Learning to Identify TV News Monologues by Style and Context

Cees G.M. Snoek and Alexander G. Hauptmann

October 2003
CMU-CS-03-193

School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213

Abstract

We focus on the problem of learning semantics from multimedia data associated with broadcast video documents. In this paper we propose to learn semantic concepts from multimodal sources based on style and context detectors, in combination with statistical classifier ensembles. As a case study we present our method for detecting the concept of news subject monologues. This approach had the best average precision performance amongst 26 submissions in the 2003 video track of the Text Retrieval Conference benchmark. Experiments were conducted with respect to individual detector contribution, ensemble size, and ranking mechanism. It was found that the combination of detectors is decisive for the final result, although some detectors might appear useless in isolation. Moreover, by using a probabilistic ranking, in combination with a large classifier ensemble, results can be improved even further.

This research was performed while the first author was a visiting scientist at the Informedia project, Carnegie Mellon University. He was sponsored by the Dutch ICES/KIS Multimedia Information Analysis project and TNO Institute of Applied Physics (TPD). This work was also supported in part by the Advanced Research and Development Activity (ARDA) under contract number MDA904-02-C-0451 and by the National Science Foundation under Cooperative Agreement No. IRI-9817496. The first author is with the Computer Science Institute, University of Amsterdam, Kruislaan 403, 1098 SJ Amsterdam, The Netherlands (e-mail: cgmsnoek@science.uva.nl).

Keywords: multimodal analysis, semantic learning, classifier ensembles, broadcast video, news subject monologue, style detectors, context detectors, TRECVID benchmark.

1 Introduction

Advancement in optical fiber technology and growing availability of low-cost digital multimedia recording devices, enables world wide capture, delivery, and exchange of large amounts of video assets over the Internet. This overwhelming amount of digital video data becoming available, will trigger the need for automatic indexing tools that can provide on-the-fly content-based annotation, ultimately allowing for effective and efficient browsing, filtering, and retrieval of specific video segments. Unfortunately, however, the progress in content-based multimedia analysis has not kept pace with the developments in multimedia recording, storage, and transmission technologies.

Automatic techniques for video indexing suffer from the fact that it is very hard to infer content-based semantics based on the low-level features that can be extracted from visual, auditory, and textual data. In an effort to bridge this *semantic gap*, there has recently been a shift from unimodal to multimodal video analysis, i.e. combining all available information sources in the analysis. See [20] for an overview of reported methods on multimodal video analysis. The combination of clues from various modalities can narrow the semantic gap, since more information is available and errors from different sources do not necessarily correlate, and thus cancel out. However, an inconsistency between automatically extracted indexes and their interpreted semantics remains. Besides using a combination of multimodal sources, analysis should therefore exploit the inherent context that is available in multimedia data, as recognized by [12, 14], where context refers to simultaneous, not sequential, co-occurrence of various semantic concepts. The key observation to help overcome the semantic gap is that semantic concepts that appear in a video document do not occur in isolation.

To model both content and context, [12] proposes to use layered Bayesian Networks. The authors assume a hierarchical organization of multimedia semantics, based on analysis by low-level, med-level and high-level detectors. Those levels represent respectively: objects, related objects or events, and related events or stories. The detectors are represented by nodes in the network, and the arcs model the relationships. Nodes and arcs are associated with conditional probability densities. Based on the joint probability distribution of the lower levels, semantics at a higher level can be inferred. The high-level semantics are learned by using hierarchical priors that combine the content and context layers. By inclusion of a context layer the authors report improved accuracy in detection of talk shows and financial news broadcasts based on detection of various multimedia elements like faces, overlayed text, speech, and specific keywords.

Another probabilistic approach is presented in [14]. The authors propose to model semantic concepts through probabilistic detectors, for example aeroplane, skydiving, and bird detectors. Those concept detectors are referred to as multijects. To infer contextual semantics, e.g. outdoor, the multijects are integrated into a network, referred to as multinet. By combining the individual probabilities of all multijects into a multinet using factor graphs, the framework is applicable to all sorts of multimedia data and a variety of semantic indexes. However, the experiments are only applied to visual concepts that are related to the setting of the multimedia data, e.g. rocky terrain, water-body, and forestry.

The focus of previous work on contextual learning of semantics in multimedia data is generally based on a type of probabilistic framework. Drawbacks of this framework are implicit independence assumption of the individual pieces and the difficulty of finding prior probabilities. Moreover, they require a large amount of jointly labelled training data, and hence a large effort in manual annotation for ground truth. To prevent this reliance on independence assumptions and prior probability estimates, we propose to model concepts and context with a pool of discrete detectors. To infer semantics based on those detectors we propose to use statistical classifier ensembles. To evaluate our method, an experiment was carried out within the content-based video retrieval track (TRECVID) of the 2003 Text Retrieval Conference (TREC) [22]. The TRECVID data contained about 120 hours of ABC and CNN news broadcasts and about 13 hours of recent C-SPAN public hearings. Based on this corpus a total of 17 semantic concepts were defined.

Among the semantic concepts to be detected were concepts like outdoors, zoom-in, vegetation, female speech, aircrafts, and news subject monologues. Given the data, the latter is one of the most interesting concepts from both a users and analysis perspective. A user that is browsing a news archive is most likely searching for past news topics, events, and people. Hence a query mechanism that allows for people based search, like proposed in [18], is highly desirable. Moreover, summarizing story segments by means of keyframes extracted from camera shots that view news subject monologues is far more informative than those showing news anchors, reporters, or commercials. From an analysis point of view there are also major challenges involved. Whereas some concepts can be detected solely on unimodal analysis, the news subject monologue requires analysis of multimodal information and inclusion of context. To clarify this statement, we first take a look at the original TRECVID definition for this task [22]:

Definition 1 (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.*

Hence, this task requires that we detect a person that talks for a while, is not affiliated to the news broadcaster, and is not promoting a commercial message. Based on unimodal analysis this would require very sophisticated detectors that possess a very high recognition accuracy. Given the current state-of-the-art this seems impossible. Therefore, we opt for a true multimedia analysis approach.

The rest of this paper is organized as follows. We will first discuss related work on multimodal person detection in section 2. The detectors are discussed in section 3. In section 4 we will highlight the construction of the classifier ensembles. Experiments and results are demonstrated in section 5.

2 Related Work in Multimodal People Detection

Multimodal people detection methods combine cues from various modalities. This is interesting for many domains, for example the security domain where the task is to reliably identify people based on biometrics like recognized faces and voices [6]. Here, we focus on multimodal analysis methods that aim at content based classification of video segments containing people.

An interactive system that allows to label the main characters in feature films is presented in [24]. The authors introduce a set of similarity measures based on alignment of closed captions and movie scripts, presence of frontal faces, background similarity, temporal coherence, and information from movie encyclopedia's. By computing a combined similarity score a ranked result is presented to the user, who provides the system with positive and negative feedback to label similar labelled shots simultaneously, resulting in a considerable reduction in annotation effort.

In contrast to [24], the Name-It system [18] is fully automatic and associates detected faces and names in CNN Headline News. This is achieved by calculating a co-occurrence factor that combines the analysis results of face detection, face tracking, face recognition and named entity recognition in both transcript and overlayed text. The authors demonstrate that a multimodal approach improves upon using the analysis methods in an individual fashion, despite disappointing performance of some of the methods.

Early integration methods for multimedia analysis, like the one proposed in [18], were based on heuristics. In [21], a framework was proposed that exploits the powerful properties of statistical classifiers by representing multimedia data as time interval relations. To demonstrate the effectiveness of their approach a monologue detector for Dutch broadcast news was developed that combined various multimodal detector results with synchronization relations. However, this method doesn't differentiate monologues from anchors, reporters, and commercials.

The monologue detector presented in [10] reported the best average precision performance in the TREC 2002 video track benchmark. The method combines detected frontal faces and speech with a mutual information based synchrony metric. The method is only partly applicable for the TREC 2003 data set, since detection of a talking face is not enough. Hence, it must be extended to include context for a corpus that contains mostly broadcast news.

In contrast to the above mentioned methodologies for multimodal people detection, our method is different with respect to the explicit inclusion of context in the analysis process. Moreover, the news subject monologues are learned from the video data by an ensemble of statistical classifiers.

3 Detector based Analysis

Video created in a production environment, like broadcast news, requires an author or editor who carefully combines multimodal layout and content elements with a certain style to express a semantic intention. When we want to analyze those produced assets and extract the semantics, this process should be reversed [20].

The reverse analysis process starts with a layout segmentation, and detection of content elements like people, objects, and setting. For this first phase of the analysis we used state-of-the-art analysis components, see table 1 for an overview. Based on those components we introduce *style detectors*, that are able to analyze parts of the author's intention. In combination with extracted context clues, obtained by *context detectors*, this will provide us with the building blocks for inference of the intended semantics of the author.

To circumvent the problems introduced by using a probabilistic or real-valued output for each individual detector we require that the output of a detector is discrete, i.e. binary or categorical. We refer to this discrete result as a feature. For the TRECVID benchmark all results were based on the layout scheme defined by a common camera shot segmentation, therefore all features are synchronized to the granularity of a camera shot.

Because the broadcasts from different channels were created by different authors, thresholds for individual detectors can be expected to vary between stations. All detectors are optimized based on experiments using the training set. Also note that the set of detectors is

Table 1: *Components for multimodal analysis of news subject monologues.*

Analysis component	*Modality*	*Reference*
Camera shot segmentation	Visual	[16]
Motion estimation	Visual	[21]
Frontal face detection	Visual	[19]
Video OCR	Visual & Textual	[17]
Named entity recognition	Textual	[25]
Speaker recognition	Auditory	[9]
Speech recognition	Auditory & Textual	[9]

tailored for the news domain, for C-SPAN we only used a subset.

In this section we will first discuss the feature detectors used for extraction of style, followed by those for extraction of context. We end with a discussion on other features that could be considered for the task of news subject monologue detection.

3.1 Style Detectors

To communicate a predefined intention, an author of a video document has a certain style for arranging layout and content elements. Based on detected layout and content elements this style concept can be reconstructed. We focus here on style detectors that allow us to detect people, since the main content of a news subject monologue is a talking human being.

3.1.1 Face Features

One of the most reliable cues for the presence of a person, is the detection of a human face in the visual modality. Ideally, a face detector should be applied to all images in the visual modality. However, due to the cost of this operation it might be unfeasible to perform it on every frame. Therefore, we have applied a frontal face detector [19] on every 15th frame of each camera shot.

To express the viewer with a sense of being far away from, or close to, the *mise-en-scène* of the camera shot, an author uses a technique called framing [4]. For extraction of this style element we use a set of style detectors that are based on face detection. For each analyzed frame in a camera shot we count the number of faces present, and for each face we derive its location and the camera distance used. This results in a total of fifteen style features per camera shot. If more than one face is detected in a frame the people feature is set. For the location of a detected face we divide an image frame into four equally sized regions: *topleft*, *bottomleft*, *topright*, and *bottomright*. If a face falls completely within one of those four regions the feature for that region is set. If a face covers parts of the topleft and bottomleft 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 those locations the *center* location feature is set. 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 location features. As an estimate for the camera distance we use a face-frame ratio, where the size of

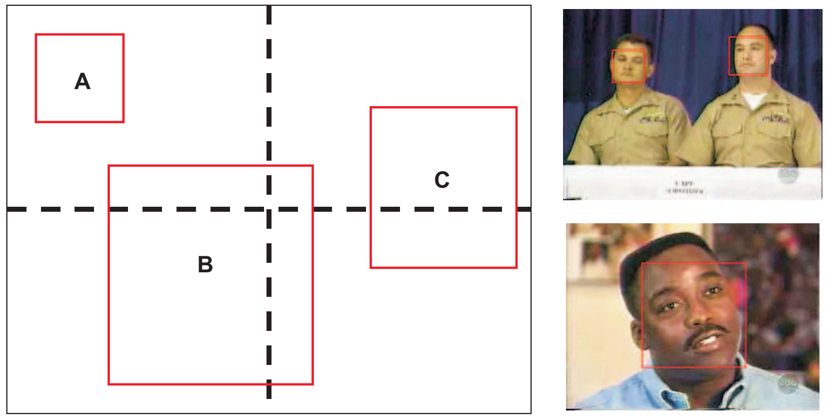

Figure 1: *Left: an image frame with three detected faces, long shot face A is located topleft, close-up face B is located center, and medium shot face C is located right. Right: two example image frames with detected faces.*

a face is related to the size of the frame. Based on this ratio we distinguished seven discrete camera distance features can be distinguished, ranging from extreme long shot to extreme close up.

To summarize, consider the example in figure 1, face A is at long shot distance located in the topleft location, face B is at close-up distance located in the center area and face C is at medium shot distance located in the right area. To aggregate the frame based face features into a camera shot, we require that a feature is true for forty percent of the analyzed frames in a camera shot.

3.1.2 Speaker Features

Besides the visual presence of a person, a news subject monologue requires that someone is talking. However, the presence of speech is not a very informative feature by itself. Given the fact that the data set contains eminent presence of talking people which are not necessarily news subjects, like news anchors, reporters and talking people in commercials. Therefore, we have to exploit style elements that are attached to detected speech.

Based on the LIMSI speech detection and recognition system [9] we developed a voice over detector and a frequent speaker detector, see figure 2 for an example. Voice over detection consists of two phases, first we count the number of cuts in the corresponding camera shot segmentation, this results in 2, 0, 2, and 0 cuts for each speech segment in the example. Note that to account for imperfect segmentation, a margin of 25 frames was extracted from each end of a speech segment before counting. We consider a speech segment a voice over when it contains more than 1 cut. To map the voice over speech segments to camera shots we use TIME relations [21]. This results in a segmentation of camera shots that have a voice over (S_3). Frequent speaker camera shots are detected in a similar fashion: first we count the three most frequent speakers in a video document, all speech segments that are uttered by one of those frequent speakers, e.g. speaker I, are then mapped to camera shots (S_4).

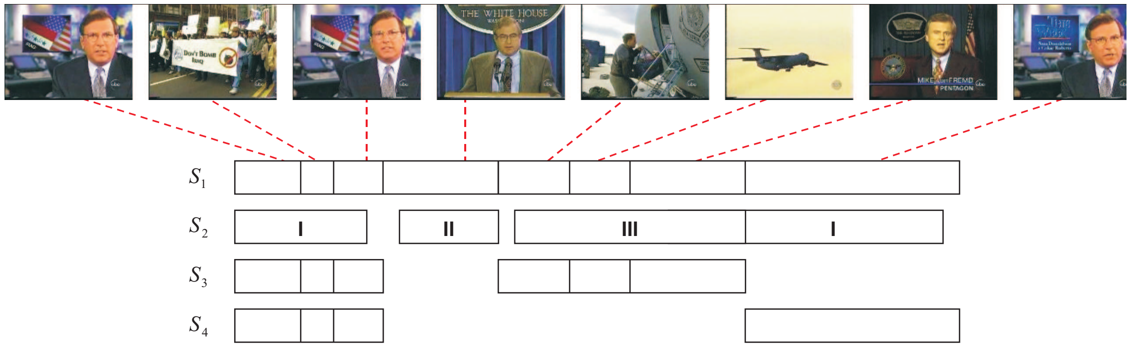

Figure 2: *Segmentation of a video document into camera shots (S_1), speech segments (S_2), voice over camera shots (S_3), and frequent speaker camera shots (S_4).*

3.1.3 Video OCR Features

When a news subject person is given broadcast time on television, it is common to display the name of this person to let the viewer know who is talking. To communicate this information an author uses overlayed text that is added during production time. Video Optical Character Recognition (OCR) [17] was applied to extract this overlayed text[1].

Unfortunately however, overlayed text is not used exclusively for annotation of people names. Other functions of overlayed text in news broadcasts include annotation of setting, objects, and events. Moreover, overlayed text is often used in commercials as a means to communicate product names, claims, and disclaimers [20]. See figure 3 for some examples of typical usage of overlayed text in the TRECVID corpus.

Because of the varying functionality of overlayed text, detection of its presence is not enough. Therefore we use the total length of the overlayed text strings that are recognized as an additional style feature. The rationale here is that overlayed text that is used to display people names is mostly shorter than overlayed text that is used for graphical shots or commercials. The text string resulting of Video OCR was also used as input for a named entity recognizer that is part of the Informedia system [25]. The categorical result is stored as a feature. It can be expected that a string that is recognized as a persons name is of higher value for news subject monologues than one that is recognized as a location. Furthermore, to differentiate news subject monologues from reporters, the detected strings where compared, using fuzzy string matching, with a database of names of CNN and ABC affiliates. If a match was found, an affiliate feature was set. The names were extracted from the corporate website of both CNN and ABC.

3.1.4 Tempo Features

To affect the overall rhythm or tempo of a video document an author can apply a variety of stylistic techniques. In [2] it was shown that editing and motion are important contributors to this style element. Therefore, we introduce two style features based on this observation.

The first tempo feature is a simple camera shot length measure. The rationale for using

[1]For CNN the ticker tape with stock information on the bottom of the image frame was ignored.

Figure 3: *Common usage of overlayed text in the TRECVID corpus. From top left to bottom right: topic annotation, location annotation, reporter annotation, financial data, news subject monologue annotation, and commercial messages.*

this feature is that a news subject monologue has a minimum duration, since it takes some time to tell something. Also, for viewers it would be very disruptive if news subject monologues were only viewed for a very short period. Short camera shots, i.e. less than 70 frames, are therefore less likely to contain a news subject monologue.

The second tempo feature measures the average amount of motion in a camera shot, based on frame difference [21]. We distinguish between three classes of motion, low motion, medium motion, and high motion. The first class is meant to detect camera shots where there is little or no motion, examples include graphical shots and still images. The second class contains some motion, but not very much. This is the typical class for news subject monologues and anchors. And finally the third class, that contains a considerable amount of either object motion, camera motion, or both. Typical report footage is a good example for this class.

3.1.5 Speech Transcript Features

Our last detector exploits style that is expressed in the uttered speech. The main usage of this feature is to get rid of talking people that cause confusion with news subject monologues, e.g. financial news reporters and commercials with monologues. The output of the LIMSI speech recognition system [9] was compared with a set of keywords that was found to have a correlation with reporters, financial news, and commercials.

For reporters it is common to pronounce their name together with the broadcast station name at the end of a report. Hence, we checked for the presence of reporter and broadcast station names in the recognized speech. In financial news reports the reporters typically talk about stock markets and shares, therefore we checked the speech transcript for occurrences of keywords like *Nasdaq*, *Dow Jones*, *multimillion*, *shares*, *stocks*, and so on. In an effort to sell a product or service, advertisers often use smiling ladies that urges the viewer of a commercial to *sign up* or call a *one eight hundred* number now. The transcript was analyzed for those phrases to prevent false positive classification of commercial footage as news subject monologues.

3.2 Context Detectors

In contrast to style detectors, which address part of the layout and content as stylized by the author, the result of a context detector is related to the overall author intention. In [20] we identified five hierarchical semantic index types, namely: author purpose, chosen genre and sub-genre, and labels that can be attached to logical units and events. When such a semantic index can be detected for a video segment, it can be used as a context clue for classification.

Context clues can enhance or limit the interpretations resulting from content-based analysis, i.e. it can have both a positive and negative influence on the final classification result. When we aim for classification of gun duels for example, a context detector that tells us that the current video document is a western movie would probably have a positive influence on the final result. On the other hand, a detected car chase will most likely have a negative impact on the final result. We focus here on context detectors that will reduce false interpretations of camera shots as news subject monologues.

A requirement for a news subject monologue is that it belongs to the news genre. However, the broadcasts from the corpus also contain a lot of footage from another genre, namely commercials. Although they may contain monologues of people promoting a product or service, those should not be labelled as news subject monologues. Therefore, we used a context detector that is able to detect commercials. We used the commercial detector that was developed for the Informedia TRECVID contribution [3]. It combines five Fisher Linear Discriminants from both a set of audio features and color features and feeds those into a Support Vector Machine for classification.

An anchor shares many characteristics with a news subject monologue, it is therefore important that we can distinguish anchors from other footage to circumvent a false interpretation. To stress this importance we used two anchor detectors. The first one is part of the Informedia system [25] and combines a set of region dependent color features that are concatenated into one vector that is feeded into a Support Vector Machine classifier. The second anchor detector was developed for TRECVID and extracts five Fisher Linear Discriminants of a set of color features, a frequent speaker identifier, and face features. The features are also combined in a Support Vector Machine for the final classification [3].

3.3 Discussion

The features presented in the previous sections are by no means exhaustive for the classification of news subject monologues. They were chosen because analysis components with reasonable performance were available. We selected components whose performance had either been proven in the literature or by our own experiments on the training set of the TRECVID corpus. Other components, i.e. face recognition and lip movement detection, were tried but were found to lack robustness. However, we believe that both components are promising for future investigation. Face recognition for example can be exploited in a similar fashion as the LIMSI speaker recognition, i.e. focus on recognition of repetition within one video document instead of generic recognition which is a very hard problem. For lip movement detection to be successful, both quality and resolution of the MPEG video's should be higher. In general, components are only useful for analysis when they perform their task with a certain reliability.

The current set of style detectors can be extended in several ways. A measure could be added to the face features that incorporates the frontal consistency of a face that is filmed.

The speaker features could be extended with a long silence detector that checks for relatively long pauses during a camera shot. For Video OCR the location of the overlayed text on the image frame together with the temporal location within the camera shot would probably be helpful. For further characterization of tempo, audio features could be beneficial. The uttered speech could be further analyzed to discover patterns in uttered word rate, e.g. do anchors and reporters speak faster on average? Also specific phrases that are more likely to have been uttered by news subject monologues than anchors or reporters are worthwhile candidates for further study.

Diverse other worthwhile extensions to the set of context detectors exist. Most obvious candidates would be detectors that are related to the specific genre of news video documents, i.e. (financial) reporter detectors, studio setting detector, weather report detector, and so on. However, less obvious detectors, such as outdoor, animal, car, sporting event and so on, might show possible hidden relations with the semantic interpretation of news subject monologues, and are therefore interesting for further research.

4 Combining Weak Detectors

All detectors share a common characteristic: they are imperfect and generate both false positive and false negative results. Each individual detector can therefore be considered as a weak classifier. In this section we will first elaborate on methods for combining weak classifiers, followed by a discussion on the important aspect of ranking the final results.

4.1 Classifier Ensembles

From the field of statistical pattern recognition the concept of classifier ensembles is well known. A classifier ensemble is believed to benefit 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 [11]. To assure this independence, one can use classifiers that are based on different characteristics of the data, e.g. multimedia detectors, or by exploiting variation in the training set, e.g. by resampling. We combine both methods by exploiting two well known classifier combination schemes, namely stacking [26] and bagging [5].

In its common use, stacking combines results of different classifiers that solve the same task. The output labels of those individual classifiers are then used as input features for a *stacked* classifier, which learns how to combine the reliable classifiers in the ensemble and makes the final decision. However, the same technique can also be used to combine classifiers that do not solve the same task per se, but are related contextually. Hence, the output of the weak learners discussed in section 3 can be used by a stacked classifier to learn new concepts based on context. As a stacked classifier we chose the Support Vector Machine (SVM) [23], which is known to be a stable classifier for various computer science problems and has also proven to be a good choice in a multimodal video indexing setting [1, 21]. In an SVM each pattern x is represented in a n-dimensional space, spanned by the n detectors. Within this space an optimal hyperplane is searched that separates the 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. The problem of finding the optimal hyperplane is a quadratic programming problem of the following form [23]:

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

Under the following constraints:

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

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. To account for the fact that our data is unbalanced, i.e. the concepts we are looking for are rare, we use an adaptation of (1) that allows us to balance the data [7]. The stacked SVM classifier is combined with bagging to create the final classifier ensemble.

Bagging, or bootstrapped aggregating, creates T redistributions from a training set of N instances by randomly replicating and deleting individual training instances. For each iteration $t \in T$ a classifier γ_t is trained. Hence, each classifier is based on a different sample of the original training set. The results of each γ_t are then aggregated to form the final classifier. For the aggregation typically the sum rule is used, as it is known to outperform other methods [13]. As noted by [5], the vital point in bagging is the instability of the classifier γ_t. It can therefore be argued whether a stacked SVM is a good choice for γ_t, since SVM is a stable classifier. When used in statistical pattern recognition, bagging aims to minimize the classification error. However, we use bagging in the context of multimedia retrieval, where we aim to find as many relevant items as possible. Those relevant items are almost always outnumbered by irrelevant items. For retrieval it is thus important that the relevant items are ranked as high as possible. This ranking mechanism can be improved by using bagging.

4.2 Ranking Results

A drawback of using an SVM in combination with stacking and bagging is that its uncalibrated classification result is not a good comparison measure for ranking. Ideally one would like to have a posterior probability, $p(\omega|x)$, that given an input pattern x returns a confidence value for a particular class ω. A simple approach to achieve this is to use a threshold τ on the uncalibrated SVM output, $\gamma_t(x)$. This results in an abstract class label $\delta_t(x)$ defined as:

$$\delta_t(x) = \begin{cases} 1, & \text{if } \gamma_t(x) \geq \tau; \\ 0, & \text{otherwise;} \end{cases} \tag{3}$$

By averaging the class labels, a simple posterior probability measure can be computed:

$$p(\omega|x) = \frac{1}{T} \sum_{t=1}^{T} \delta_t(x), \quad \forall x \in X \tag{4}$$

where T is the number of SVMs in the ensemble and X is the number of patterns. Although this results in a ranking measure, it is not very likely to be optimal, since there is no confidence value associated to $\delta_t(x)$.

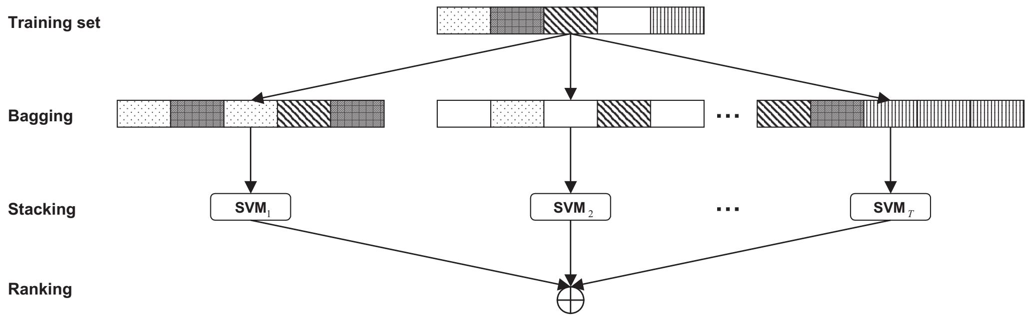

Figure 4: *Classifier ensemble architecture.*

A more popular and stable method for SVM output conversion was proposed in [15]. This solution is based on the observation that class-conditional densities between the margins are exponential, therefore a sigmoid model is suggested. In our classifier ensemble architecture the output of this model is averaged over all individual classifiers, resulting in the following posterior probability:

$$p(\omega|x) = \frac{1}{T}\sum_{t=1}^{T}\frac{1}{1+\exp(\alpha_t\gamma_t(x)+\beta_t)}, \quad \forall x \in X \tag{5}$$

where the parameters α_t and β_t are maximum likelihood estimates based on the tth redistribution of the training set [15]. Rankings based on (5) can be expected to be more stable than those based on (4). An overview of the classifier ensemble architecture is given in figure 4.

5 Results

To evaluate the viability of our approach for learning news subject monologues, we carried out a set of experiments as part of the TRECVID benchmark. The TRECVID corpus was split into an equally sized training and test set, i.e. each containing about 65 hours of broadcast video. Since annotating those videos for all 17 semantic concepts requires a vast amount of human effort, the TRECVID research community initiated a common annotation effort, guided by IBM research. Although, this common annotation effort was a very welcome initiative, it suffered from some initial problems. We observed that for unambiguous concepts, like outdoor, this common annotation worked very well. However, after manual inspection of the common ground truth for news subject monologues we found that the common annotation was less useful. Whether this was caused by ambiguity resulting from the TRECVID definition or bad annotation, is unclear. Based on the observed inconsistencies we decided to label our own ground truth for training of the classifiers. We labelled 23 ABC broadcasts, 24 CNN broadcasts, and all 19 C-SPAN broadcasts from the training set, about 29 hours in total.

In this section, we will first highlight the evaluation criteria used by TRECVID, followed by our initial results on the TRECVID benchmark. Then we will present an experiment that shows the contribution of individual detectors. After that we proceed with an experiment that evaluates both the influence of ensemble size and ranking mechanism. We end with some possible applications scenarios of news subject monologue concept detection.

5.1 Evaluation Criteria

Traditional evaluation measures from the field of information retrieval are precision and recall. Let R be the number of relevant camera shots, i.e. camera shots containing the specific semantic concept one is looking for. Let A denote the answer set, i.e. the number of camera shots that are retrieved by the classifier. Let $R \cap A$ be the number of camera shots in the intersection of the sets R and A. Then, precision is the fraction of retrieved camera shots (A) which are relevant:

$$Precision = \frac{R \cap A}{A} \tag{6}$$

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

$$Recall = \frac{R \cap A}{R} \tag{7}$$

This measure is indicative for the amount of correct classifications, false positive classifications, and false negative classifications. For evaluation within TRECVID both measures are combined in an *average precision*, AP, measure. This single-valued measure corresponds to the area under an ideal precision-recall curve and is the average of the precision value obtained after each relevant camera shot is retrieved. This metric favors highly ranked relevant camera shots. Let L be a ranked version of A. At any given index i let $R \cap L_i^1$ be the number of relevant camera shots in the top i of L, then AP is defined as:

$$AP = \frac{1}{R} \sum_{i=1}^{A} \frac{R \cap L_i^1}{i} \lambda(L_i) \tag{8}$$

Where $\lambda(L_i)$ is defined as:

$$\lambda(L_i) = \begin{cases} 1, & \text{if } L_i \in R; \\ 0, & \text{otherwise;} \end{cases} \tag{9}$$

We used the AP evaluation measure as the basic metric for the conducted experiments.

5.2 TRECVID Benchmark

To evaluate the AP for each submitted run, TRECVID uses a pooled ground truth, P, for the test set. From each submitted run a fixed number of ranked shots is taken, those are combined in 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 equation (8), we use P where $P \subset R$.

There were a total of 26 submissions for the news subject monologue detection task, the results are summarized in figure 5. The first column indicates the median performance among all submitted runs. The second column shows the AP performance of the second best system. The third column shows our run using training data based on the common ground truth. The fourth column shows our best run and the overall best performer of the 2003 TRECVID benchmark for news subject monologue detection. Note that our approach is better by more than a factor of ten when compared to the second best competing system.

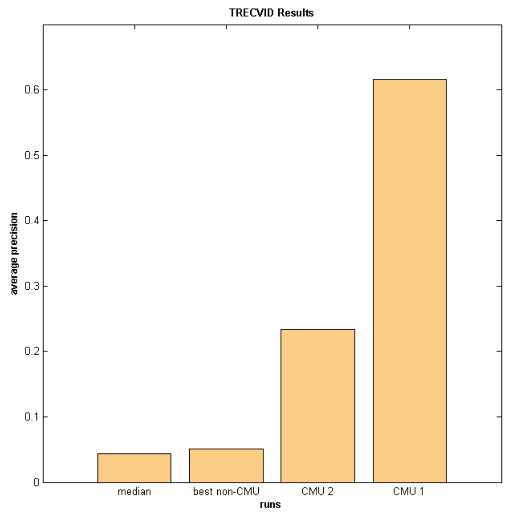

Figure 5: *TRECVID 2003 results for news subject monologue detection. The first column, labelled median, indicates the median average precision for all submitted runs. The second column, labelled best non-CMU, is the score of the second best system. The third column, labelled CMU 2, shows our performance based on the common annotation. The fourth column, labelled CMU 1, shows our best submitted run.*

Also note the drop in AP when a bad ground truth from the inconsistently labelled common annotation is used. An overview of the distribution of correct, false, and unknown labelled shots in our best run, according to the TRECVID pooled ground truth, is given in table 2.

Our best run was based on an early version of our system. The run combined an ensemble of 200 classifiers with a ranking mechanism that was based on the average of the abstract class label, as defined in (4), with a slight modification. Instead of using only the average output we used a round-robin scheduling of results per station. The scheduling was based on the estimated number of news subject monologue in the test set per station. For this estimate the prior statistics of our manually labelled ground truth were used. Based on this estimate we expected that there were about 200 news subject monologues in C-SPAN, 1300 in ABC, and 700 in CNN. The final result set was created by taking 2 shots from C-SPAN,

Table 2: *Distribution of correct, false, and unknown labelled shots using TRECVID pooled ground truth on our best submitted run.*

Evaluated shots	*10*	*20*	*50*	*100*	*200*	*500*	*1000*	*2000*
Correct	10	19	42	86	144	191	226	249
False	0	1	8	14	22	40	59	107
Unknown	0	0	0	0	34	269	715	1644

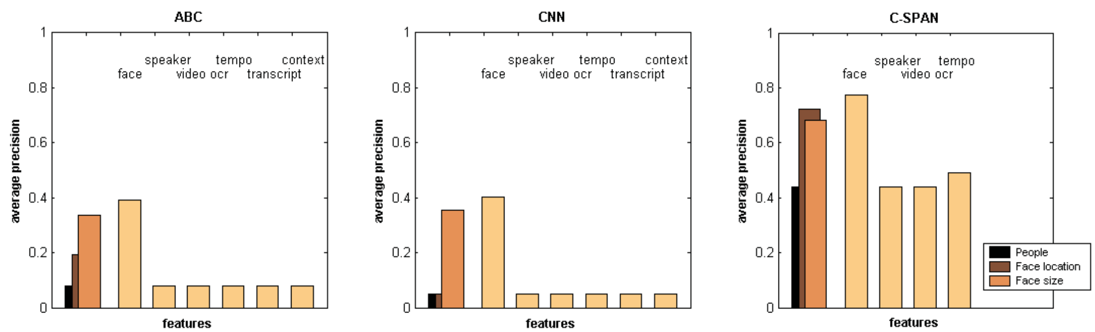

Figure 6: *The effect of using only one feature on average precision. Face features are split in people, location, and size.*

13 from ABC, and 7 from CNN in a round-robin fashion until the maximum number of camera shots was reached (2000). Although, this run was the best performing news subject monologue detector in TRECVID 2003, we believe its performance can be improved by using a probabilistic ranking mechanism.

5.3 Feature Contribution

To evaluate the feature contribution to the final classification result we performed an exhaustive search on some possible subsets of the feature types. For each combination and broadcast station one stacked SVM classifier was trained. In absence of annotation for the entire test set, those classifiers were tested on the training set. The output of the SVM was converted to a posterior probability using (5), where $T = 1$. Results for each feature combination and station were ranked based on this probability. Note that we used a subset of features for C-SPAN, i.e. containing only face features, voice over speaker feature, video OCR features, and tempo features.

First, we computed AP for individual features to show their contribution to the classification of news subject monologues, see figure 6. The results demonstrate that for both ABC and CNN, only the face features are useful as single feature. To get a better insight of the contribution of face features we split those into features that detect people, face location, and face size. Based on this division it becomes clear that face size for both ABC and CNN and face location for ABC are the only features that allow to detect news subject monologues by themselves. Based on the other features the AP result equals the result of the default ranking, i.e. they have no contribution. For C-SPAN the results are almost similar, like ABC both face location and face size are useful, unlike ABC and CNN, tempo is also able to give a better than default AP value.

In our second experiment we evaluated the effect of feature combinations. We incrementally added a feature to the best performing previous combination. Hence, for each station we started with the face features, the next feature was chosen based on the largest gain in AP, and so on. The results are visualized in figure 7. For both ABC and CNN, the speaker features resulted in the largest gain in AP when added to the face features. When adding more features, Video OCR was more important for CNN and C-SPAN. Context is important

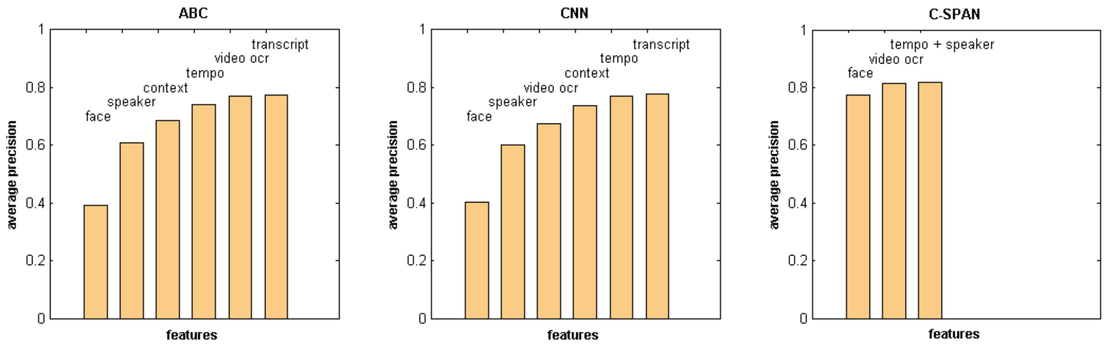

Figure 7: *The effect of incrementally adding features that contribute most to an improvement in average precision.*

for both ABC and CNN to improve results. The transcript features are of little use for ABC and CNN, tempo and speaker feature only slightly improve the overall result for C-SPAN. However, the best combination for all stations, i.e. the combination with best AP, is the combination that exploits all features simultaneously.

In our third and final experiment we repeatedly removed one feature from the total pool of features, and measured the decrease in average precision. The results are visualized in figure 8. As expected, the largest decrease is obtained when the face features are removed from the feature set. When we take a closer look to the individual face features, it shows that by themselves the people and location features have a minimal contribution to the decrease. However, when combined with the face size features the drop in AP is significant. This also holds for C-SPAN. The influence of removing other features shows a less significant drop in AP. For both ABC and CNN, context and speaker features lower AP more than the other features, for C-SPAN removing the Video OCR features reduces AP more than removing speaker and tempo. Again the influence of the transcript features is minimal for both ABC and CNN.

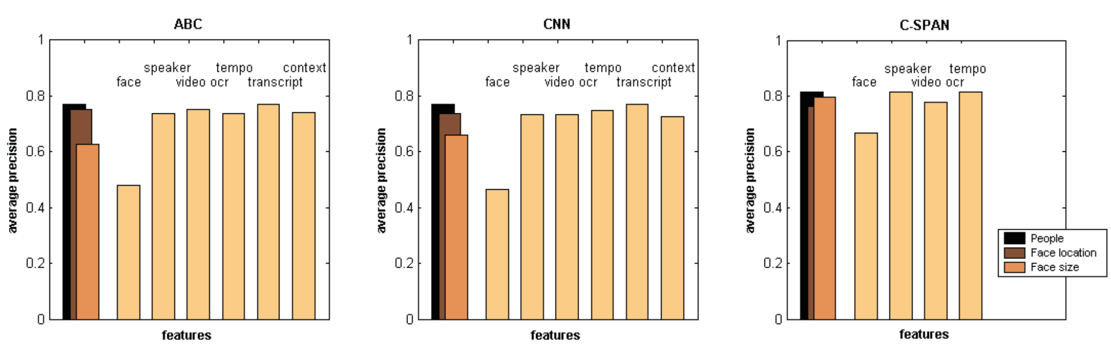

Figure 8: *The effect of removing one single feature from the total set of features on average precision. Face features are split in people, location, and size.*

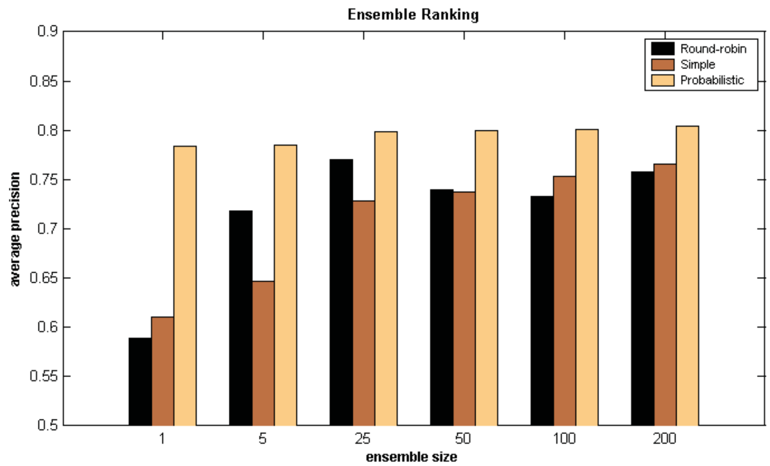

Figure 9: *Influence of ensemble size on average precision, using round-robin, simple, and probabilistic ranking.*

After this set of experiments it becomes clear that face features are important for news subject monologue identification. However, detection of a face by itself is not enough, the style of usage measured by number of faces, face location, and face style is important. Moreover, a significant gain in AP can be obtained when the face features are used in combination with other style and context features, although those features may have no significant individual contribution to the classification task at hand.

5.4 Ensemble Ranking

To show the merit of using a probabilistic ranking method, we performed an extra set of experiments using the simple (4) and probabilistic (5) ranking mechanisms proposed in section 4.2 in combination with an increasing ensemble of classifiers. For completeness we also included a run based on the round-robin ranking of our best submitted run to TRECVID.

Unfortunately TRECVID only provided a pooled ground truth, which is fine for comparison of submitted runs, but when new experiments are performed the pooled ground truth is too sparse and too much specific for the submitted runs. Due to this sparseness, highly ranked unknown labels have a very negative influence on AP, when the top 20 of a run contains a lot of unknown camera shots for example they will degrade AP the same way as when they would be false. Therefore, we modify the basic AP measure in (8) by only updating the denominator i for labels that are known, i.e. only correct and false ones. This has a positive bias on average precision, but is a more reliable metric for comparing new runs. To give a fair performance comparison we also repeated our best run of the TRECVID submission[2], and calculated the modified AP. The results are visualized in figure 9.

As the graph indicates probabilistic ranking outperforms the round-robin and simple ranking mechanisms. Their is also a clear relation between ensemble size and AP, which is

[2]Due to the random factor in the construction of the ensemble, caused by the bagging algorithm, this run is not exactly identical.

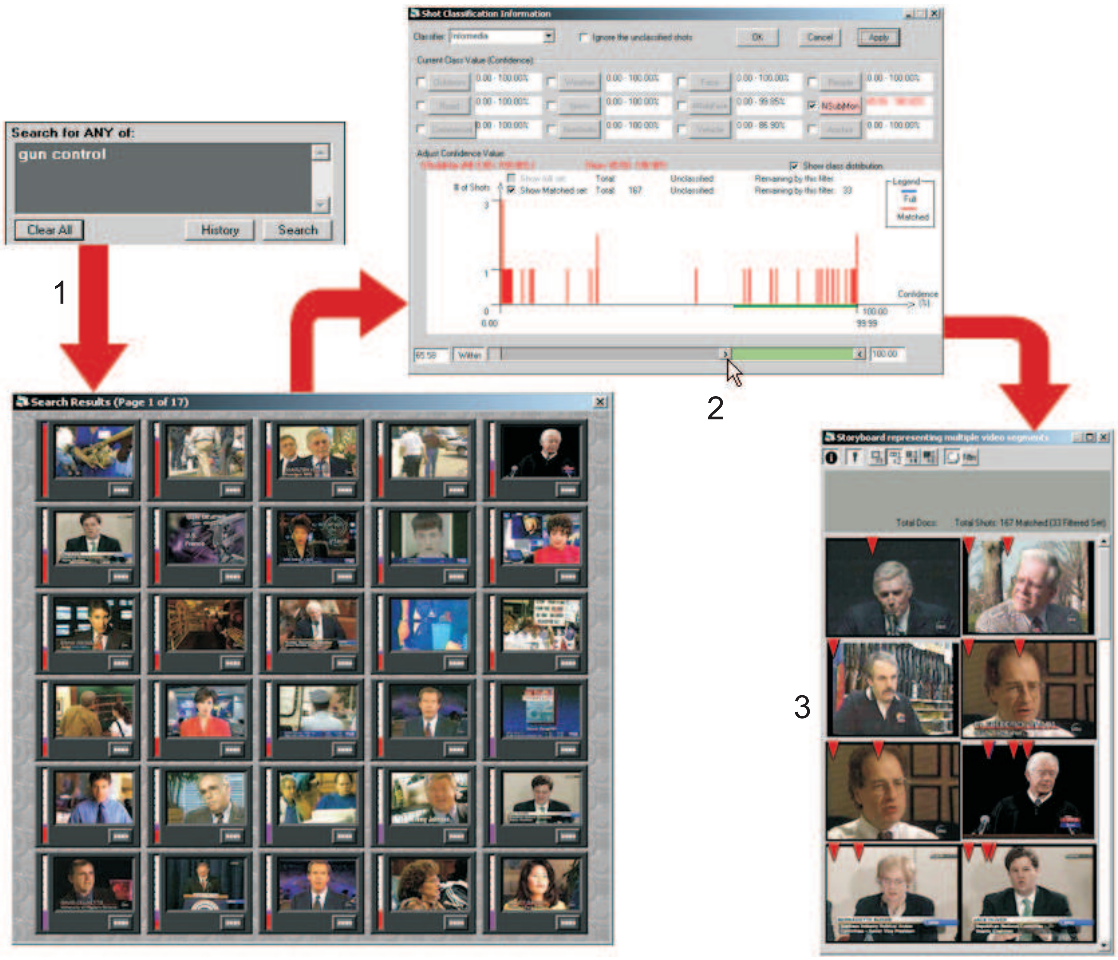

Figure 10: *Example of a search scenario. After entering a search query (1), the user is presented with an overflow of query related video segments. By applying a semantic filter (2) the result set is reduced to video segments relating to the semantic concept of interest (3).*

the most apparent for simple ranking. The round-robin ranking outperforms simple ranking for small ensemble sizes, but is outperformed by both simple and probabilistic ranking when the ensemble contains more than 50 classifiers. The best TRECVID submission, round-robin with an ensemble of 200 classifiers, is outperformed by its equivalent using simple or probabilistic ranking.

5.5 Application Scenarios

Automatic detection of semantic concepts, like news subject monologues, facilitates various innovative application scenarios. In a digital multimedia library, e.g. the Informedia system [25], a *search scenario* is important. Based on a user query the Informedia systems returns a storyboard of video segments that are related to a textual query term. However, the user can be overwhelmed with the results. Semantic filters [8] are therefore a welcome extension to the search functionality. By filtering the results on their probability of being related to a certain concept, a user might be presented a more useful query result, see figure 10 for an example search scenario. In step 1 a search query is entered to the system, resulting in an overflow of returned results. In step 2 the user chooses to filter those results by requiring a high probability for news subject monologues. Finally, in step 3, the user is presented a manageable set of query related news subject monologue results.

In an attempt to make pervasive broadcasting devices more personalized they could be extended based on an *alert scenario*. Where given a certain user profile, the system alerts the user with possible interesting video segments containing news subject monologues or other semantic concepts of interest.

For archiving purposes, for example by intelligence agencies or broadcast channels, a *logging scenario* would be worthwhile to consider as application. Storing the detected semantics together with the video sources, allows for future mining of the data and reduces the amount of effort necessary by human intervention.

6 Conclusions and Future Research

Multimedia content and context should be combined to narrow the semantic gap. We have used the news subject monologue task of the 2003 TRECVID benchmark as a case study to demonstrate that by using style and context detectors, in combination with statistical classifier ensembles, semantic concepts can be learned reliably. Based on conducted experiments we conclude that the combination of various detectors is decisive for the final classification result, although some detectors might appear useless in isolation. Our TRECVID submission resulted in the best average precision for this task amongst 26 contributions. Moreover, we were able to improve upon this result by exploiting a probabilistic ranking in combination with a large number of classifiers in the ensemble.

With respect to future research, we consider three possible types of extensions: enlarging the scope of detected news subject monologues, extension of the methodology to other semantic concepts, and exploration of other classifier combination and ranking schemes.

The current system can be extended in several ways. More style and context detectors can be added to improve results. Other domains can be considered, e.g. talk shows or documentaries, or the type of news subject monologue can be specified, e.g. interview or speech. By inclusion of more textual sources an even richer description can be given to news subject monologues by adding explicit names or topics.

An interesting extension of our methodology would be to investigate whether the same approach can be applied to other examples of semantic concepts, using similar and new types of style and context detectors. There are various possible semantic concept candidates for future research, a good start would be the ones defined in the TRECVID benchmark, since annotations and shared features for those tasks are available.

Other classifier combination and ranking schemes, e.g. Borda count, can be explored, and the number of classifiers in the ensemble can be increased. The number of possibilities left to explore is therefore quite large, and their impact on detecting semantic concepts will eventually boost progress in content based multimedia analysis.

Acknowledgement

The authors thank the people at Informedia and Carnegie Mellon University for their hospitality and assistance. Special thanks to Robert Baron and Dorbin Ng for general Informedia system support. Ming-yu Chen and Norman Papernick are acknowledged for developing the commercial and anchor detectors.

References

[1] B. Adams, A. Amir, C. Dorai, S. Ghosal, G. Iyengar, A. Jaimes, C. Lang, C.-Y. Lin, A. Natsev, M.R. Naphade, C. Neti, H.J. Nock, H.H. Permuter, R. Singh, J.R. Smith, S. Srinivasan, B.L. Tseng, T.V. Ashwin, and D. Zhang. IBM research TREC-2002 video retrieval system. In *Proceedings of the 11th Text Retrieval Conference*, Gaithersburg, USA, 2002.

[2] 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.

[3] R.V. Baron, M.-Y. Chen, M. Christel, P. Duygulu, R. Jin, W.-H.Lin, T. Ng, N. Papernick, C.G.M. Snoek, G. Tzanetakis, H.D. Wactlar, J. Yang, R. Yang, and A.G. Hauptmann. Informedia at TRECVID 2003: Analyzing and searching broadcast news video. In *Proceedings of the 12th Text Retrieval Conference*, Gaithersburg, USA, 2003. To appear.

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

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

[6] R. Brunelli and D. Falavigna. Person identification using multiple cues. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 17(10):955–966, 1995.

[7] 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/`.

[8] M. Christel and C. Huang. Enhanced access to digital video through visually rich interfaces. In *Proceedings of the IEEE International Conference on Multimedia & Expo*, Baltimore, USA, 2003.

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

[10] G. Iyengar, H.J. Nock, and C. Neti. Audio-visual synchrony for detection of monologues in video. In *IEEE International Conference on Multimedia & Expo*, pages 329–332, Baltimore, USA, 2003.

[11] 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.

[12] R.S. Jasinschi, N. Dimitrova, T. McGee, L. Agnihotri, J. Zimmerman, D. Li, and J. Louie. A probabilistic layered framework for integrating multimedia content and context information. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 2057–2060, Orlando, Florida, 2002.

[13] J. Kittler, M. Hatef, R.P.W. Duin, and J. Matas. On combining classifiers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(3):226–239, 1998.

[14] 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.

[15] J. Platt. Probabilistic outputs for support vector machines and comparison to regularized likelihood methods. In *Advances in Large Margin Classifiers*, pages 61–74, 1999.

[16] G.M. Quénot, D. Moraru, L. Besacier, and P. Mulhem. CLIPS at TREC-11: Experiments in video retrieval. In *Proceedings of the 11th Text Retrieval Conference*, Gaithersburg, USA, 2002.

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

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

[19] H. Schneiderman and T. Kanade. Object detection using the statistics of parts. *International Journal of Computer Vision*. To appear.

[20] C.G.M. Snoek and M. Worring. Multimodal video indexing: A review of the state-of-the-art. *Multimedia Tools and Applications*. To appear. [Online]. Available: `http://www.science.uva.nl/~cgmsnoek/pub/mmta.pdf`.

[21] C.G.M. Snoek and M. Worring. Multimedia event based video indexing using time intervals. Technical Report 2003-01, Intelligent Sensory Information Systems Group, University of Amsterdam, 2003.

[22] TRECVID 2003 Guidelines. `http://www-nlpir.nist.gov/projects/tv2003/tv2003.html`.

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

[24] J. Vendrig and M. Worring. Interactive adaptive movie annotation. *IEEE Multimedia*, 10(3):30–37, 2003.

[25] 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.

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