, pages 81–86, Princeton, USA, 1993.

[78] H. Lee and A.F. Smeaton. Designing the user-interface for the Físchlár digital video library. *Journal of Digital Information*, 2(4), 2002.

[79] R. Leonardi and P. Migliorati. Semantic indexing of multimedia documents. *IEEE Multimedia*, 9(2):44–51, 2002.

[80] D. Li, I.K. Sethi, N. Dimitrova, and T. McGee. Classification of general audio data for content-based retrieval. *Pattern Recognition Letters*, 22(5):533–544, 2001.

[81] H. Li, D. Doermann, and O. Kia. Automatic text detection and tracking in digital video. *IEEE Transactions on Image Processing*, 9(1):147–156, 2000.

[82] R. Lienhart, C. Kuhmünch, and W. Effelsberg. On the detection and recognition of television commercials. In *IEEE Conference on Multimedia Computing and Systems*, pages 509–516, Ottawa, Canada, 1997.

[83] C.-Y. Lin, B.L. Tseng, and J.R. Smith. Video collaborative annotation forum: Establishing ground-truth labels on large multimedia datasets. In *Proceedings of the TRECVID Workshop*, NIST Special Publication, Gaithersburg, USA, 2003.

[84] W.-H. Lin and A.G. Hauptmann. News video classification using SVM-based multimodal classifiers and combination strategies. In *ACM Multimedia*, Juan-les-Pins, France, 2002.

[85] B.S. Manjunath, P. Salembier, and T. Sikora, editors. *Introduction to MPEG-7: Multimedia Content Description Interface*. Wiley, 2002.

[86] C.D. Manning and H. Schütze. *Foundations of Statistical Natural Language Processing*. The MIT Press, Cambridge, USA, 1999.

[87] K. Minami, A. Akutsu, H. Hamada, and Y. Tonomura. Video handling with music and speech detection. *IEEE Multimedia*, 5(3):17–25, 1998.

[88] 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*, pages 26–30, Grenoble, France, 2000.

[89] A. Mohan, C. Papageorgiou, and T. Poggio. Example-based object detection in images by components. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(4):349–361, 2001.

[90] S. Moncrieff, C. Dorai, and S. Venkatesh. Detecting indexical signs in film audio for scene interpretation. In *IEEE International Conference on Multimedia & Expo*, pages 1192–1195, Tokyo, Japan, 2001.

[91] F. Nack and A.T. Lindsay. Everything you always wanted to know about MPEG-7: Part 1. *IEEE Multimedia*, 6(3):65–77, 1999.

[92] F. Nack and A.T. Lindsay. Everything you always wanted to know about MPEG-7: Part 2. *IEEE Multimedia*, 6(4):64–73, 1999.

[93] J. Nam, M. Alghoniemy, and A.H. Tewfik. Audio-visual content-based violent scene characterization. In *IEEE International Conference on Image Processing*, volume 1, pages 353–357, Chicago, USA, 1998.

[94] J. Nam, A.E. Cetin, and A.H. Tewfik. Speaker identification and video analysis for hierarchical video shot classification. In *IEEE International Conference on Image Processing*, volume 2, pages 550–553, Washington DC, USA, 1997.

[95] M.R. Naphade. On supervision and statistical learning for semantic multimedia analysis. *Journal of Visual Communication and Image Representation*, 15(3):348–369, 2004.

[96] M.R. Naphade and T.S. Huang. A probabilistic framework for semantic video indexing, filtering, and retrieval. *IEEE Transactions on Multimedia*, 3(1):141–151, 2001.

[97] M.R. Naphade and T.S. Huang. Extracting semantics from audiovisual content: The final frontier in multimedia retrieval. *IEEE Transactions on Neural Networks*, 13(4):793–810, 2002.

[98] M.R. Naphade, I.V. Kozintsev, and T.S. Huang. A factor graph framework for semantic video indexing. *IEEE Transactions on Circuits and Systems for Video Technology*, 12(1):40–52, 2002.

[99] M.R. Naphade and J.R. Smith. On the detection of semantic concepts at TRECVID. In *ACM Multimedia*, New York, USA, 2004.

[100] 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*, 9(1):137–141, 2000.

[101] L. Nigay and J. Coutaz. A design space for multimodal systems: concurrent processing and data fusion. In *INTERCHI'93 Proceedings*, pages 172–178, Amsterdam, The Netherlands, 1993.

[102] NIST. TRECVID Video Retrieval Evaluation, 2001–2004. `http://www-nlpir.nist.gov/projects/trecvid/`.

[103] D.W. Oard. The state of the art in text filtering. *User Modeling and User-Adapted Interaction*, 7(3):141–178, 1997.

[104] 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.

[105] 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*, volume 2670, pages 373–384, San Jose, USA, 1996.

[106] 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.

[107] J. Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann, San Mateo, USA, 1988.

[108] 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.

[109] 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.

[110] S. Pfeiffer, S. Fischer, and W. Effelsberg. Automatic audio content analysis. In *ACM Multimedia*, pages 21–30, Boston, USA, 1996.

[111] S. Pfeiffer, R. Lienhart, and W. Effelsberg. Scene determination based on video and audio features. *Multimedia Tools and Applications*, 15(1):59–81, 2001.

[112] T.V. Pham and M. Worring. Face detection methods: A critical evaluation. Technical Report 2000-11, Intelligent Sensory Information Systems, University of Amsterdam, 2000.

[113] J.C. Platt. Probabilities for SV machines. In A.J. Smola, P.L. Bartlett, B. Schölkopf, and D. Schuurmans, editors, *Advances in Large Margin Classifiers*, pages 61–74. MIT Press, 2000.

[114] Praja, December 2001. `http://www.praja.com`.

[115] G.M. Quénot, D. Moraru, L. Besacier, and P. Mulhem. CLIPS at TREC-11: Experiments in video retrieval. In E.M. Voorhees and L.P. Buckland, editors, *Proceedings of the 11th Text REtrieval Conference*, volume 500-251 of *NIST Special Publication*, Gaithersburg, USA, 2002.

[116] J.R. Quinlan. *C4.5: Programs for Machine Learning*. Morgan Kaufmann, 1993.

[117] S. Raaijmakers, J. den Hartog, and J. Baan. Multimodal topic segmentation and classification of news video. In *IEEE International Conference on Multimedia & Expo*, Lausanne, Switzerland, 2002.

[118] L.R. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989.

[119] M. Rautiainen, T. Ojala, and T. Seppänen. Analysing the performance of visual, concept and text features in content-based video retrieval. In *ACM SIGMM International Workshop on Multimedia Information Retrieval*, pages 197–204, New York, USA, 2004.

[120] H.A. Rowley, S. Baluja, and T. Kanade. Neural network-based face detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(1):23–38, 1998.

[121] Y. Rui, A. Gupta, and A. Acero. Automatically extracting highlights for TV baseball programs. In *ACM Multimedia*, pages 105–115, Los Angeles, USA, 2000.

[122] E. Sahouria and A. Zakhor. Content analysis of video using principal components. *IEEE Transactions on Circuits and Systems for Video Technology*, 9(8):1290–1298, 1999.

[123] G. Salton and M.J. McGill. *Introduction to Modern Information Retrieval*. McGraw-Hill, New York, USA, 1983.

[124] 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.

[125] T. Sato, T. Kanade, E.K. Hughes, M.A. Smith, and S. Satoh. Video OCR: Indexing digital news libraries by recognition of superimposed caption. *Multimedia Systems*, 7(5):385–395, 1999.

[126] S. Satoh, Y. Nakamura, and T. Kanade. Name-It: Naming and detecting faces in news videos. *IEEE Multimedia*, 6(1):22–35, 1999.

[127] 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*, volume 3022, pages 176–187, San Jose, USA, 1997.

[128] R.E. Schapire. The strength of weak learnability. *Machine Learning*, 5(2):197–227, 1990.

[129] 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.

[130] H. Schneiderman and T. Kanade. Object detection using the statistics of parts. *International Journal of Computer Vision*, 56(3):151–177, 2004.

[131] F.J. Seinstra, C.G.M. Snoek, D. Koelma, J.M. Geusebroek, and M. Worring. User transparent parallel processing of the 2004 NIST TRECVID data set. In *International Parallel and Distributed Processing Symposium*, Denver, USA, 2005.

[132] 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*, pages 46–53, Boston, USA, 2000.

[133] 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*, pages 618–620, 1998.

[134] A.F. Smeaton, W. Kraaij, and P. Over. TRECVID 2003 - an introduction. In *Proceedings of the TRECVID Workshop*, NIST Special Publication, Gaithersburg, USA, 2003.

[135] A.F. Smeaton, H. Lee, and K. McDonald. Experiences of creating four video library collections with the Físchlár system. *International Journal on Digital Libraries*, 4(1):42–44, 2004.

[136] 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*, 22(12):1349–1380, 2000.

[137] J.R. Smith and S.-F. Chang. Visually searching the web for content. *IEEE Multimedia*, 4(3):12–20, 1997.

[138] J.R. Smith et al. MARVEL: Multimedia analysis and retrieval system, April 2005. White Paper.

[139] C.G.M. Snoek. Camera distance classification: Indexing video shots based on visual features. Master's thesis, Universiteit van Amsterdam, October 2000.

[140] C.G.M. Snoek and M. Worring. Time interval maximum entropy based event indexing in soccer video. In *Proceedings of the IEEE International Conference on Multimedia & Expo*, volume 3, pages 481–484, Baltimore, USA, 2003.

[141] C.G.M. Snoek and M. Worring. Multimedia event-based video indexing using time intervals. *IEEE Transactions on Multimedia*, 7(4):638–647, 2005.

[142] C.G.M. Snoek and M. Worring. Multimodal video indexing: A review of the state-of-the-art. *Multimedia Tools and Applications*, 25(1):5–35, 2005.

[143] C.G.M. Snoek, M. Worring, J.M. Geusebroek, D.C. Koelma, and F.J. Seinstra. The MediaMill TRECVID 2004 semantic video search engine. In *Proceedings of the TRECVID Workshop*, NIST Special Publication, Gaithersburg, USA, 2004.

[144] C.G.M. Snoek, M. Worring, J.M. Geusebroek, D.C. Koelma, F.J. Seinstra, and A.W.M. Smeulders. The semantic value chain: A unifying architecture for generic indexing of multimedia archives. Submitted for publication.

[145] C.G.M. Snoek, M. Worring, and A.G. Hauptmann. Learning rich semantics from produced video by style analysis. Submitted for publication.

[146] C.G.M. Snoek, M. Worring, and A.G. Hauptmann. Detection of TV news monologues by style analysis. In *Proceedings of the IEEE International Conference on Multimedia & Expo*, Taipei, Taiwan, 2004.

[147] C.G.M. Snoek, M. Worring, D.C. Koelma, and A.W.M. Smeulders. A lexicon-driven paradigm for interactive multimedia retrieval. Submitted for publication.

[148] R.K. Srihari. Automatic indexing and content-based retrieval of captioned images. *IEEE Computer*, 28(9):49–56, 1995.

[149] 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.

[150] 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.

[151] Y. Tonomura, A. Akutsu, Y. Taniguchi, and G. Suzuki. Structured video computing. *IEEE Multimedia*, 1(3):34–43, 1994.

[152] B.T. Truong and S. Venkatesh. Determining dramatic intensification via flashing lights in movies. In *IEEE International Conference on Multimedia & Expo*, pages 61–64, Tokyo, Japan, 2001.

[153] 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.

[154] 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*, 11(4):522–535, 2001.

[155] B.L. Tseng, C.-Y. Lin, M.R. Naphade, A. Natsev, and J.R. Smith. Normalized classifier fusion for semantic visual concept detection. In *IEEE International Conference on Image Processing*, volume 2, pages 535–538, Barcelona, Spain, 2003.

[156] 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*, volume 3972, San Jose, USA, 2000.

[157] A. Vailaya, A.K. Jain, and H.-J. Zhang. On image classification: City images vs. landscapes. *Pattern Recognition*, 31:1921–1936, 1998.

[158] V.N. Vapnik. *The Nature of Statistical Learning Theory*. Springer-Verlag, New York, USA, 2th edition, 2000.

[159] J. Vendrig and M. Worring. Systematic evaluation of logical story unit segmentation. *IEEE Transactions on Multimedia*, 4(4):492–499, 2002.

[160] Virage, December 2001. `http://www.virage.com`.

[161] H.D. Wactlar, M.G. Christel, Y. Gong, and A.G. Hauptmann. Lessons learned from building a terabyte digital video library. *IEEE Computer*, 32(2):66–73, 1999.

[162] Y. Wang, Z. Liu, and J. Huang. Multimedia content analysis using both audio and visual clues. *IEEE Signal Processing Magazine*, 17(6):12–36, 2000.

[163] T. Westerveld. Image retrieval: Content versus context. In *Content-Based Multimedia Information Access, RIAO 2000 Conference*, pages 276–284, Paris, France, 2000.

[164] T. Westerveld, A.P. de Vries, A. van Ballegooij, F. de Jong, and D. Hiemstra. A probabilistic multimedia retrieval model and its evaluation. *EURASIP Journal on Applied Signal Processing*, 2003(2):186–197, 2003.

[165] E. Wold, T. Blum, D. Keislar, and J. Wheaton. Content-based classification, search, and retrieval of audio. *IEEE Multimedia*, 3(3):27–36, 1996.

[166] D.H. Wolpert. Stacked generalization. *Neural Networks*, 5:241–259, 1992.

[167] M. Worring, G.P Nguyen, L. Hollink, J. van Gemert, and D.C. Koelma. Interactive search using indexing, filtering, browsing and ranking. In *Proceedings of the TRECVID Workshop*, NIST Special Publication, Gaithersburg, USA, 2003.

[168] L. Wu, J. Benois-Pineau, and D. Barba. Spatio-temporal segmentation of image sequences for object-oriented low bit-rate image coding. *Image Communication*, 8(6):513–544, 1996.

[169] 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*, pages 928–931, Tokyo, Japan, 2001.

[170] R. Yan, J. Yang, and A.G. Hauptmann. Learning query-class dependent weights for automatic video retrieval. In *ACM Multimedia*, New York, USA, 2004.

[171] M.-H. Yang, D.J. Kriegman, and N. Ahuja. Detecting faces in images: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(1):34–58, 2002.

[172] 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*, volume 3022, pages 45–58, 1997.

[173] D. Yow, B.-L. Yeo, M.M. Yeung, and B. Liu. Analysis and presentation of soccer highlights from digital video. In *Asian Conference on Computer Vision*, Singapore, 1995.

[174] H.-J. Zhang, A. Kankanhalli, and S.W. Smoliar. Automatic partitioning of full-motion video. *Multimedia Systems*, 1(1):10–28, 1993.

[175] H.-J. Zhang, S.Y. Tan, S.W. Smoliar, and Y. Gong. Automatic parsing and indexing of news video. *Multimedia Systems*, 2(6):256–266, 1995.

[176] H.-J. Zhang, J. Wu, D. Zhong, and S.W. Smoliar. An integrated system for content-based video retrieval and browsing. *Pattern Recognition*, 30(4):643–658, 1997.

[177] T. Zhang and C.-C. Jay Kuo. Hierarchical classification of audio data for archiving and retrieving. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 6, pages 3001–3004, Phoenix, USA, 1999.

[178] D. Zhong and S.-F. Chang. Structure analysis of sports video using domain models. In *IEEE International Conference on Multimedia & Expo*, pages 920–923, Tokyo, Japan, 2001.

[179] Y. Zhong, H.-J. Zhang, and A.K. Jain. Automatic caption localization in compressed video. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(4):385–392, 2000.

[180] W. Zhou, S. Dao, and C.-C. Jay Kuo. On-line knowledge- and rule-based video classification system for video indexing and dissemination. *Information Systems*, 27(8):559–586, 2002.

[181] W. Zhou, A. Vellaikal, and C.-C. Jay Kuo. Rule-based video classification system for basketball video indexing. In *ACM Multimedia*, Los Angeles, USA, 2000.

[182] 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*, pages 1036–1039, Tokyo, Japan, 2001.

Author Index

Samenvatting*

Dit proefschrift levert een bijdrage aan het vakgebied dat zich bezighoudt met het automatisch begrijpen van multimedia, hier verder multimedia-leren genoemd. Ons ultieme doel is het structuren van de multimedia-chaos door de semantische kloof te overbruggen tussen berekenbare data-eigenschappen enerzijds en de semantische interpretatie van deze data door een gebruiker anderzijds. We maken daartoe eerst een onderscheid tussen geproduceerde en ongeproduceerde multimedia, in het bijzonder videodocumenten. We gaan uit van de aanname dat een geproduceerde video het resultaat is van een auteursgedreven productieproces. Dit proces dient als een metafoor voor multimedia-leren. We presenteren een geleidelijke uitwerking van deze metafoor voor multimedia-leren. In deze uitwerking geven we een uitgebreid overzicht van het veld, een theoretische grondslag voor multimedia-leren, state-of-the-art benchmark validatie en praktische toepassingen voor het semantisch ontsluiten van video. De auteursgedreven methodologie voor het semantisch indexeren van multimedia is de voornaamste bijdrage van dit proefschrift.

In Hoofdstuk 2 leggen we de basis voor de auteursmetafoor. We introduceren een multimediaal raamwerk waarbij we een videodocument bezien vanuit het perspectief van de auteur. Binnen het raamwerk beschouwen we lay-out, inhoud en de semantische index als de significante componenten. Door een videodocument te zien als het resultaat van een auteursproces zijn we in staat om de visuele, auditieve en tekstuele media consistent te integreren. Bovendien vormt het raamwerk het leidende principe voor het identificeren van indextypen waarvoor automatische methoden in de literatuur bestaan. Het raamwerk verenigt en categoriseert deze methoden en dient als een blauwdruk voor een generiek en flexibel semantisch video-indexeersysteem dat is gebaseerd op multimediale analyse.

Het gebruik van meerdere typen media voor semantisch indexeren vormt een probleem met betrekking tot synchronisatie en integratie van temporele contextaanwijzingen. Om dit integratieprobleem aan te pakken, introduceren we in Hoofdstuk 3 het Time Interval Multimedia Event (TIME) raamwerk. Het raamwerk gaat expliciet

*Summary, in Dutch.

om met context en synchronisatie en omdat het raamwerk gebaseerd is op statistiek resulteert het in een robuuste methode voor multimediale integratie. We focusseren op het probleem van het combineren van auteurselementen, in de vorm van inhoud en lay-out segmentaties, in een algemeen analyseraamwerk. We modelleren deze lay-out- en inhoudsegmentaties als tijdsintervallen om de beperkingen van bestaande methoden voor media-integratie te overwinnen. De tijdsintervalrepresentatie stelt ons in staat om de integratie van temporele context en synchronisatie op een juiste manier af te handelen. Daarnaast laten we zien dat een aantal statistische classificatoren toepasbaar zijn voor het semantisch indexeren van video gebaseerd op een tijdsinterval patroonrepresentatie. Om de effectiviteit van TIME aan te tonen is het geëvalueerd op twee domeinen, te weten voetbal en nieuws. Het domein voetbal is gekozen om zijn afhankelijkheid van context. Het domein nieuws is gekozen om zijn afhankelijkheid van synchronisatie. We hebben drie statistische classificatoren, met variërende complexiteit, vergeleken en hebben aangetoond dat er een duidelijke relatie bestaat tussen de engte van de semantische kloof en de complexiteit van de benodigde classificator. Bovendien, hebben we laten zien dat het TIME raamwerk, inclusief synchronisatie en context, significant betere resultaten behaald dan de in de literatuur voorkomende 'standaardmethoden' voor multimediale analyse.

Zodra we in staat zijn om multimediale informatiebronnen correct te fuseren, zijn we in Hoofdstuk 4 klaar om de notie van stijl toe te voegen aan het repertoire van multimedia-leren. Naast lay-out en inhoud, identificeren we opname en context als belangrijke aspecten van de auteursstijl. We beschrijven een generiek en flexibel raamwerk voor geproduceerde video dat in staat is om rijke semantische concepten te filteren uit multimediale bronnen gebaseerd op stijlanalyse. Met rijke semantiek bedoelen we dat stijl in vele opzichten benut wordt door de auteur. Het raamwerk stelt ons in staat om verscheidene rijke semantische concepten in geproduceerde video te classificeren. We maken daartoe gebruik van een vaste kern van lay-out-, inhoud- en opname-detectoren samen met variërende context-detectoren die gecombineerd worden in een ensemble van statistische classificatoren. Resultaten op 120 uur videodata van de TRECVID 2003 benchmark laten zien dat het de combinatie van stijlelementen is die het beste resultaat geeft voor het indexeren van geproduceerde video. Bovendien demonstreren we dat de accuraatheid van het voorgestelde raamwerk voor classificatie van verscheidene rijke semantische concepten state-of-the-art is.

In Hoofdstuk 5 weiden we verder uit over de auteursgedreven analysemethodologie. We introduceren een generieke methode voor semantisch indexeren, gebaseerd op de auteursmetafoor, die we de semantische waardeketen noemen. Om de semantische kloof te overbruggen verenigt de keten ons werk van Hoofdstuk 2, 3 en 4 met recente ontwikkelingen in het vakgebied, in een algemene systeemarchitectuur. De architectuur is gebouwd op verschillende gespecialiseerde detectoren, multimediale analyse, hypothese selectie en automatisch leren en bovendien omvat het de noties van inhoud, stijl en context. De semantische waardeketen extraheert semantische concepten vanuit videodocumenten door drie opeenvolgende analyseschakels te doorlopen, genaamd: de inhoudschakel, de stijlschakel en de contextschakel. De semantische waardeketen bepaalt automatisch per concept een optimale configuratie van analyseschakels en op basis hiervan komen we tot een techniektaxonomie voor detectoren van semantische

concepten. Experimenten met een lexicon van 32 semantische concepten demonstreren dat de semantische waardeketen in staat is om op generieke wijze video te indexeren. Daarnaast is de semantische waardeketen succesvol geëvalueerd binnen de TRECVID 2004 benchmark als beste performer voor de semantische concept-detectietaak. De resultaten laten zien dat de semantische waardeketen in staat is om generiek te indexeren met state-of-the-art performance.

De semantische kloof dicteert dat slechts een beperkt lexicon van semantische concepten automatisch geleerd kan worden, dus uiteindelijk is betrokkenheid van de gebruiker essentieel. Daarom focusseren we in Hoofdstuk 6 op interactieve ontsluiting van multimedia. We presenteren een paradigma voor lexicongedreven ontsluiting om multimedia-archieven toegankelijk te maken. Het fundament van het paradigma wordt gevormd door het lexicon van 32 semantische concepten, zoals gedetecteerd in Hoofdstuk 5. Gebaseerd op dit lexicon, wordt gebruikers semantische toegang tot multimedia-archieven geboden in de vorm van zoekmogelijkheden op een conceptueel niveau. Daarnaast wordt gebruikers een entree geboden in de vorm van similariteit, door gebruik te maken van tekstuele en visuele voorbeelden. Interactie met de verschillende zoekinterfaces wordt afgehandeld door een video zoekmachine, die feedback geeft in de vorm van storyboard resultaten. Het lexicongedreven paradigma combineert leren, similariteit en interactietechnieken om de semantische kloof in multimedia ontsluiting te overbruggen. Het paradigma is geëvalueerd binnen de interactieve zoektaak van de TRECVID 2004 benchmark, gebruikmakend van een archief van 184 uur aan nieuwsuitzendingen. Uit de experimenten blijkt dat het lexicongedreven zoekparadigma hoogst effectief is voor interactieve ontsluiting van multimedia. Daarnaast demonstreren we dat het paradigma resulteert in de best mogelijke zoekresultaten wanneer gebruikers kennis hebben van de concepten in het lexicon en hun te verwachten performance.

De in Hoofdstukken 3, 4, 5 en 6 ontwikkelde technologie leidt vanzelfsprekend naar de concretisering van semantische zoekmachines voor video. In Hoofdstuk 7 presenteren we een algemene architectuur voor een dergelijke zoekmachine, die bestaat uit een archief van televisie-uitzendingen, een indexeermachine, componenten voor index afgeleidde diensten en een zoekinterface. We leggen de nadruk op verscheidene aspecten van de algemene architectuur door middel van vier prototypesystemen: *Goalgle, News RePortal, Viper* en het *MediaMill* systeem.

Aan het eind van dit proefschrift zijn we klaar om onszelf af te vragen of we geslaagd zijn in het beantwoorden van de fundamentele vraag: *hoe de semantische kloof te overbruggen voor geproduceerde video?* De geleidelijke uitwerking van de auteursmetafoor levert ons een effectief oplossingspad. Voor automatische analyse resulteert het in de semantische waardeketen, gepresenteerd in Hoofdstuk 5. We bereiken een voorlopig eindpunt in onze zoektocht naar multimedia-leren wanneer we de semantische waardeketen combineren met het paradigma voor interactieve multimedia ontsluiting in Hoofdstuk 6. *Een combinatie van automatische auteursgedreven analyse en gebruikersinteractie, resulteert in de meest effectieve benadering om de semantische kloof te overbruggen.*

Concluderend: met de auteursmetafoor hebben we het vakgebied voor automatisch begrijpen van multimedia vooruit gestuwd met een effectieve methodologie die

de semantische kloof substantieel nauwer maakt. We zijn vol vertrouwen dat een verlengde verkenning langs de door ons ingeslagen weg, in de vorm van toekomstig onderzoek, de alom aanwezige multimedia-chaos verder zal structureren.

Dankwoord[†]

Het promotietraject dat leidde tot dit proefschrift is voor mij als een ontdekkingsreis door een wonderlijk land geweest. Door de behulpzaamheid van velen is de reis bovendien aangenaam verlopen. Onderweg heb ik het meeste opgestoken van mijn co-promotor. Marcel stippelde de route uit en hielp me onderweg weer op het juiste pad te geraken als ik een onbeduidend zijweggetje dreigde te nemen, of even geen zin had om verder te *trec*-en. Door zijn vraag: *"Wat als anderen stijl toevoegen aan hun analyse?"*, kwam uiteindelijk ook het belangrijkste inzicht. Het antwoord was immers door de wekelijkse één-op-één sessies al voorgeprogrammeerd: *"Draai het om!"*. Marcel, mede door je professionele begeleiding en prettige manier van samenwerken is het reisverslag voltooid, waarvoor mijn hartelijke dank.

Tegen het einde van de reis mocht ik nog even met promotor Arnold op pad. Het werd een dollemansrit langs de richel van de semantische kloof. Arnold leerde me op de valreep (nog) beter kijken, luisteren en schrijven. Het is een geruststellende gedachte om te beseffen dat er ook na deze zoektocht nog veel te leren valt. Arnold, bedankt voor de inspiratie, de schrijflessen en het faciliteren van de reis.

Besides the figurative journey, there was also a literal trip. I am grateful to Howard for giving me the opportunity to be part of the Informedia team at Carnegie Mellon University. Thanks to Alex's supervision the period in the USA has been an eye opener; not only for what can be achieved, but also for what can not. In that sense it forms a landmark in my PhD process. I would like to thank all the Informedia folks for their hospitality, assistance, and the many delicious pizza's. Moreover, I thank the team for the generous permission to use their ibox software after my return to Amsterdam. Special thanks go to Pinar and Dorbin for the evenings out and the Pittsburgh tours. I hope we will meet again someday soon.

Met medereizigers en ISIS collegae is het behalve gezellig keuvelen en koffiedrinken ook uitstekend samenwerken. Ik ben veel dank verschuldigd aan Jeroen V. voor het delen van zijn kennis en kunde betreffende video-analyse. Ook Jan-Mark, voor de complexere beeldbewerking, en Frank S., voor de razendsnelle verwerking daarvan,

[†] Acknowledgements, mostly in Dutch.

hebben een belangrijk steentje bijgedragen. Dennis verdient een eervolle vermelding. Naast vaste waarde aan de koffietafel, heeft hij ook een substantiële invloed gehad op de Hoofdstukken 5 en 6 in de vorm van ongeëvenaarde software support en het ontwerp van de 'golden demo'. Dennis mijn oprechte dank hiervoor. Kamergenoten van het lab worden bedankt voor het creëren van een prettige werkomgeving en het beschikbaar stellen van hun machines in de zomermaanden. De mensen op de lange lijst met overige namen huldig ik voor hun behulpzaamheid bij dringende kwesties, zoals het verhelpen van LaTeX en matlab probleempjes, en hun contributie aan de fijne sfeer binnen het instituut.

TNO financierde de reis in de vorm van klinkende munt en een expertisecentrum onder de noemer MediaMill. De belangstelling en vrijheid die me bij het onderzoek werd gegeven door Jurgen en Nellie heb ik daarbij altijd zeer op prijs gesteld. Als beginnend AiO heb ik bovendien veel gehad aan de kennis van Jan Baan, wiens spullen tot op de dag van vandaag waardevolle tools zijn binnen onze video-indexeerpraktijk. De overige MediaMillers ben ik een bedankje schuldig voor hun hulp bij de diverse demo's die in de loop der jaren zijn ontwikkeld.

Als promovendus word je maar al te vaak bezig gehouden met zaken die verre staan van het bedrijven van wetenschap, maar daarom zeker niet onbelangrijk zijn. Het is een bijzondere luxe om dan terug te kunnen vallen op de goede zorgen van toegewijde mensen. Vanaf deze plaats wil ik daarom de mannen van support, Liesbeth en bovenal Virginie in het zonnetje zetten.

Vrienden en vriendinnen roem ik voor hun interesse in de vorderingen van het onderzoek en hun volharding om me achter de PC vandaan te krijgen. Mede-kopstukken Richard A. en Niels B. wisten mij gedurende het traject geregeld te herinneren aan het feit dat er meer in het leven is dan video indexing en het 'semantische gat'. Het spijt me dat het bezoeken van foute kroegen en het aflopen van vage feestjes er in de beslissende fase van het onderzoek niet meer van gekomen is. Jullie rol als paranimf is in dat opzicht een goedmakertje voor de verloren uurtjes.

Het voltooien van een proefschrift is geen eindpunt. Het is slechts het passeren van een tussenstation op een grotere reis. Ik prijs mezelf bij die tocht gelukkig met de nimmer aflatende steun en goede zorgen van mijn (schoon)familie. In het bijzonder wil ik mijn ouders Paul en Woltje bedanken voor het vervullen van een voorbeeldfunctie de afgelopen jaren en het van jongs af aan stimuleren van mijn keuzes. Dit proefschrift is daarom ook een beetje van jullie.

Tot slot, zonder de onvoorwaardelijke liefde, het eindeloze begrip en engelengeduld van één persoon was de reis door wonderland nooit volbracht: Marga, lest best.

- Cees