Learning Rich Semantics from News Video Archives by Style Analysis

CEES G. M. SNOEK and MARCEL WORRING
University of Amsterdam
and
ALEXANDER G. HAUPTMANN
Carnegie Mellon University

We propose a generic and robust framework for news video indexing which we founded on a broadcast news production model. We identify within this model four production phases, each providing useful metadata for annotation. In contrast to semiautomatic indexing approaches which exploit this information at production time, we adhere to an automatic data-driven approach. To that end, we analyze a digital news video using a separate set of multimodal detectors for each production phase. By combining the resulting production-derived features into a statistical classifier ensemble, the framework facilitates robust classification of several *rich* semantic concepts in news video; rich meaning that concepts share many similarities in their production process. Experiments on an archive of 120 hours of news video from the 2003 TRECVID benchmark show that a combined analysis of production phases yields the best results. In addition, we demonstrate that the accuracy of the proposed style analysis framework for classification of several rich semantic concepts is state-of-the-art.

Categories and Subject Descriptors: H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing—*Indexing methods*; I.2.6 [**Artificial Intelligence**]: Learning—*Concept learning*

General Terms: Algorithms, Management, Performance

Additional Key Words and Phrases: Benchmark evaluation, multimedia understanding, multimodal detectors, news video indexing, semantic classification, statistical pattern recognition, style analysis

1. INTRODUCTION

The advancement in optical fiber technology, coupled with a growing availability of low-cost multimedia recording devices, enables the massive capture, delivery, and exchange of large amounts of digital video. This overwhelming amount of digital video data will trigger the need for automatic indexing tools. Ideally, these tools should provide on-the-fly content-based annotation at a conceptual level. Once the video is semantically annotated, this allows for effective and efficient browsing, filtering, and retrieval of specific video fragments. Unfortunately, automatic techniques for indexing video with conceptual labels suffer from the fact that it is hard to infer semantics based on features extracted from the data. This

Part of this research was performed while C. G. M. Snoek was a visiting scientist at the Informedia group of Carnegie Mellon University. This research is sponsored by the ICES/KIS MIA project and TNO.

Contact author's address: C. G. M. Snoek, M. Worring, Informatics Institute, University of Amsterdam, Kruislaan 403, 1098 SJ Amsterdam, The Netherlands, email: cgmsnoek@science.uva.nl.

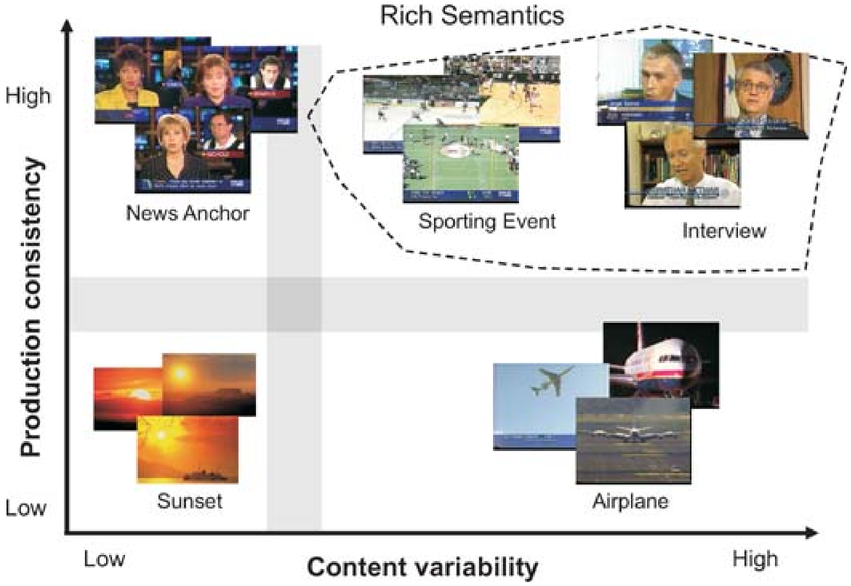

Fig. 1. Relation between content variability and production consistency. Rich semantics have both a high variability in content and a high consistency in production.

semantic gap [Smeulders et al. 2000] has hampered the development of a generic index solution. Thus, the progress in content-based multimedia analysis has not kept pace with the multimedia enabling technology push.

To tackle the semantic video indexing problem, two solution paths have emerged. The first road to an effective semantic video index circumvents the problem of mapping features to semantics by actively involving the creator of the video in the annotation process [Barry and Davenport 2003; Davis 2003; Nack and Putz 2004]. While this approach is fruitful for present day video production, it is unsuited for the large amount of digital video stored in video repositories already. In addition, there might be several reasons why authors are unwilling to share valuable production information. Therefore, the second solution path for semantic indexing of video focuses analysis on the raw multimedia data of the final video product. Most work in this direction emphasizes an analysis based on content only. This approach yields adequate results for concepts that are easy to distinguish because of their large similarity in (visual) content, for example, tigers [Smith and Li 1998] and soccer games [Xie et al. 2004]. For concepts that have more variability in their content such as buildings, sporting events, and dialogues, analysis methods based on content only are too fragile. We study a particular class of concepts—we call *rich semantics*—which share many similarities in their production process while being very different content-wise. Both the concepts sporting events and dialogues, for example, are often recorded in different settings but from a fixed camera distance. Rich semantics are illustrated in Figure 1. The question arises whether it is possible to combine the two video indexing approaches by extracting rich semantic concepts from video archives, using features which are related to the entire production process.

Others have also probed whether video production techniques are applicable to the problem of semantic video indexing. There is a large body of work on segmenting feature films and situational comedies into logical story units using techniques from film art [Adams et al. 2002; Hanjalic et al. 1999; Sundaram and Chang 2002; Truong et al. 2005; Vendrig and Worring 2002]. In general, these approaches demonstrate the feasibility of production-derived features for semantic indexing by focusing on one specific technique from film art only. Then, the authors are able to map features to semantics based on knowledge-based classification rules. In Adams et al. [2002] for example, the authors use *tempo* to identify events and dramatic sections, while in Sundaram and Chang [2002] the 180° *rule* is exploited to detect dialogues. However, the generic applicability and robustness of these indexing methods is limited because of their rule dependency.

To cope with both generic applicability and robustness, the focus in semantic video indexing has recently shifted to machine learning based approaches, for example, Naphade et al. [2002], Adams et al. [2003], Amir et al. [2003], Hauptmann et al. [2003], and Snoek and Worring [2005a]. These approaches are able to index video based on a large collection of automatically derived decision rules while exploiting multiple features. This has resulted in increased robustness for automatic classification of various semantic concepts. However, none of these indexing methods use production-derived features.

In this article, we take the idea of using production-derived features for semantic video indexing a step further. We focus on the domain of news video. This class of video adheres to a strict production format, but its content changes on an hourly basis. At the same time, it is also a class of video which is actively archived by several broadcast stations and cultural heritage institutions. To study the problem of semantic news video indexing, we first derive a model of the news video production process. Based on this model, we propose a generic and robust framework for news video indexing. Our framework is generic because we learn various rich semantic concepts in news from a fixed set of features related to the news video production model. The framework guarantees robustness by integrating all production-derived features into a statistical classifier ensemble. The framework is unique in its combined usage of production-derived features and machine learning—we call *style analysis*—for the purpose of generic and robust rich semantic indexing of news video archives.

The organization of the remainder of this article is as follows. First, we discuss related work in more detail. Then, we proceed with a model for news video production in Section 3. We present the style analysis framework for news video indexing in Section 4. We discuss an implementation of the framework in Section 5. An extensive evaluation on the 2003 NIST TRECVID video retrieval benchmark [Smeaton et al. 2003] demonstrating the applicability of our framework is presented in Section 6.

2. RELATED WORK IN SEMANTIC VIDEO INDEXING

Initial work on semantic video indexing started with visual analysis only. A good example of this exploratory work is Zhang et al. [1997]. The authors focus on the parsing and indexing of news video. Based on extracted motion features, they are able to classify concepts such as crowds and talking heads. Because of the exploratory nature of this work, experiments are carried out on a small-scale video archive only. In addition, the multimodal nature of news video is ignored.

Large scale news video indexing using multimodal analysis has been pioneered by the Informedia project at Carnegie Mellon University [Wactlar et al. 1999]. Their approach focuses 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. The current system is shifting towards the usage of more advanced learning schemes [Hauptmann et al. 2003]. However, production-derived features are still largely ignored.

Naphade et al. [2002] were among the early adopters of advanced pattern recognition techniques for semantic classification of video. In Naphade et al. [2002] 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, for example, outdoor is based on detection of vegetation and sky. By combining the individual probabilities of all multijects into a Multinet using factor graphs [Naphade et al. 2002], 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 such as rocky terrain, body of water, and forestry. This indicates that this method is mostly suited for nonrich semantics.

An extended and truly multimodal version of Naphade et al. [2002] was presented in Adams et al. [2003] Amir et al. [2003], 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 Tseng et al. [2003]). All multimodal concepts then serve as the input for the Multinet that uses the combination of concepts for a final semantic classification. This approach has demonstrated good results on the concept detection task of the NIST TRECVID benchmark. 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. 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 various modalities. Finally, the approach ignores production-derived features. Hence, the current pipeline approach is not optimal for the detection of rich semantic concepts in news video.

A framework for synchronization of multiple modalities and the inclusion of temporal clues was proposed in Snoek and Worring [2005a]. Viewing the result of individual detectors as time intervals allows for the combination of various detectors 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 largely ignores production-derived features.

Combining the previously described, by explicitly modeling multimodal production-derived features into a machine learning framework, we are able to detect rich semantics in large-scale news video archives more accurately.

3. NEWS VIDEO PRODUCTION MODEL

To arrive at a style analysis framework for news video indexing, we first consider broadcast news production. A news video is the work of an author who conceives an initial idea for news coverage and finally produces a result, semantically reflecting this idea as well as possible. To communicate a semantic intention by means of a video, an author has an arsenal of techniques to choose from [Boggs and Petrie 2000; Bordwell and Thompson 1997]. The choice for a specific set of techniques is restricted only by the imagination of the author and the format of the news video to be produced.

In practice, the author will not make all possible technical decisions in isolation. The author relies for specific tasks on a production team of specialists. For the general domain of professional video, people distinguish between the *preproduction* phase, the *production* phase, and the *postproduction* phase in the creation process [Bordwell and Thompson 1997; Davis 2003; Nack and Putz 2004]. The production phase can be specified further into a production *design* phase and a production *recording* phase [Bordwell and

Thompson 1997]. Thus, we identify the following four phases for news video production:

—*Preproduction*: identification of news topics and production planning;

—*Production design*: provision of guidelines for the arrangement of setting, objects, and people;

—*Production recording*: shooting of the audiovisual material;

—*Postproduction*: editing of audiovisual footage and adding of special effects;

The blueprint of any news broadcast is the production planning, as prepared in the preproduction phase. In this phase, several journalists identify news topics of interest. Guided by the preproduction phase, the production design phase defines the content of a news video by arranging people, objects, and setting, where possible. Choices for the news studio include the number of anchors, decoration of the studio, and type of weather map. However, production design is not limited exclusively to studio design, for example, for outdoor footage care is taken in posting reporters on the right spot. The production recording phase guides the capturing of the news video content into a multimedia format. One recording team works in the studio and another one on location. Both teams take care of specific recording circumstances such as camera framing, lightning, and the balance and combination of microphones. The editor is responsible for the assembly and synchronization of individual pieces in the postproduction phase. The editor assures that the voiceover of the reporter is in line with the visible content. In addition, the task of the editor is to add descriptive information to the news video in the form of overlayed text. The author is responsible for the complex interplay of all production phases to communicate a semantic intention within the news video.

As observed by Nack and Putz [2004], each broadcast news production phase provides useful metadata. An author may exploit this information for rich semantic annotation of broadcast news at production time. Unfortunately, creation of daily news episodes is often a once-only production. As a consequence, precious metadata is lost when it is not immediately stored. This is often the case for news episodes already archived in repositories. For rich semantic indexing of news video archives, a data-driven analysis is the only viable alternative.

4. A STYLE ANALYSIS FRAMEWORK FOR NEWS VIDEO INDEXING

4.1 Style Detectors

A broadcast news episode is the result of an authoring process. Hence, analysis of news video should focus on its author-driven production process. Since a news video is often available in a raw data format only, we need to identify as many of the production choices as possible using detectors. The creative choices made during the video production process are commonly referred to as the author's style [Boggs and Petrie 2000]. Therefore, we refer to detectors that aim to reconstruct production choices as *style* detectors. We group these detectors into four sets based on the phases identified for news video production in Section 3. We cannot analyze phases directly from the data. Therefore, a style detector can, at best, approximate the result of each creative phase involved in news video production. Thus, we analyze a news video using four sets of style detectors which are related to the entire author-driven production process.

The layout results after the editor is finished with the news video in the postproduction phase. Therefore, the first set of style detectors we use for analysis are *layout* detectors. We follow our previous work, Snoek and Worring [2005b], for its definition:

Definition 4.1 *Layout Detectors*. Layout detectors are the set of style detectors $\mathcal{L}$ that yields an approximation of the sensor shots, transition edits, and special effects of a news video.

As layout is modality-specific, the set of layout detectors is limited by the number of modalities involved. Examples of layout detectors include, but are not limited to, shot segmentation, tempo, overlayed text, and voiceovers. When an editor chooses to use a special effect, this has no consequences for the sensor shot used. Thus, for layout, detectors for various editing elements act independently of each other.

In the production recording phase, people use 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. Hence, the second set of style detectors we use for analysis are *capture* detectors. We define:

Definition 4.2 *Capture Detectors*. Capture detectors are the set of style detectors $\mathcal{T}$ that yields an approximation of the recording parameters used to transfer an observed scene into the sensory format of a news video.

Like layout detectors, the number of possibilities for capture detectors are bounded, but this time by the degrees of freedom of the recording sensors and devices. For camera and microphone sensors, examples include their distance, angle, and motion. The fact that the production 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.

The production design defines the 3D content world of a news video. The content is obtained by arranging people, objects, and setting. For its detection we rely on *content* detectors. We define:

Definition 4.3 *Content Detectors*. Content detectors are the set of style detectors $\mathcal{C}$ that yields an approximation of the people, objects, and setting appearing in a news video.

In contrast to layout and capture detectors, the set of possible content detectors is unlimited in theory (see Snoek and Worring [2005b] for an overview). Examples of content detectors include face detectors, specific object detectors, and explosion sound detectors. Because of the numerous possibilities for content detectors, a dependent combination may exist. Consider, for example, a news topic on the problems of foreigners in a particular country, where people and setting are clearly dependent. In general, however, we assume content detectors act independently of each other.

For news video archives the production planning of individual episodes is seldom available. At the same time, however, their reconstruction based on raw data seems almost impossible. We need a means to enhance or limit the number of possible semantic interpretations of a news video segment. As an estimate of the preproduction phase, we therefore exploit the notion of context [Dey 2001; Dourish 2004]. We rely on *context* detectors for its detection. We adapt the definition of Dey [2001] and define:

Definition 4.4 *Context Detectors*. Context detectors are the set of style detectors $\mathcal{S}$ that yields an approximation of the information that can be used to characterize the situation of a rich semantic concept in news video.

Similar to content detectors, the possibilities for context detectors are unrestricted. We again refer to Snoek and Worring [2005b] for a general overview. Examples of context detectors for news include indoor detectors, reporter detectors, and commercial detectors. For context detectors a dependent combination may also exist; but we assume that context detectors act independently of each other in the analysis.

A style detector-based analysis of production phases results in the mapping of a news video to the detector set $\{\mathcal{L}, \mathcal{C}, \mathcal{T}, \mathcal{S}\}$. As an author exploits all production phases to communicate a semantic intention, the individual style detectors need to be combined into a common representation. This involves synchronization, since detector results from the various modalities are not necessarily aligned. Synchronization has largely been ignored in video indexing literature. It is typically solved by aligning all detection results to a camera shot layout segmentation, although better schemes exist [Snoek and Worring 2005a]. For style analysis, we define:

Definition 4.5 *Style Vector*. A style vector is a vector $\vec{s}_i$ that contains the synchronized result of the detector ensemble $\{\mathcal{L}, \mathcal{C}, \mathcal{T}, \mathcal{S}\}$, with the desired property that individual components are independent, where i indicates the segmentation used.

The style vector, resulting from the synchronization and concatenation of individual detectors, contains production-derived features from the entire broadcast news creation process. It forms the basis for the generic and robust learning of rich semantics from produced news video.

4.2 Semantic Classifier

We perceive detection of rich semantics from production-derived features as a supervised pattern recognition problem. We aim to detect a rich semantic class ω based on an ensemble of independent style detectors, represented in a style vector $\vec{s}_i$, using the probability $p(\omega|\vec{s}_i)$. This requires a classifier combination scheme. This scheme combines the results of several independent classifiers or detectors that solve the same task. However, there is no reason to assume that the same technique cannot be used to combine classifiers that do not solve the same task *per se*, but are related semantically, that is, they 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 style 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 [Jain et al. 2000]. Since we have designed $\vec{s}_i$ as an ensemble of independent style detectors, a classifier combination scheme is a natural choice for learning rich semantics from news video.

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

Definition 4.6 *Style Model*. A style model is a model resulting after applying a classifier combination scheme to a set of style vectors.

In literature various classifier combination schemes exist, for example, bagging [Breiman 1996], boosting [Schapire 1990], and stacking [Wolpert 1992]. Bagging and boosting resample a data set to obtain an ensemble (or series) of independent classifiers. They differ from stacking schemes in the way they combine the individual results. Both bagging and boosting focus on the data and exploit independence by combining classifiers that are trained on different samples of the 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 detectors, an assurance for independence exists and there is no need for resampling. Hence, for our purpose, that is, the detection of rich semantics, stacking is a good choice.

The probabilistic output $p(\omega|\vec{s}_i)$ obtained from a stacked classifier allows us to define new context detectors. Suppose we construct a style model for the detection of political reporters in broadcast news using n style detectors for each style vector. When we apply this model to a set of shot-segmented news broadcasts, it results in a probability of occurrence of political reporters for each shot. We can then add this concept, together with its probability, as a context detector and use it for a style model containing $n+1$ style detectors for each style vector, which detects segments where a politician speaks on a topical subject. Besides positive correlation, negative correlation is also helpful as a context detector. It aids in preventing false positive classification of semantically different concepts that share similarity in their production process, for example, political and financial reporters. Moreover, it can be exploited

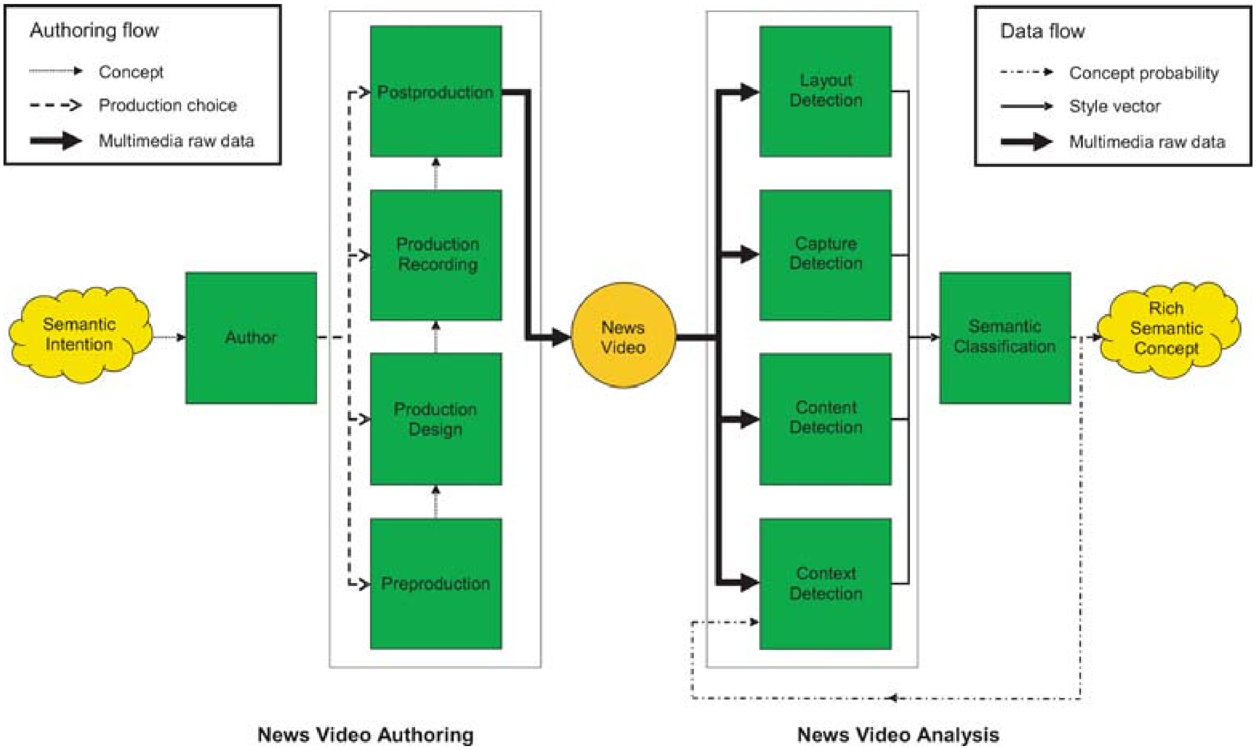

Fig. 2. Style analysis framework for news video indexing.

for the detection of rich semantics that are defined by what it is not, for example, nonpolitical news events. Note that by iteratively adding concepts to the style vector in the form of context detectors, 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 detectors are added to the style vector can be defined by domain knowledge, experimentation, or feature selection techniques [Jain et al. 2000].

Summarizing the aforementioned, we aim to learn rich semantic concepts from news video using a restricted set of style detectors which addresses aspects of the entire broadcast news production process. The semantic classifier used for detection of rich semantics exploits supervised learning in combination with stacking. The complete style analysis framework for indexing news video, in terms of rich semantics, is shown in Figure 2.

5. AN EXPERIMENT ON A NEWS VIDEO ARCHIVE

We participated in the 2003 NIST TRECVID video retrieval benchmark [Smeaton et al. 2003] to demonstrate the applicability of our framework for automatic concept detection. The news video archive for the 2003 benchmark totaled 120 hours of ABC World News Tonight and CNN Headline News from the first half of 1998.[1] NIST split the archive into an equally sized training and test set, each containing about 60 hours of produced news video in MPEG-1 format. Together with the video archive came automatic speech recognition results donated by LIMSI [Gauvain et al. 2002]. CLIPS-IMAG [Quénot et al. 2002]

[1] We ignore the 13 hours of C-SPAN video from the official 2003 TRECVID benchmark data set in our experiments, since these are not related to the news domain.

provided a camera shot segmentation. These shots serve as the basic unit of testing and performance assessment within the TRECVID benchmark.

In total, 17 semantic concepts were defined by NIST to be detected in this archive. Most concepts can be classified as content-based, relating to setting, objects, and people. We focus our evaluation on the classification of rich semantic news concepts which share many similarities in their production processes, namely: *news subject monologue*, *nonstudio setting*, *sporting event*, and *weather news*. To show the limitations of the framework, we also include results for two nonrich semantic concepts, that is, *aircraft* and *vegetation*. NIST defined all concepts as follows [Smeaton et al. 2003]:

—*News subject monologue*: The 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 acceptable if short;

—*Nonstudio setting*: The segment is not set in a TV broadcast studio;

—*Sporting event*: The segment contains video of one or more organized sporting events;

—*Weather news*: The segment reports on the weather;

—*Aircraft*: The segment contains at least one aircraft of any sort; and

—*Vegetation*: The segment contains living vegetation in its natural environment.

Note that the interpretation of concept definitions is subject to manual judgment as performed by NIST. We automatically detect all six concepts, in a generic fashion, using an implementation of the style analysis framework.

5.1 Semantic Classifier Implementation

As a stacked semantic classifier we chose the support vector machine (SVM) [Vapnik 2000; Chang and Lin 2001], 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 [Adams et al. 2003; Snoek and Worring 2005a].

As the SVM adheres to a supervised learning paradigm, we needed positive examples for every concept under consideration. We manually labeled a subset of the training set, of about 24 hours in total, to obtain these examples for each semantic concept.

We performed a parameter search for each concept to optimize the settings for the SVM in our classification scheme. This parameter search accounts for balancing positive and negative examples. In addition, it optimizes the classifier, in case the data is not perfectly separable, using so-called slack variables [Vapnik 2000]. Both data balancing and the use of slack variables are required for the detection of rich semantics, since rich semantics are never evenly balanced in the data and never perfectly separable. We used ten-fold cross validation [Jain et al. 2000] on the training set for the parameter search.

To allow for the combination of style models, we converted the classification result of the SVM, that is, the margin, to a calibrated result. Ideally, we 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 Platt [2000]. 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:

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

where the parameters α and β are maximum likelihood estimates based on the training set [Platt 2000]. The stacked SVM allows for semantic classification using the probabilistic output $p(\omega|\vec{s}_i)$.

5.2 Style Detector Implementation

For all four production phases discussed in Section 3, style detectors were developed. We do not claim to be complete in the current set of detectors; rather, we advocate that for each of the four production phases style, detectors should be included. We selected the style detectors based on established observations from years of media science research, for example, Boggs and Petrie [2000] and Bordwell and Thompson [1997], which also provide suggestions for additional detectors that could be added to the style analysis framework. The current set largely builds upon the Informedia news video analysis toolbox, which has proven its utility in the news domain in previous work [Wactlar et al. 1999; Hauptmann et al. 2003]. We chose 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 [Chang and Lin 2001]. Moreover, this weakens individual detector classifiers even more, which has a positive side effect on the classifier combination scheme. To make detectors discrete we used two procedures. On the numerical output, thresholds were applied. We mapped categorical output to a discrete number. We optimized all detectors and thresholds based on experiments using the training set. We synchronized all discrete detector results, referred to as production-derived features, to the granularity of the provided shot segmentation. We refer to the electronic appendix for specific implementation details of all style detectors.

For the layout $\mathcal{L}$, the length of a camera shot was used as a feature that characterizes tempo [Adams et al. 2002]. The presence of overlayed text, added by the editor at production time, was detected by a text localization algorithm [Sato et al. 1999]. A microphone segmentation using speech and silence detection was based on the results provided by the LIMSI speech detection system [Gauvain et al. 2002]. We obtained a voiceover detector by combining the speech segmentation with the camera shot segmentation. The total set of layout detectors is given by: $\mathcal{L} = \{$*shot length, overlayed text, silence, voiceover*$\}$.

On the content $\mathcal{C}$, a frontal face detector [Schneiderman and Kanade 2004] was applied to detect people. For each analyzed frame in a shot we counted the number of faces, and for each face we derived one of seven possible locations. In addition, we measured the average amount of object motion in a camera shot [Snoek and Worring 2005a]. Based on speaker identification [Gauvain et al. 2002] we were able to identify each of the three most frequent speakers. Each camera shot was checked for the presence on the basis of speech from one of the three. For all semantic concepts under consideration, we learned a list of positive and negative correlated keywords using the training set. Stopwords were removed using SMART's English stoplist [Salton and McGill 1983]. 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 [Sato et al. 1999] were checked on length and used as input for a named entity recognizer [Wactlar et al. 1999]. The total set of content detectors is given by: $\mathcal{C} = \{$*faces, face location, object motion, frequent speaker, positive keywords, negative keywords, overlayed text length, video text named entity*$\}$.

From the size of detected faces [Schneiderman and Kanade 2004] 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 closeup. When no face was detected the camera distance was set as unknown. In addition to camera distance, several types of camera work were detected [Baan et al. 2001], for example, 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 [Baan et al. 2001], which was either high, medium, or low. The total set of capture detectors is given by: $\mathcal{T} = \{$*camera distance, camera work, camera motion*$\}$.

The possibilities for detectors of context $\mathcal{S}$ are endless. We restricted ourselves to the ones available in the Informedia toolbox. 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 able to detect commercials [Hauptmann et al. 2003]. News anchors also share many characteristics with news subject monologues, so it is therefore important that we can distinguish between anchors to circumvent a false interpretation. Moreover, anchors aid in the detection of studio setting. We applied an anchor detector to stress this importance [Hauptmann et al. 2003]. 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 basic set of context detectors is given by: $\mathcal{S} = \{$*commercial, news anchor, news reporter*$\}$.

Based on a concatenation of $\{\mathcal{L}, \mathcal{C}, \mathcal{T}, \mathcal{S}\}$ into a style vector $\vec{s}_i$, we were able to train a style model. Since we aimed for the detection of six semantic concepts and we wanted to exploit context, we had to define order. The order was chosen based on domain knowledge and training set performance. The content-based concepts aircraft and vegetation are unlikely to perform well using style analysis. We added these concepts last to prevent disturbance of the style vector with noisy detection results. Based on training set experiments, we found that aircraft is the most difficult to detect. Hence, this concept was chosen as the last one. Vegetation was chosen as fifth. We chose sporting event as the fourth one, because we used a limited set of specific detectors for this semantic concept. 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. For detection of a nonstudio setting, both weather news and news subject monologues are useful. Hence, we chose nonstudio setting as third. We found that the order of weather news and news subject monologues is not important in terms of performance. We chose to detect news subject monologues first. We added all concepts iteratively to the context. We assigned a semantic concept to a style vector, or not, based on the probability $p(\omega|\vec{s}_i)$ for each individual style model where we used a threshold of 0.5 on $p(\omega|\vec{s}_i)$.

6. RESULTS

6.1 Evaluation Criteria

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

$$\textit{precision at } 100 = \frac{1}{100} \sum_{k=1}^{100} \lambda(l_k) \ , \tag{2}$$

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 is proportional to the area under a recall-precision curve. This value is the average of the precision over all relevant shots. Hence, it combines precision and recall into one performance value. 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:

$$\textit{average precision} = \frac{1}{R} \sum_{k=1}^{A} \frac{R \cap L^k}{k} \lambda(l_k) \ . \tag{3}$$

The average precision is the basic metric to evaluate experiments within the TRECVID benchmark. TRECVID relies on a pooled ground truth P to reduce labor-intensive manual judgments of all

Table I. New Pooling and Judging Statistics (for six semantic concepts from the 2003 TRECVID benchmark after our evaluation)

	News Subject Monologue	Nonstudio Setting	Sporting Event	Weather News	Vegetation	Aircraft
Pooled depth	100	350	150	100	150	150
Unlabelled	28	324	66	0	116	140
Judged true	28	304	30	0	25	4
Original true	266	2429	585	166	1095	258
New true	294	2733	615	166	1120	262

submitted runs. From each submitted run a fixed number of ranked shots is taken which is combined into a list of unique shots. Every submission is then evaluated based on the results of assessing this merged subset, that is, instead of using R in (3), 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. This is because 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 be recalculated using G for both the original submitted runs and new runs.

6.2 Benchmark Comparison

We compared the concept detection results of our style analysis framework with nine present-day systems participating in TRECVID 2003. A total of 188 system settings were submitted to TRECVID 2003 for the six semantic concepts considered. 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 [Smeaton et al. 2003]. We used the procedure explained in Section 6.1. Results of this evaluation are summarized in Table I. We obtained the final submissions from each participating group. 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 Figure 3.

The style analysis framework works particularly well for news subject monologues, improving upon the other approaches by more than a factor ten (see also Snoek et al. [2004]). This clearly demonstrates the potential of style analysis for detection of rich semantic concepts. For the nonstudio setting, our method is slightly better than the other systems. Although other systems obtain a twice as good average precision performance for the 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. The average precision results for weather news are comparable to approaches among the best. As expected, the style analysis framework fails for the content-based concepts vegetation and aircraft. This indicates that the current set of content detectors is not complete enough to cover a wide range of content-based concepts. However, for the four rich semantic concepts, the benchmark results show that the style analysis framework allows for generic indexing with performance comparable to the state-of-the-art.

6.3 Influence of Production Phases on Detection of Rich Semantic Concepts

To gain insight into the importance of production phases for news video indexing, we trained classifiers for the four rich semantic concepts using style analysis. In addition, we trained classifiers for each

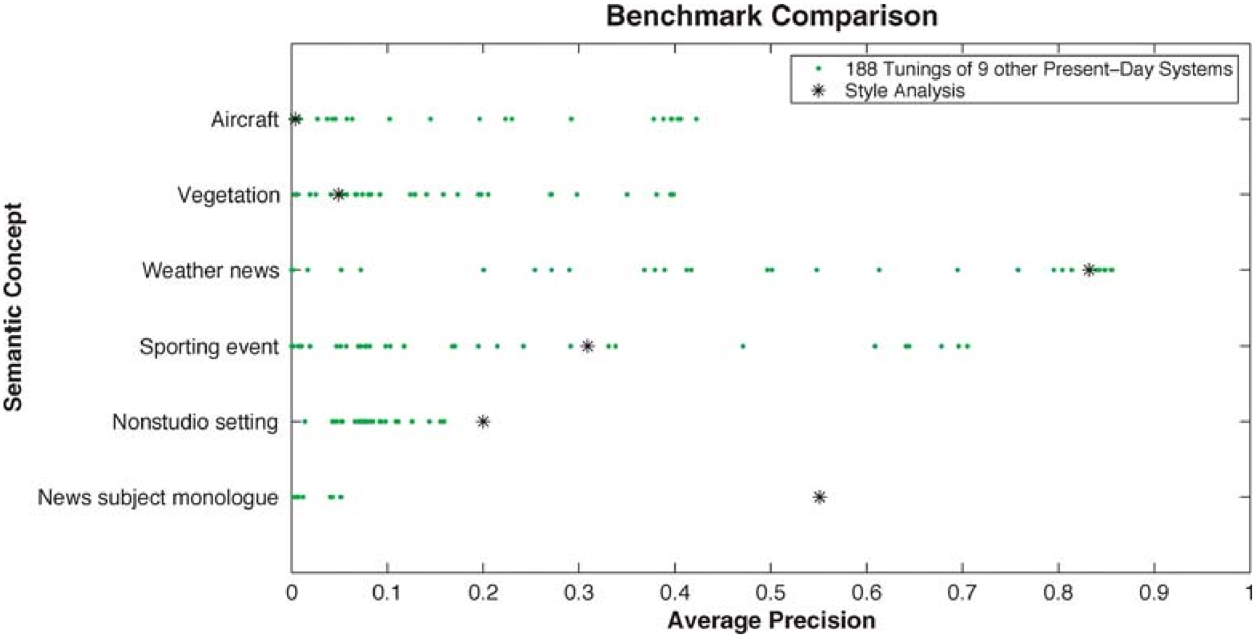

Fig. 3. Comparison of style analysis results for six semantic concepts from the 2003 TRECVID benchmark with nine other present-day indexing systems.

rich semantic concept using style detectors for the four production phases in isolation. This resulted for each of the four rich semantic concepts in a classifier based on layout detectors, content detectors, capture detectors, context detectors, or style analysis. For all classifier combinations, we evaluated the precision at 100. In Figure 4 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 analyzed production phases that yields the best results. The precision at 100 scores are, respectively, 0.94 for news subject monologues, 0.98 for the nonstudio setting, 0.85 for the sporting event, and 0.99 for weather news. As expected, content detectors are 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 in 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 good precision at 100 accuracy, but here, a combined style analysis also yields the best result. Besides content detectors, capture detectors are important for the effective detection of rich semantics. The graphs for the news subject monologue, the nonstudio setting, and to a lesser extent, sporting events, support this observation. For weather news, the current set of capture detectors plays no role of importance. Layout is somewhat useful in isolation when aiming for detection of news subject monologues and the nonstudio setting. For sporting event and weather news, layout is less useful. The current set of context detectors is too limited. Except for the nonstudio setting, usage of context in isolation is not sufficient to classify rich semantics. The results support our claim that for generic and robust rich semantic indexing of news video archives, analysis should exploit all production phases.

6.4 Discussion

When we take a closer look at the results, we conclude that the news subject monologue detector achieves a high accuracy. In part, the individual detectors will contribute to this. The influence of the entire production process, however, is evident also. Production recording, modeled by capture detectors, is a very informative production phase for news subject monologues. Typically, an author records news subject monologues in closeup with a static camera. Some concepts, such as anchors and reporters, share similarities with news subject monologues in their production design, that is, content detector results, and production recording, that is, capture detector results. Context is therefore required to reduce the number of false positives. By adding layout detectors, specifically the presence of overlayed text, we improve the results even further. Another interesting observation stems from analyzing the top 100 results for each production phase. Shots from ABC dominate results for production design, that is, content detector results, and postproduction, that is, layout detector results. In contrast, shots from CNN dominate the top 100 lists for production recording, that is, capture detector results, and preproduction, that is, context detector results. This suggests that the authors of both news broadcasts stress different production phases for the creation of news subject monologues.

We can classify weather news accurately by using only content detectors. 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 relatively easy based on textual content only. It is, however, again the combination of production phases 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 production process between the various weather reports. It is therefore not surprising that shots from CNN dominate detection results. Weather news in ABC is much harder to detect.

The nonstudio setting is relatively easy to detect in both ABC and CNN news with all approaches. The fact that this class of concepts is rather large accounts for this behavior. Detection is possible with all four style detectors in isolation, even by context alone. This can be explained because the nonstudio 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 nonstudio 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 the absence of frontal faces. The type of camera work used for production recording also aids in correct classification of sporting events. Context detectors and layout detectors are not useful in isolation for the detection of sporting events. Similar to weather news, CNN broadcasts sporting events in a separate report. This production similarity makes detection of sporting events easier for CNN than for ABC.

In terms of precision, content detectors are the most dominant style detectors for all four rich semantic concepts. This is visible in Figure 4, where content detectors parallel style analysis for the highest ranked results. This makes analysis based on content detectors useful for applications that require only a limited set of correct results. However, when a large set of correct results is required, we should use all style detectors. As can be observed from Figure 4, it is the combination of production phases that strengthens recall and overall performance.

In contrast to the previously described benefits, one obvious criticism of the framework is that many of the individual style detectors require an additional development effort when compared to methods using only low-level content features such as color, texture, and motion. However, for rich semantic concepts it comes with the reward of increased performance. As the TRECVID comparison shows, the performance can be as much as ten times better than standard approaches using low-level content features only. When performance matters, the additional style detector development effort pays off.

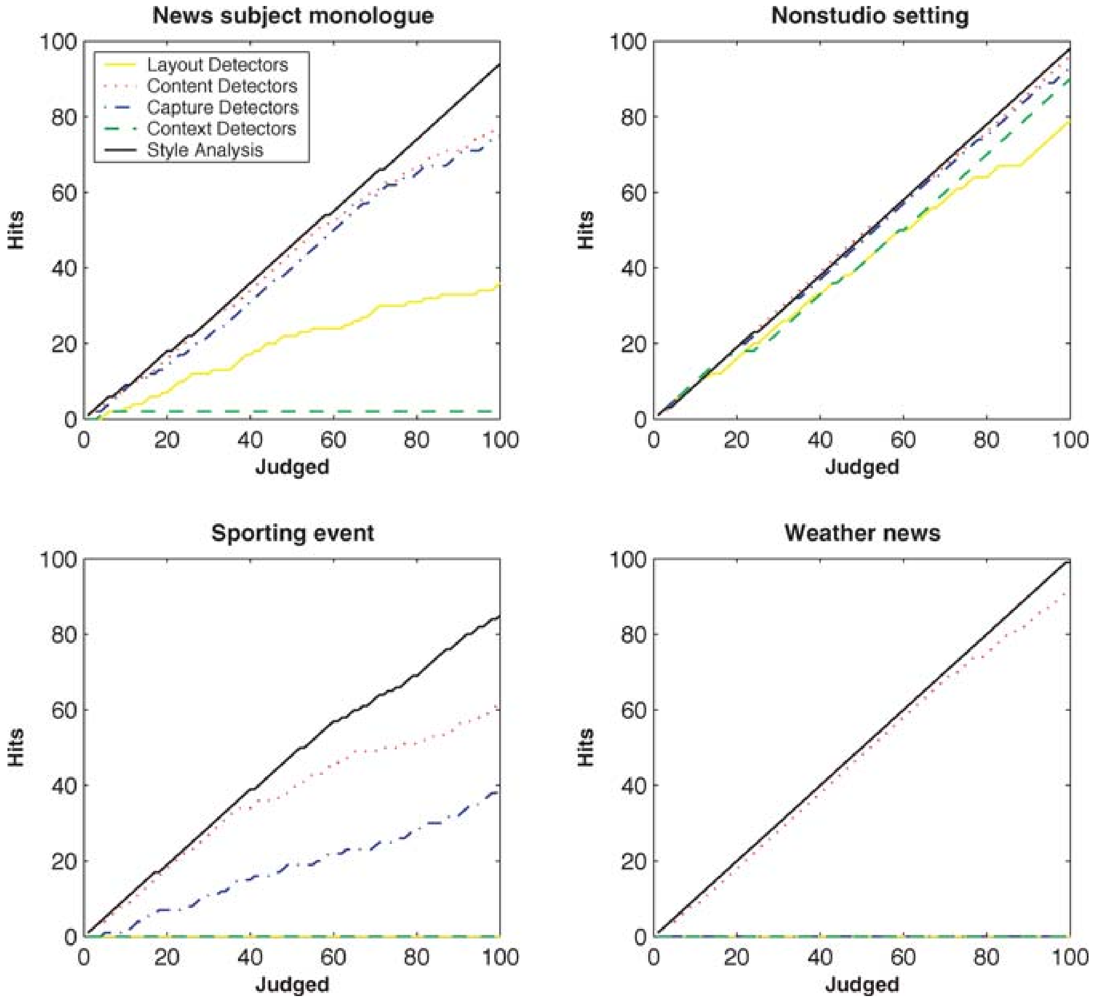

Fig. 4. Influence of production phases, modeled as style detectors, on rich semantic concept detection performance. The number of hits are plotted as a function of the number of camera shots judged, for the first 100 results of each analyzed rich semantic concept.

Other rich semantic concepts for which the proposed framework is likely to perform well are specific instances of the ones already discussed. We can split the news subject monologue, for example, into public speech, press conference, and interview. In addition, the detection of political reporters and financial reporters should also be possible. Semantic concepts that only have similarity in content, for example, boat, train, and building, are much harder to detect since our framework works particularly well if similarity in more than one production phase exists. We visualized the top 25 results for each rich semantic concept using the style analysis framework in Figure 5.

7. CONCLUSION

In this article, we propose to model the entire broadcast news production process for the purpose of automatically indexing news video archives. We present the style analysis framework, a generic and

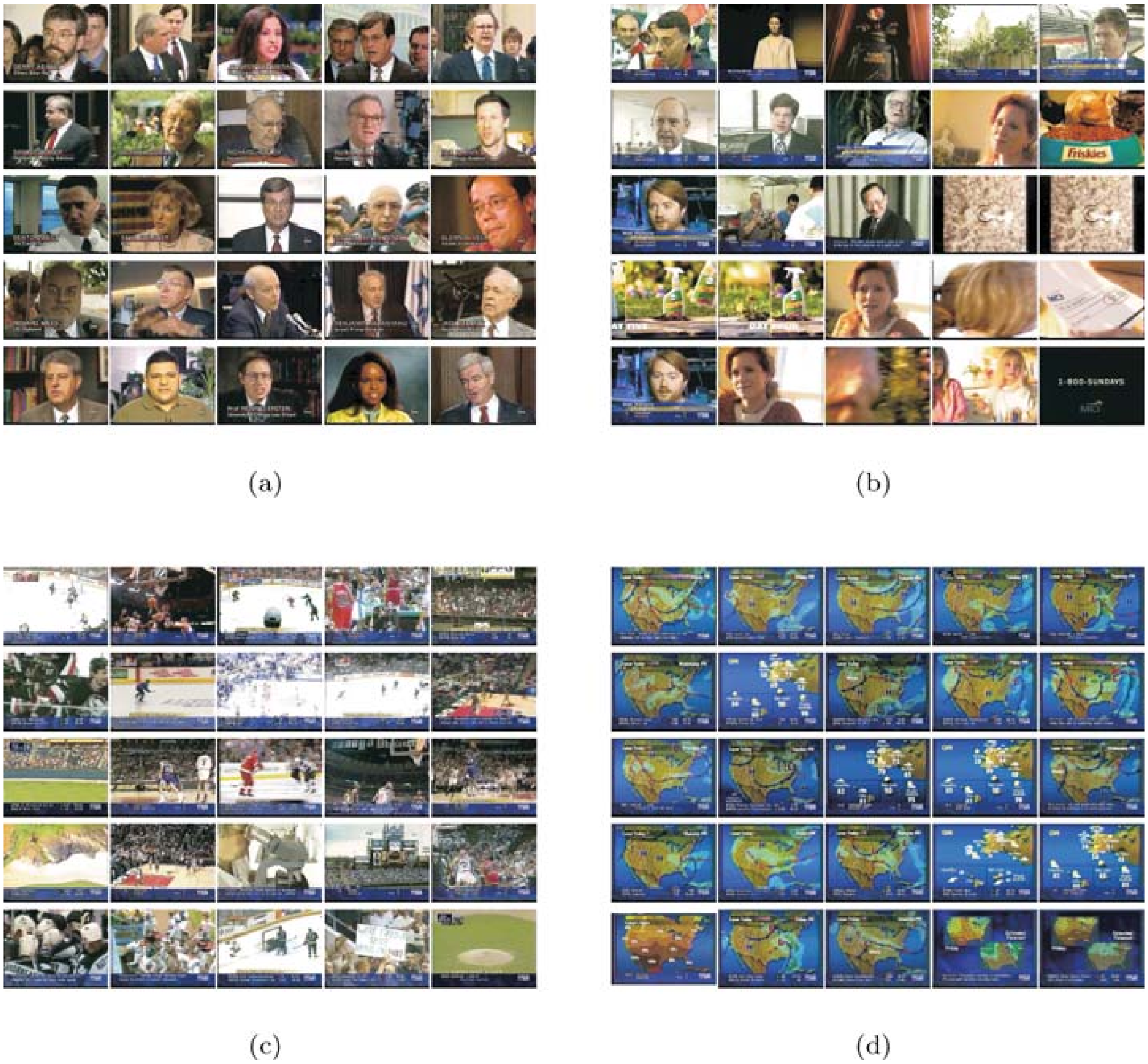

Fig. 5. Top 25 results for (a) news subject monologue, (b) nonstudio setting, (c) sporting event, and (d) weather news using style analysis. Key frames are ordered left to right, top to bottom.

robust framework for news video indexing. Within the broadcast news production process, we identify four production phases. For each production phase we extract multimodal production-derived features using style detectors, that is, layout detectors, content detectors, capture detectors, and context detectors. To combine style detector results and learn the rich semantics, the framework utilizes a classifier combination scheme. This machine learning scheme facilitates enrichment of semantics by iteratively updating the set of context detectors. By combining the style detectors in an iterative classifier combination scheme, the framework allows for rich semantic indexing.

An experiment on an archive of 120 hours of news video, in which we detect several rich semantic concepts, demonstrates that style analysis allows for the generic indexing of rich semantic concepts in news video. For all concepts analyzed, the set of content detectors is the most decisive. In terms of performance, content detectors are specifically useful when aiming for good precision on a limited set of retrieved items. However, a combined style analysis, which models the entire broadcast news production process, yields the best overall performance for both precision and recall. This makes our framework a good candidate when aiming for a 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. As expected, the framework is unsuited for concepts that rely on content only, such as vegetation and aircraft. However, for concepts that share many similarities in their production process, style analysis pays off. The results show that the proposed framework obtains an accuracy favorable for detecting news subject monologues, the nonstudio setting, and weather news, and only lags behind dedicated sporting event detection algorithms. We consider this another strong indicator of the approach.

At present, the style analysis framework is limited to the news domain. However, other video domains also adhere to a phased production process. Obvious examples are feature films, situational comedies, and documentaries. It is therefore likely that these produced video domains will profit from style analysis also.

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 *mise-en-scène*. We are convinced that their impact on the detection of rich semantic concepts in video will boost progress in multimedia analysis.

ELECTRONIC APPENDIX

The electronic appendix for this article can be accessed in the ACM Digital Library.

ACKNOWLEDGMENTS

The authors are grateful to Ming-yu Chen of Carnegie Mellon University for providing commercial and anchor detection results. In addition, they acknowledge the benchmark organization and evaluation effort by NIST and the TRECVID coordinators. Finally, they thank the anonymous reviewers for their valuable comments and suggestions.

REFERENCES

ADAMS, B., DORAI, C., AND VENKATESH, S. 2002. Toward automatic extraction of expressive elements from motion pictures: Tempo. *IEEE Trans. Multimedia 4*, 4, 472–481.

ADAMS, W. H., IYENGAR, G., LIN, C.-Y., NAPHADE, M., NETI, C., NOCK, H., AND SMITH, J. 2003. Semantic indexing of multimedia content using visual, audio, and text cues. *EURASIP J. Appl. Signal Process. 2003*, 2, 170–185.

AMIR, A., BERG, M., CHANG, S.-F., HSU, W., IYENGAR, G., LIN, C.-Y., NAPHADE, M., NATSEV, A., NETI, C., NOCK, H., SMITH, J., TSENG, B., WU, Y., AND ZHANG, D. 2003. IBM research TRECVID-2003 video retrieval system. In *Proceedings of the TRECVID Workshop*. NIST Special Publication. Gaithersburg, Md.

BAAN, J., BALLEGOOIJ, A., GEUSEBROEK, J., HIEMSTRA, D., DEN HARTOG, J., LIST, J., SNOEK, C., PATRAS, I., RAAIJMAKERS, S., TODORAN, L., VENDRIG, J., DE VRIES, A., WESTERVELD, T., AND WORRING, M. 2001. Lazy users and automatic video retrieval tools in (the) lowlands. In *Proceedings of the 10th Text REtrieval Conference*, E. Voorhees and D. Harman, eds. NIST Special Publication, vol. 500-250. Gaithersburg, Md.

BARRY, B. AND DAVENPORT, G. 2003. Documenting life: Videography and common sense. In *Proceedings of the IEEE International Conference on Multimedia & Expo*. Baltimore, Md.

BOGGS, J. AND PETRIE, D. 2000. *The Art of Watching Films*, 5th ed. Mayfield Publishing Mountain View, Calif.

BORDWELL, D. AND THOMPSON, K. 1997. *Film Art: An Introduction*, 5th ed. McGraw-Hill, New York.

BREIMAN, L. 1996. Bagging predictors. *Mach. Learn. 24*, 2, 123–140.

CHANG, C.-C. AND LIN, C.-J. 2001. *LIBSVM: A Library for Support Vector Machines*. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm/.

DAVIS, M. 2003. Editing out video editing. *IEEE Multimedia 10*, 2, 54–64.

DEY, A. 2001. Understanding and using context. *Personal Ubiquitous Comput. Journal 5*, 1, 4–7.

DOURISH, P. 2004. What we talk about when we talk about context. *Personal Ubiquitous Comput. 8*, 1, 19–30.

GAUVAIN, J., LAMEL, L., AND ADDA, G. 2002. The LIMSI broadcast news transcription system. *Speech Commun. 37*, 1–2, 89–108.

HANJALIC, A., LAGENDIJK, R., AND BIEMOND, J. 1999. Automated high-level movie segmentation for advanced video-retrieval systems. *IEEE Trans. Circuits Syst. Video Technol. 9*, 4, 580–588.

HAUPTMANN, A., BARON, R., CHEN, M.-Y., CHRISTEL, M., DUYGULU, P., HUANG, C., JIN, R., LIN, W.-H., NG, T., MORAVEJI, N., PAPERNICK, N., SNOEK, C., TZANETAKIS, G., YANG, J., YAN, R., AND WACTLAR, H. 2003. Informedia at TRECVID 2003: Analyzing and searching broadcast news video. In *Proceedings of the TRECVID Workshop*. NIST Special Publication. Gaithersburg, Md.

JAIN, A., DUIN, R., AND MAO, J. 2000. Statistical pattern recognition: A review. *IEEE Trans. Pattern Anal. Mach. Intel. 22*, 1, 4–37.

NACK, F. AND PUTZ, W. 2004. Saying what it means: Semi-Automated (news) media annotation. *Multimedia Tools Appl. 22*, 3, 263–302.

NAPHADE, M., KOZINTSEV, I., AND HUANG, T. 2002. A factor graph framework for semantic video indexing. *IEEE Trans. Circuits Syst. Video Technol. 12*, 1, 40–52.

PLATT, J. 2000. Probabilities for SV machines. In *Advances in Large Margin Classifiers*, A. Smola, et al., eds. MIT Press, Cambridge, Mass., 61–74.

QUÉNOT, G., MORARU, D., BESACIER, L., AND MULHEM, P. 2002. CLIPS at TREC-11: Experiments in video retrieval. In *Proceedings of the 11th Text REtrieval Conference*, E. Voorhees and L. Buckland, eds. NIST Special Publication, vol. 500-251. Gaithersburg, Md.

SALTON, G. AND MCGILL, M. 1983. *Introduction to Modern Information Retrieval*. McGraw-Hill, New York.

SATO, T., KANADE, T., HUGHES, E., SMITH, M., AND SATOH, S. 1999. Video OCR: Indexing digital news libraries by recognition of superimposed caption. *Multimedia Syst. 7*, 5, 385–395.

SCHAPIRE, R. 1990. The strength of weak learnability. *Mach. Learn. 5*, 2, 197–227.

SCHNEIDERMAN, H. AND KANADE, T. 2004. Object detection using the statistics of parts. *Int. J. Comput. Vision 56*, 3, 151–177.

SMEATON, A., KRAAIJ, W., AND OVER, P. 2003. TRECVID 2003—An introduction. In *Proceedings of the TRECVID Workshop*. NIST Special Publication. Gaithersburg, Md.

SMEULDERS, A., WORRING, M., SANTINI, S., GUPTA, A., AND JAIN, R. 2000. Content based image retrieval at the end of the early years. *IEEE Trans. Pattern Anal. Mach. Intell. 22*, 12, 1349–1380.

SMITH, J. AND LI, C.-S. 1998. Decoding image semantics using composite region templates. In *Proceedings of the IEEE CVPR-98 Workshop on Content-Based Access to Image, Video Databases*. Santa Barbara, Calif.

SNOEK, C. AND WORRING, M. 2005a. Multimedia event-based video indexing using time intervals. *IEEE Trans. Multimedia 7*, 4, 638–647.

SNOEK, C. AND WORRING, M. 2005b. Multimodal video indexing: A review of the state-of-the-art. *Multimedia Tools Appl. 25*, 1, 5–35.

SNOEK, C., WORRING, M., AND HAUPTMANN, A. 2004. Detection of TV news monologues by style analysis. In *Proceedings of the IEEE International Conference on Multimedia & Expo*. Taipei, Taiwan.

SUNDARAM, H. AND CHANG, S.-F. 2002. Computable scenes and structures in films. *IEEE Trans. Multimedia 4*, 4, 482–491.

TRUONG, B., VENKATESH, S., AND DORAI, C. 2005. Extraction of film takes for cinematic analysis. *Multimedia Tools Appl. 26*, 3, 277–298.

TSENG, B., LIN, C.-Y., NAPHADE, M., NATSEV, A., AND SMITH, J. 2003. Normalized classifier fusion for semantic visual concept detection. In *Proceedings of the IEEE International Conference on Image Processing*. vol. 2. Barcelona, Spain, 535–538.

VAPNIK, V. 2000. *The Nature of Statistical Learning Theory*, 2nd ed. Springer Verlag, New York.

VENDRIG, J. AND WORRING, M. 2002. Systematic evaluation of logical story unit segmentation. *IEEE Trans. Multimedia 4*, 4, 492–499.

WACTLAR, H., CHRISTEL, M., GONG, Y., AND HAUPTMANN, A. 1999. Lessons learned from building a terabyte digital video library. *IEEE Computer 32*, 2, 66–73.

WOLPERT, D. 1992. Stacked generalization. *Neural Netw. 5*, 241–259.

XIE, L., XU, P., CHANG, S.-F., DIVAKARAN, A., AND SUN, H. 2004. Structure analysis of soccer video with domain knowledge and hidden Markov models. *Pattern Recogn. Lett. 25*, 7, 767–775.

ZHANG, H.-J., WU, J., ZHONG, D., AND SMOLIAR, S. 1997. An integrated system for content-based video retrieval and browsing. *Pattern Recogn. 30*, 4, 643–658.

Received April 2005; revised December 2005; accepted December 2005

Online Appendix to:
Learning Rich Semantics from News Video Archives by Style Analysis

CEES G. M. SNOEK and MARCEL WORRING
University of Amsterdam
and
ALEXANDER G. HAUPTMANN
Carnegie Mellon University

A. INTRODUCTION

In this appendix we discuss the implementation of the various style detectors as presented in Snoek et al. [2006]. Each style detector uses an existing software implementation as a basis. The output of such a base detector is then aggregated and synchronized to a camera shot. We categorize the resulting production-derived features based on experimentally obtained thresholds. Together, these three components define a style detector. All style detectors follow the basic architecture as presented in Figure 6.

B. LAYOUT DETECTORS

B.1 Shot Length

An author uses variation in the shot length to affect the overall rhythm of a produced video [Adams et al. 2002]. We determine shot length based on a camera shot segmentation obtained from a camera shot detector [Quénot et al. 2002]. 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, *medium take* if it contains 70 to 300 frames, and *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 displayed in Figure 7.

B.2 Overlayed Text

Overlayed text is added by the author at production time to provide the viewer with additional descriptive information, for example, the annotation of people in broadcast news. Its presence is an important indicator for layout style. We apply a video optical character recognition system [Sato et al. 1999] to localize and extract overlayed text in a video frame[2]. As 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 five characters, otherwise, we consider it absent. The overlayed text detector scheme is portrayed in Figure 8.

[2] For CNN the ticker tape with stock information was ignored.

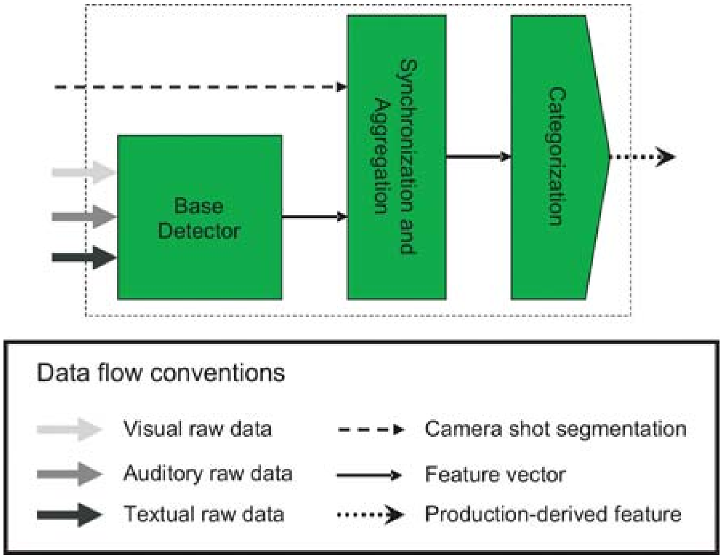

Fig. 6. Basic architecture and data flow within a style detector.

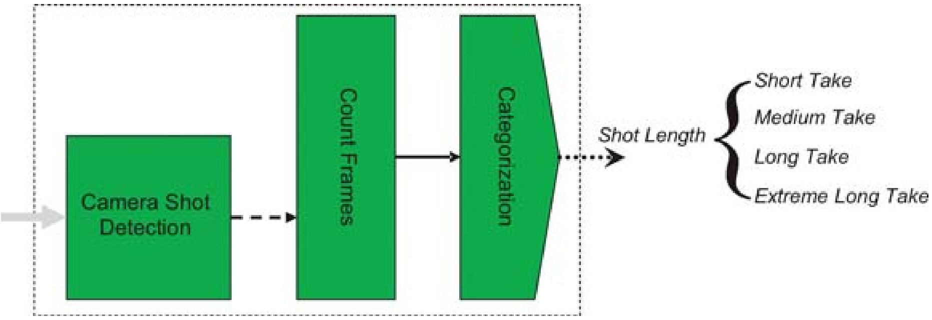

Fig. 7. A shot length detector using the conventions of Figure 6.

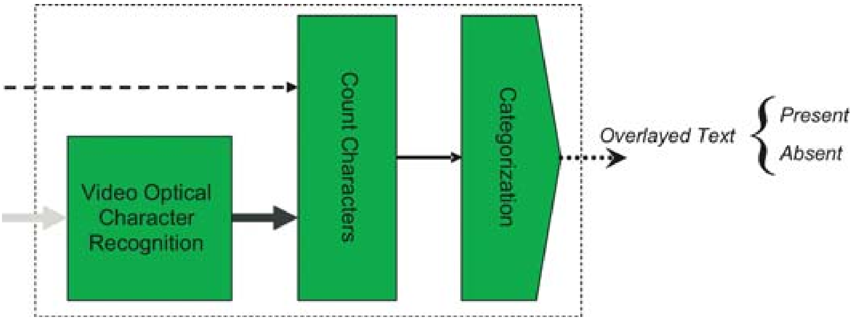

Fig. 8. An overlayed text detector.

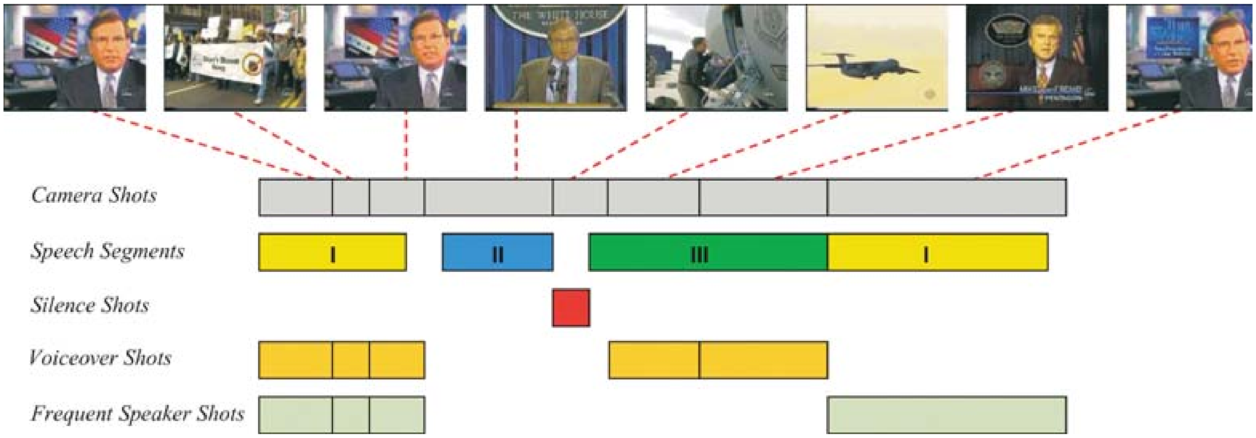

Fig. 9. Segmentation of a news video into camera shots, speech segments with speaker identifier, silence shots, voiceover camera shots, and frequent speaker camera shots.

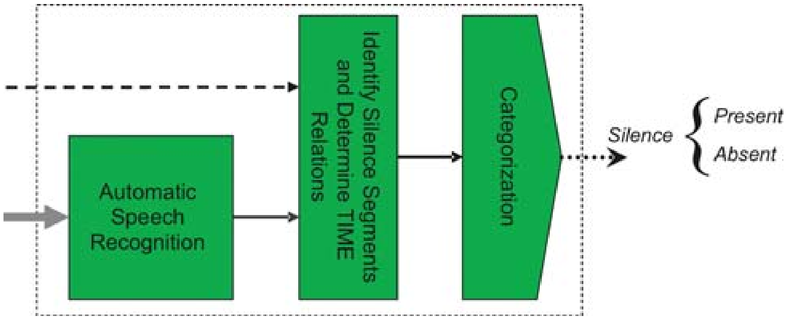
Automatic Speech Recognition
Identify Silence Segments and Determine TIME Relations
Categorization
Silence
Present
Absent

Fig. 10. A silence detector.

B.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 [Gauvain et al. 2002]. 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 a camera shot level. For this purpose we exploit the TIME relations proposed in Snoek and Worring [2005]. We ignore the *NoRelation*, *precedes* and *precedes_i* relations, as these are interesting for temporal clues only. The shot segmentation is considered the reference segmentation. We use a value of ten 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 Figure 9 for an example). The silence detector scheme is shown in Figure 10.

B.4 Voiceover

An author uses a voiceover when the content of the news video is not self-descriptive and requires additional information, for example, in sport news. Voiceover detection is also based on the automatic

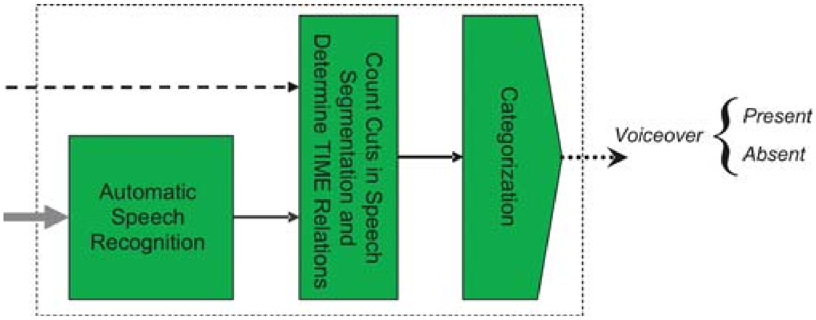
Automatic Speech Recognition
Count Cuts in Speech Segmentation and Determine TIME Relations
Categorization
Voiceover
Present
Absent

Fig. 11. A voiceover detector.

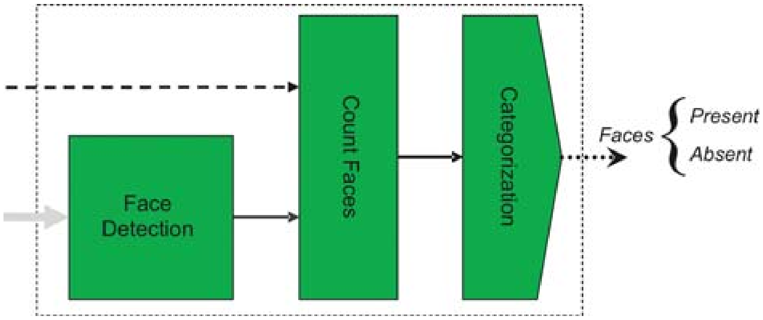

Fig. 12. A faces detector.

speech recognition results from Gauvain et al. [2002]. 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 voiceover segment when it contains more than one cut (this is illustrated in Figure 9). To map the voiceover segments to camera shots, we use the same TIME relations as before. However, for T_1 we now use a value of 25 frames. If a TIME relation between a camera shot and a voiceover segment exists, we consider a voiceover present in the shot, else absent. The voiceover detector scheme is presented in Figure 11.

C. CONTENT DETECTORS

C.1 Faces

Human beings are a prominent content element in news video. To detect the presence of people we apply the face detector of Schneiderman and Kanade [2004]. 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 illustrated in Figure 12.

C.2 Face Location

Since people are important in news video an author takes great care in filming them, that is, 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

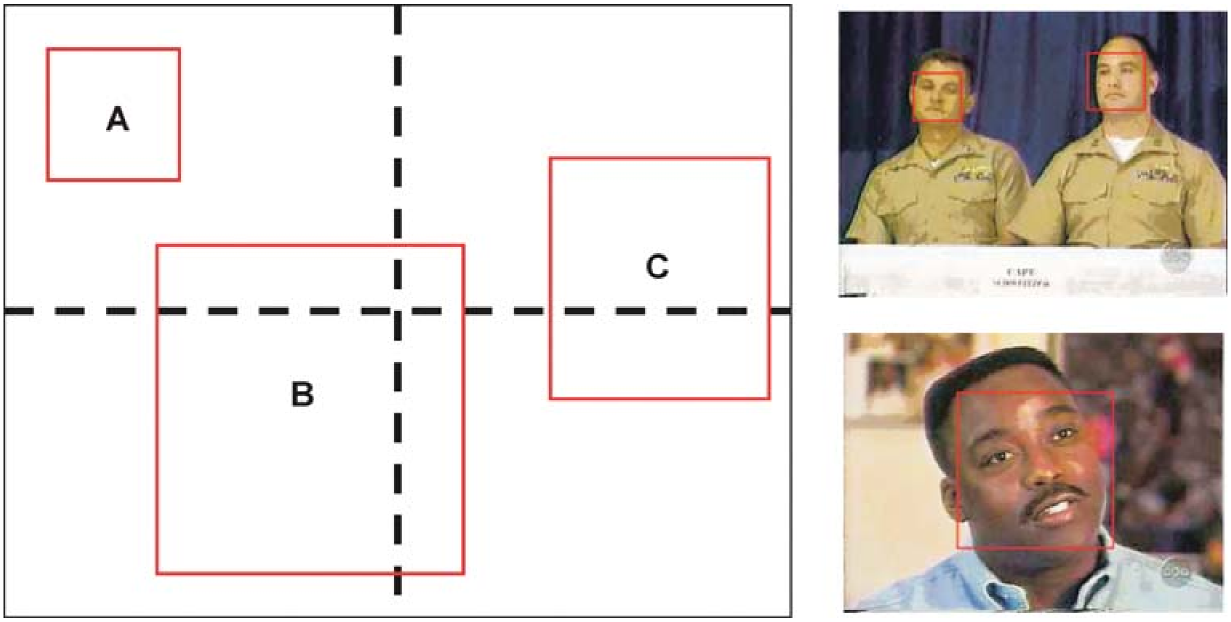

Fig. 13. Left: An image frame with three detected faces, face A is located topleft, face B is located center, and face C is located right. Right: Two example image frames with detected faces.

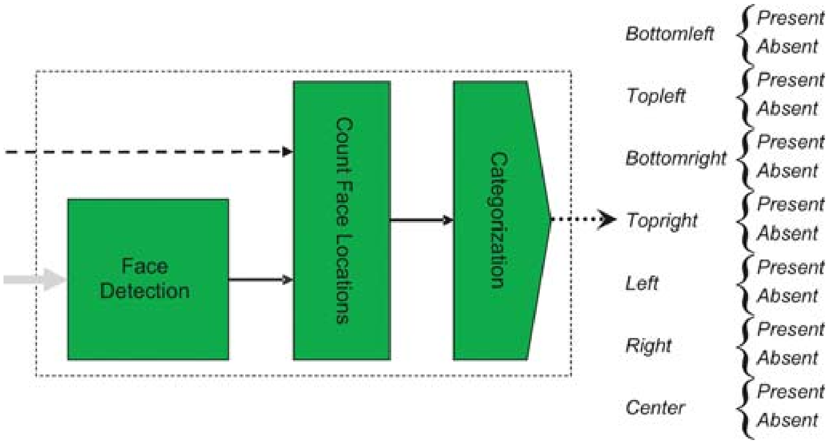

Fig. 14. A face location detector.

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, as shown in Figure 13. Note that we do not distinguish between top and bottom and that the larger the face as shown 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 frames in a camera shot. If this is the case, the feature is set as present. The face location detector scheme is portrayed in Figure 14.

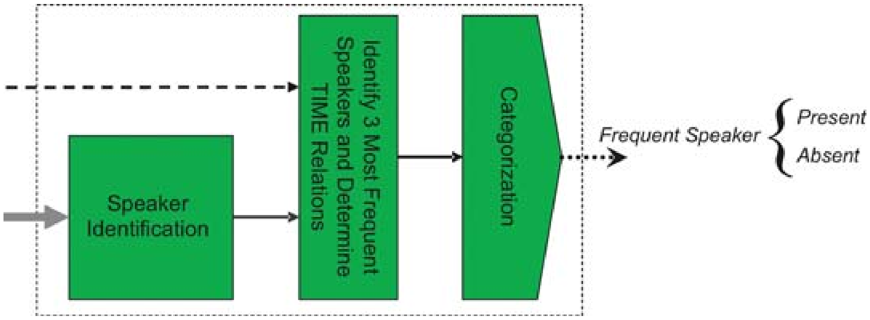

Fig. 15. A frequent speaker detector.

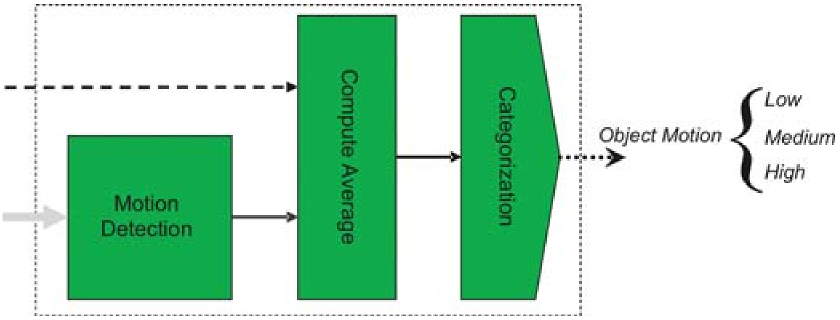

Fig. 16. An object motion detector.

C.3 Frequent Speaker

Another clue for the presence of people is their speech. We use the speaker identification results from an automatic speech recognition system [Gauvain et al. 2002]. The system provides an identifier for each recognized speaker per analyzed video. Because the identifiers are unique for a single news video only, the recognized speakers do not scale to an entire archive. Moreover, because the performance of speaker identification degrades when a large number of speakers appear in a video, we do not blindly trust the results. To accommodate for both issues we determine the three most frequent speakers per news video. Again, we refer to Figure 9 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 shown in Figure 15.

C.4 Object Motion

Specific object detectors help when you know what to look for. If not, the 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 [Joly and Kim 1996]. 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 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 displayed in Figure 16.

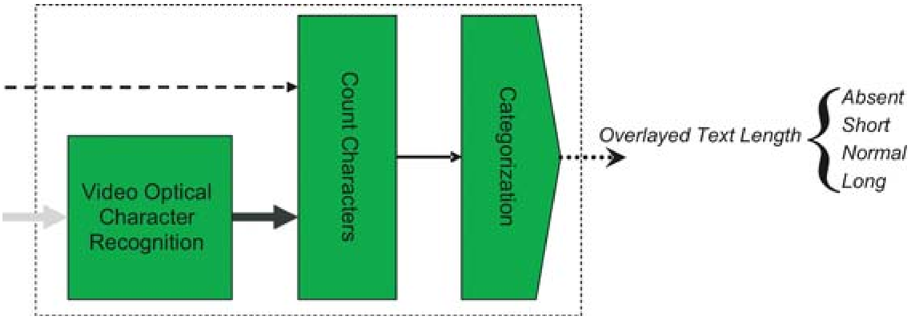

Fig. 17. An overlayed text length detector.

C.5 Overlayed Text Length

Whereas the 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 [Sato et al. 1999] 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, and in all other cases it is classified as long. The overlayed text length detector scheme is depicted in Figure 17.

C.6 Video Text Named Entity

Besides the length of recognized overlayed text strings, it is interesting to know the type of annotation, for example, 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 Bikel et al. [1999], which is described in Yan et al. [2004]. 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 the 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 the presence of named entities. This results in four features which are initially set to absent. If one of the strings in the shot contains one of the four named entities, the respective feature is set to present for the shot. The video text named entity detector scheme is pictured in Figure 18.

C.7 Positive and Negative Keywords

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 [Salton and McGill 1983]. Given a training 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

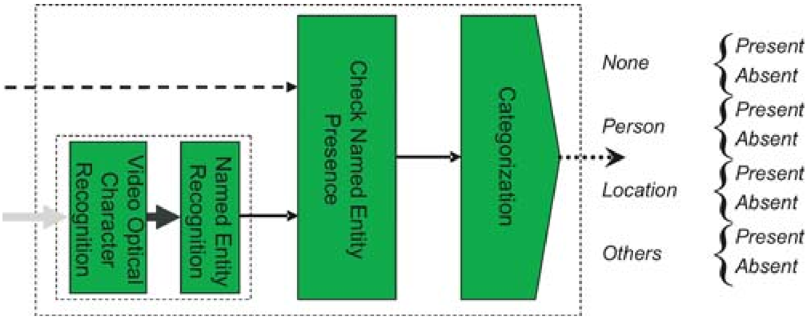

Fig. 18. A video text named entity detector.

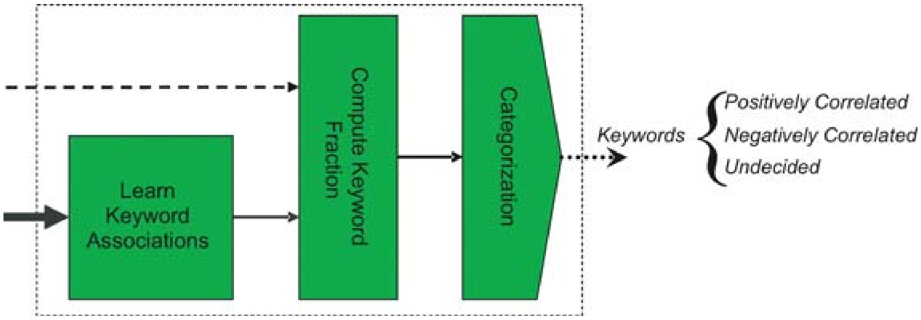

Fig. 19. Positive and negative keyword detection.

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 shown in Figure 19.

D. CAPTURE DETECTORS

D.1 Camera Distance

As an estimate for the camera distance we use a frame/face ratio proposed in Snoek [2000]. 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 [Schneiderman and Kanade 2004]. Based on the frame/face ratio we distinguish seven camera distance features: *extreme long shot, long shot, medium long shot, medium shot, medium closeup, closeup,* and *extreme closeup*. For every detected face we determine the camera distance. We aggregate all camera distances per analyzed frame to obtain a decision at shot level. 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 illustrated in Figure 20.

D.2 Camera Work

Several kinds of camera work exist, each creating its own specific effect. For computation of camera work we use an algorithm based on the one reported in Joly and Kim [1996] and Tonomura et al. [1994]. Within a camera shot, the algorithm classifies all frames as belonging to one of six types of camera

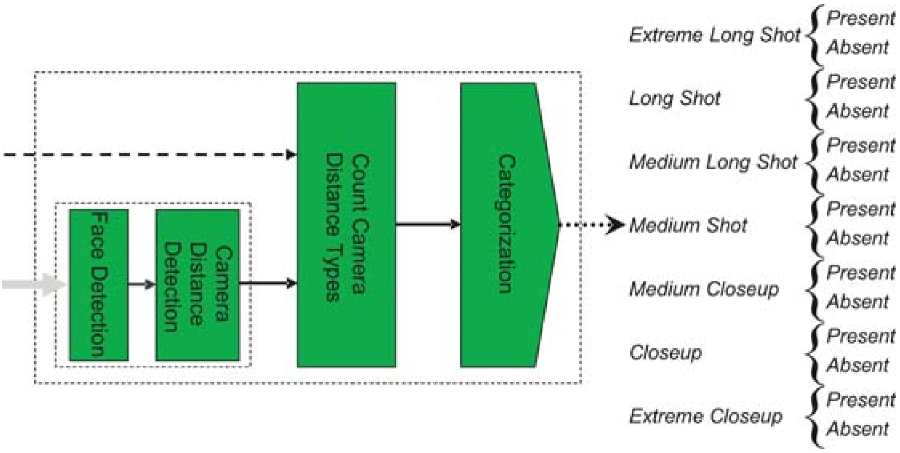

Fig. 20. A camera distance detector.

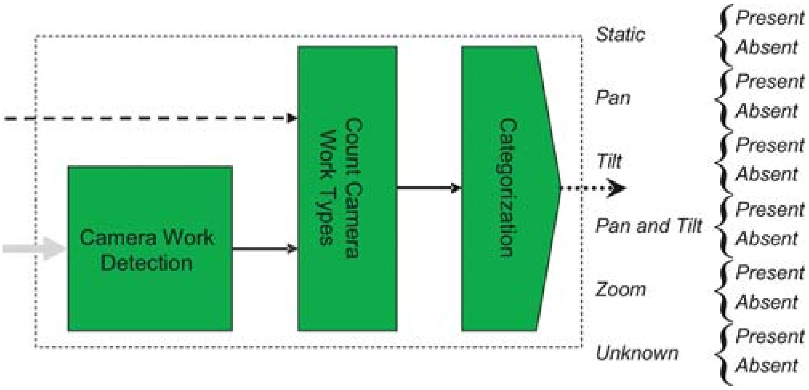

Fig. 21. A camera work detector.

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 work detector scheme is portrayed in Figure 21.

D.3 Camera Motion

Besides the detection of type of camera work, the aformentioned 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 displayed in Figure 22.

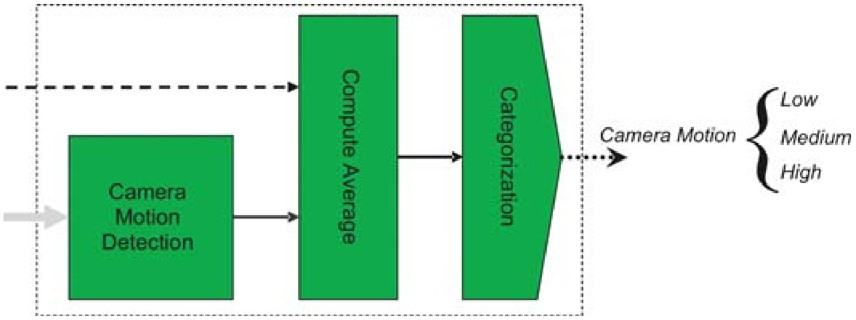

Fig. 22. A camera motion detector.

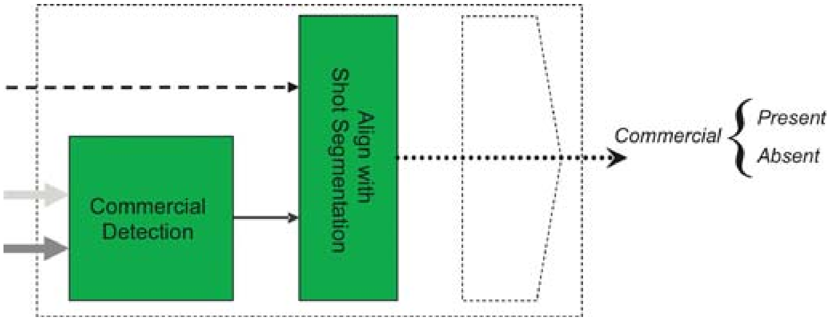

Fig. 23. A commercial detector.

E. CONTEXT DETECTORS

E.1 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 people are not interested in them. We detect commercials to prevent a false classification of news related concepts in commercials. We apply the commercial detector proposed in Hauptmann et al. [2003]. It labels key frames as either a commercial or not. The output is aligned with the camera shot segmentation. The commercial detector scheme is portrayed in Figure 23.

E.2 News Anchor

In broadcast news an author adds a news anchor to summarize the news and to connect news stories. In general, the visual content of shots containing anchors is of limited interest. However, because anchors speak on a large number of topics their textual content may trigger a lot of false positive classifications of other concepts. We apply a news anchor detector [Hauptmann et al. 2003] to prevent this misclassification. It labels key frames as either a news anchor or not. The output is aligned with the camera shot segmentation. The news anchor detector scheme is illustrated in Figure 24.

E.3 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 a news reporter detector. It compares each recognized

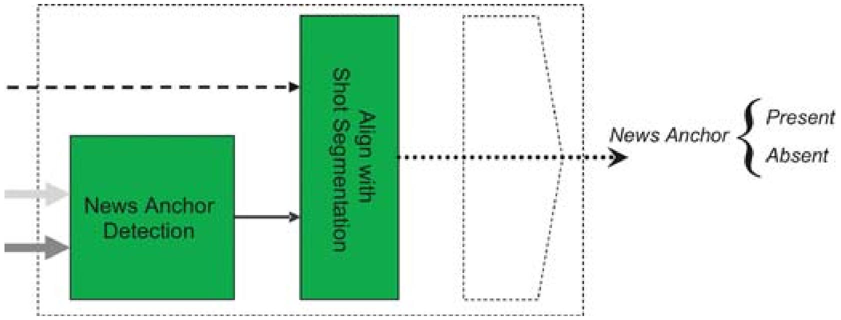

Fig. 24. A news anchor detector.

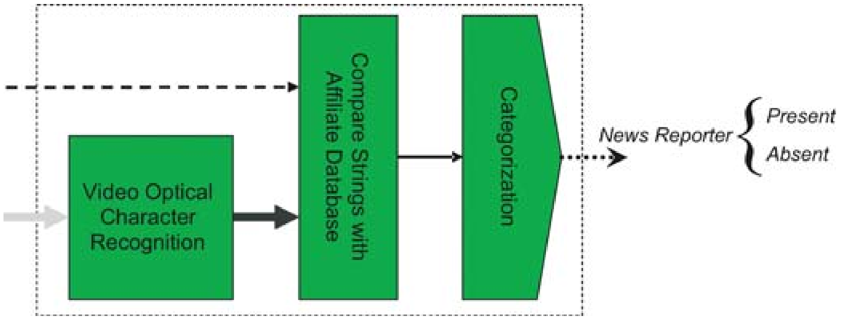
Video Optical Character Recognition
Compare Strings with Affiliate Database
Categorization
News Reporter
Present
Absent

Fig. 25. A news reporter detector.

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 shown in Figure 25.

REFERENCES

Adams, B., Dorai, C., and Venkatesh, S. 2002. Toward automatic extraction of expressive elements from motion pictures: Tempo. *IEEE Trans. Multimedia 4*, 4, 472–481.

Bikel, D., Schwartz, R., and Weischedel, R. 1999. An algorithm that learns what's in a name. *Mach. Learn. 34*, 1–3, 211–231.

Gauvain, J., Lamel, L., and Adda, G. 2002. The LIMSI broadcast news transcription system. *Speech Commun. 37*, 1–2, 89–108.

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

Joly, P. and Kim, H.-K. 1996. Efficient automatic analysis of camera work and microsegmentation of video using spatiotemporal images. *Signal Process. Image Commun. 8*, 4, 295–307.

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

Salton, G. and McGill, M. 1983. *Introduction to Modern Information Retrieval*. McGraw-Hill, New York.

Sato, T., Kanade, T., Hughes, E., Smith, M., and Satoh, S. 1999. Video OCR: Indexing digital news libraries by recognition of superimposed caption. *Multimedia Syst. 7*, 5, 385–395.

Schneiderman, H. and Kanade, T. 2004. Object detection using the statistics of parts. *Intl. J. Comput. Vision 56*, 3, 151–177.

Snoek, C. 2000. Camera distance classification: Indexing video shots based on visual features. M.S. thesis, Univ. van Amsterdam.

Snoek, C. and Worring, M. 2005. Multimedia event-based video indexing using time intervals. *IEEE Trans. Multimedia 7*, 4, 638–647.

SNOEK, C., WORRING, M., AND HAUPTMANN, A. 2006. Learning rich semantics from news video archives by style analysis. *ACM Trans. Multimedia Comput. Commun. Appl. 2*, 2, 91–108.

TONOMURA, Y., AKUTSU, A., TANIGUCHI, Y., AND SUZUKI, G. 1994. Structured video computing. *IEEE Multimedia 1*, 3, 34–43.

YAN, R., YANG, J., AND HAUPTMANN, A. 2004. Learning query-class dependent weights for automatic video retrieval. In *ACM Multimedia*. New York, USA.

Received April 2005; revised December 2005; accepted December 2005