A Hybrid Approach to Improving Semantic Extraction of News Video

A.G. Hauptmann, M.-Y. Chen, M. Christel, W.-H. Lin, and J. Yang
School of Computer Science, Carnegie Mellon University, Pittsburgh USA
{alex, mychen, christel, whlin, juny}@cs.cmu.edu

Abstract

In this paper we describe a hybrid approach to improving semantic extraction from news video. Experiments show the value of careful parameter tuning, exploiting multiple feature sets and multilingual linguistic resources, applying text retrieval approaches for image features, and establishing synergy between multiple concepts through undirected graphical models. No single approach provides a consistently better result for every concept detection, which suggests that extracting video semantics should exploit multiple resources and techniques rather than a single approach.

1. Introduction

Increasingly, the detection of a large number of semantic concepts is being seen as an intermediate step in enabling semantic video search and retrieval [1]. Early video retrieval systems [2, 3] usually modeled video clips with a set of (low-level) detectable features generated from different modalities. These low-level features like histograms in the HSV, RGB, and YUV color space, Gabor texture or wavelets, structure through edge direction histograms and edge maps can be accurately and automatically extracted from video. However, because the semantic meaning of the video content cannot be captured faithfully by these low-level features, these systems had a very limited success in retrieving video for complex and semantically-rich queries. Several studies have confirmed the difficulty of addressing information needs with such low-level features [4, 5].

To fill this "semantic gap", one approach is to utilize a set of intermediate "textual descriptors that can be reliably applied to visual content (e.g., outdoors, faces, animals, etc.) [6]. Many researchers have been developing automatic semantic concept classifiers such as those related to people (face, anchor, etc), acoustics (speech, music, significant pause), objects (image blobs, buildings, graphics), location (outdoors/indoors, cityscape, landscape, studio setting), genre (weather, financial, sports) and production (camera motion, blank frames) [7]. The task of automatic semantic concept detection has been investigated by many studies in recent years [8-13], showing that these classifiers could, with enough training data, reach the level of maturity needed to be helpful filters for video retrieval [14, 15].

Since so far only very few high-level concepts can machine reliably extracted, the quest for developing better concept classifiers is never ending. Instead of focusing on single approach we test a wide collection of approaches in improving video concept extraction. The hybrid approach described in Section 4 spans a wide range of classifier development process from parameter tuning (Section 4.1), combining multiple features (Section 4.2), exploiting relationship between multiple concepts (Section 4.3), and fusing multiple linguistic resources (Section 4.4). The hybrid approach, unlike previous work that focus on only developing better features or fusing techniques, may provide a more holistic answer to the question: at what level will we obtain most improvement for extracting video semantics? We tested the hybrid approach on a well-established testbed, TRECVID (Section 2) using common low-level features (Section 3). The findings and future direction are summarized in Section 5.

2. TRECVID Semantic Concept Detection

The main forum for studying video retrieval, and in the last few years, video retrieval aided by semantic concepts, has been organized by the National Institute of Standards and Technology (NIST) in the form of the TRECVID video retrieval evaluations [16]. In 2001, NIST started the TREC Video Track (now referred to as TRECVID [17]) to promote progress in content-based video retrieval via an open, metrics-based evaluation, where the video corpora have ranged from documentaries, advertising films, technical/educational material to multi-lingual broadcast news. As the largest video collections with manual annotations available to the research community, the TRECVID collections have become the standard large-scale testbeds for the task of multimedia retrieval [18]. These evaluations provide a standard collection available to all participants, separated into training and development sets.

Currently, TREVID focuses on the video news domain because it is structured video and contains a

broad range of information. The TRECVID 2006 collection contains three different languages: Arabic, Chinese and English. The video collection comes from 11 different sources, including different Arabic news, a variety of Chinese news, CNN, NBC and MSNBC for English news. The development and test sets each contains about 160 hours of video. Each video in the collection is decomposed into shots, which are used as the basic units of the video content. We use the keyframes as defined by the TRECVID benchmark [18], allowing more standardized testing and comparison. We split this data into two parts, with half used for training the models (development set) and the other half is used to evaluate the models (testing set). Since many experiments require parameter tuning, the development set is further split into two parts, with roughly 65% used for training the basic labels (including cross-validation), and the rest for tuning the combination parameters.

Since the video contains rich information from visual, audio, and text extracted from screen and speech recognition, in our experimental setting, we use a number of color, texture, and text features extracted from key-frame images in each shot, described below.

In the following experiments, we utilized the semantic labels of 39 concepts from Large Scale Concept Ontology for Multimedia Understanding (LSCOM) [19] workshop. This set of manual annotation labels for the development set is publicly available.

3. Low-level features

Low-level features constitute the most "atomic" building blocks of our analysis. They are used as the initial features in a variety of machine learning approaches.

Our experiments to detect high-level features are based on 4 different types of low-level features: color moment feature, Gabor texture feature, local image features, and text (transcript) feature, briefly described as follows:

- **Color moment & Gabor texture**: Columbia University [20] provided color and texture features. To generate the color moment feature, each image (key-frame) is divided into 5x5 grids, and each grid is described by the mean, standard deviation, and third root of the skewness of each color channel in the LUV color space. This results in a 225-dimension (5x5x3x3) color moment feature. Texture feature comes from the Gabor filter, which denotes an image by mean and standard deviation from the combination of four scales and six orientations [21].
- **Local features**: The local feature of each image is computed from the local interest points (as known as keypoints) detected from the image. We use the keypoints [22] provided by City University of Hong Kong, which are detected using the DoG detector and depicted by SIFT descriptors [23]. Details on experiments with these keypoints are described later in this paper.
- **Text features**: Text features have been shown to successfully complement visual features in constructing effective multi-modal visual classifiers. Extracting text features on a multilingual corpus, such as TRECVID'06, however, faces an additional problem: how should we effectively combine information from multiple languages? One straightforward solution is to translate multilingual text (e.g., ASR transcripts) into a common target language (e.g., English), and we can proceed classifier learning and evaluation protocols as if there were no multiple languages. The advantage of this approach is that number of training examples in English will be abundant. The disadvantage, however, is that automatic translation systems inevitably introduce errors in addition to errors from automatic speech recognition systems. To leverage abundant training examples and discriminative power from native languages, we explored multilingual text features for learning text-based visual classifiers. Based on our experience, the parameter setting of SVM is critical to the performance. Therefore, we perform grid search of the parameter space using cross-validation to find the optimal parameters for each concept in the training set, particularly the gamma parameter of the kernel function and the cost parameter.

4. A hybrid approach

4.1. Importance of Classifier Tuning

Our basic approach uses support vector machines (SVM) with radial basis kernel function (RBF) on the training set to train baseline classifiers for all concepts based on various combinations of low-level features. are used in the training of baseline classifiers.

Table 1 shows that the optimal parameter setting achieves an average of 27% improvement (0.2633 to 0.3352) over the default setting in terms of mean average precision (MAP) on 39 concepts.

Table 1: Comparison between default and optimal SVM parameters.

Semantic Concepts	MAP	
	SVM-Default Parameters	SVM Optimal Parameters
Airplane	0.0135	0.1469
Animal	0.3863	0.4978
Boat/Ship	0.2131	0.1699
Building	0.3048	0.3481
Bus	0.0088	0.0778
Car	0.3151	0.4458
Charts	0.1265	0.1815
Computer/TV-screen	0.3525	0.4971
Corp.-Leader	0.0059	0.0103
Court	0.0879	0.1882
Crowd	0.5288	0.5818
Desert	0.0602	0.109
Entertainment	0.0975	0.2999
Explosion/Fire	0.1413	0.2504
Face	0.7752	0.8634
Flag-US	0.1227	0.1344
Gov'nt-Leader	0.1822	0.2672
Maps	0.4816	0.4805
Meeting	0.1708	0.2578
Military	0.2049	0.2711
Mountain	0.1718	0.2512
Natural-Disaster	0.0403	0.0521
Office	0.0895	0.1181
Outdoor	0.4816	0.7954
People-arching	0.0759	0.1695
Person	0.8531	0.9004
Police/Security	0.0078	0.0121
Prisoner	0.1693	0.1546
Road	0.2481	0.3023
Sky	0.6502	0.6526
Snow	0.1725	0.2232
Sports	0.4481	0.5478
Studio	0.7541	0.8389
Truck	0.0251	0.0341
Urban	0.1127	0.1651
Vegetation	0.3203	0.3969
Walk/Run	0.1635	0.2491
Waterscape/ Waterfront	0.3171	0.4421
Weather	0.5887	0.6869
Average	**0.2633**	**0.3351**

Table 1 shows how the optimal SVM parameters provide improvements for each individual feature set over the default parameters in the fusion set based only on color moment feature. Of the 39 semantic concepts, 37 improved as a result, one (Maps) was virtually unchanged, and only one (Boats/Ships) decreased due to overfitting. The results underscore the strong need for careful tuning and parameter normalization.

4.2. Using Multiple Features

In the experiments with the TRECVID 2006 feature classification data we also explored the use of image local features as an alterative of the global color/texture features for detecting semantic concepts in video data. Local feature points can capture aspects of an object in a picture, and are often less sensitive to variations in lighting and viewpoint. Local features describe the regions around the salient keypoints detected in an image. We propose to explore a text categorization approach to the problem of shot classification based on vector-quantized keypoint features or visual-word features. That is, we treat visual words in images as words in documents, and apply techniques widely used in text categorization (or generally, in information retrieval) to the concept classification problems. These include choosing vocabulary size, feature weighting methods such as tf and tf-idf, stop word removal, and so on. These techniques seek for the most effective bag-of-word representation for text categorization, and in this case the most effective "bag of visual words" representation for scene classification. Therefore, a major contribution of this section is to provide the beginnings of a comparative study of various implementation choices related to image representation based on local keypoint features. Although some of these techniques have been already adopted in scene classification, such as stop word removal and tf-idf weighting [22, 24], their effectiveness has been so far taken for granted without empirical evidence showing that they indeed enhance the performance.

Each image is represented as an unordered collection of real-valued keypoint descriptors with varying cardinality. This representation, however, creates difficulties for supervised classifiers which demand feature vectors of fixed dimension as input. The solution is to cluster the keypoint descriptors in their feature space into a large number of clusters using clustering algorithms such as K-means [25], and encode each keypoint by the index (an integer) of the cluster it is assigned to. This process is described as the generation of a vocabulary (or codebook), where the index of each cluster can be seen as a visual word in

the vocabulary. Each image can be thus represented by a histogram-like vector of the count of each visual word in the image (i.e., the number of keypoints in each cluster). The dimension of this feature is determined by the number of clusters, or the vocabulary size, which usually varies from hundreds to tens of thousands or even more. In this way, we transform descriptors of image keypoints into a discrete, high-dimensional "bag of visual words" representation of the whole image, which is analogous to the "bag-of-words" representation of text documents.

Given its similarity to the "bag-of-keywords" representation of text documents, we applied text categorization methods for classifying video data by the presence (or absence) of semantic concepts, and studied the influence of feature dimension, weighting and normalization, feature selection, spatial information to the classification performance. Experiments show that using local features achieves comparable performance to that of the global features, and significantly higher performance when these two types of feature are used together.

In a text corpus, the size of word vocabulary is determined by the language, while for images the size of the visual word vocabulary is specified as the number of keypoint clusters in the vocabulary generation process. Choosing the right vocabulary size involves the trade-off between the discriminative power of the feature and its generalization ability. When a small vocabulary is used, the resulting visual-keyword feature lacks discriminative power because two keypoints can be assigned into the same cluster, even if they are not very similar. As the vocabulary size increases, the feature becomes more discriminative but also less generalizable and forgiving to noises, since similar keypoints can be assigned to different clusters. we experiment with vocabulary containing 200, 1000, 5000, 20000, 80000, and 320,000 visual words, which cover most of the vocabulary sizes ever used in existing work. Note that even 320,000 is not terribly huge as the number of clusters given that the dimension of the keypoint descriptor space. A single partition at each dimension of the original descriptor space will result in 236 clusters for the 36-dimensional PCA-SIFT features, or 2128 clusters for the 128-dimensional SIFT features.

The main observation, as summarized in Table 2, is that the performance of scene classification improves significantly as the vocabulary size (or feature dimension) increases. The MAP achieved by linear-kernel SVM almost triples when the vocabulary sizes increases from 200 to 80,000 or 320,000. The increase with RBF-kernel SVM is not as dramatic but still remarkable. The performance starts to level off or even slightly drop after the vocabulary size reaches 80,000 for linear kernel or 20,000 for RBF kernel.

Table 2: The MAP of concept classification using region-based visual-term features computed at various spatial partitions. The percentage in the parenthesis shows the relative improvement over the performance at 1x1 partition.

Vocabulary Size	Spatial Partitioning			
	1x1	2x2	3x3	4x4
200 (RBF SVM)	0.137	0.258 (+89%)	0.267 (+95%)	0.272 (+99%)
1,000 (RBF SVM)	0.235	0.249 (+6%)	0.291 (+24%)	0.286 (+22%)
5,000 (RBF SVM)	0.245	0.279 (+14%)	0.285 (+16%)	0.268 (+9%)
20,000 (RBF/Linear)	0.271	0.280 (+3%)	0.290 (+7%)	0.293 (+8%)
80,000 (Linear SVM)	0.280	0.290 (+4%)	0.290 (+4%)	0,288 (+3%)

Interesting observations can be made by comparing the performance of the two kernel functions. For small vocabularies, the RBF kernel has a clear advantage over the linear one, but this advantage is reversed once the vocabulary size reaches 80,000. This suggests that the visual words in a small vocabulary are highly correlated, but they become more independent and gain the nice property of linear separability (of data) as the vocabulary gets larger. Finally, the results of combining local features represented as visual words and the more standard color/texture features can be found in Table 3.

Practical insights emerge from our experiments. Some are consistent with the findings in text categorization, some are not. Some of the common implementation choices in scene classification are shown to be ineffective. Overall, we find these representation issues critical to the scene classification performance: 1) a vocabulary much larger than the ones currently used is preferred; 2) binary features that indicate the presence/absence of visual words are as effective as tf or tf-idf features that encode the word count information; 3) normalizing the feature vector into unit length hurts the classification performance; 4) frequent visual words are not "stop words" but the informative ones; 5) feature selection can reduce the vocabulary by more than half without loss of performance; 5) the benefit of spatial information is much more significant with small vocabularies than with large vocabularies.

Our experiments yielded deeper insights into these findings by exploring their connections with the properties of visual words. We find the distribution of visual words in a video corpus bears many similarities yet important differences to the word distribution in a

text corpus. This explains some of our experiment results, such as why there are no "stop visual words" and why feature selection can reduce the vocabulary size without hurting the performance. We also show that the classification performance of local keypoint features (visual words) is comparable to that of global color/texture feature, and combining the two features leads to a further improvement of 10-20%.

4.3. Exploiting Multiple-Concept Relationships

Previous experience has shown that the semantic concepts are not independent of each other. Thus, exploiting relationships between multiple semantic concepts in video could be an effective approach to enhancing the concept detection performance. In TRECVID 2006, we tried to use a multi-concept fusion technology called Multiple Discriminative Random Fields (MDRF).

Figure 1 illustrates the framework of MDRF on top of a single video shot, which consists of three different semantic concepts, such as *"building"*, *"tree"*, and *"sky"*. We construct an undirected graphical model to represent the relationships between concepts and the video shot, and also the relationships between various concepts. Figure 2 illustrates such a graphical model. Mathematically, MDRF is stated as:

$$p(Y \mid X) = \frac{1}{Z} \exp\left(\sum_{i \in S} A_i(y_i, W, X) + \sum_{i \in S} \sum_{j \in N_i} I_{ij}(y_i, y_j, V, X) \right) \quad (1)$$

where $Y = (y_1, y_2, \ldots, y_n)$ is the vector of multiple concept labels, with y_i denoting the label of ith concept. In this work, each semantic concept is either present or absent in the shot, i.e. $y_1 = \{-1, 1\}$. X (in R^c) is the observation or feature vector extracted from the video shot. $A_i(y_i, W, X)$ is called the *association potential* function. In MDRF, association potential provides links between concept labels and observation, as a normal classifier does. $I_{ij}(y_i, y_j, V, X)$ is the *interaction potential* function. The interaction potential tries to model the interactions between various concepts with observation. For example, if there are some shots in training set that have both the "*sky*" and "*tree*" concept, the bluish and greenish color feature (which are typical for the two concepts) will be emphasized in the learning process via the interaction potential. When a new shot comes out with big blue which is easy to be recognized with unclear green area, the tree detector will benefit from the interaction potential to detect the tree concept. $\theta = \{W, V\}$ are the parameters of the model. W is the parameter of the association potential, and V is the parameter of the interaction potential. In Eq.(1), the summation of association potentials corresponds to the set of individual classifiers for each concept, and the summation of interaction potentials models the relationship of each concept pair.

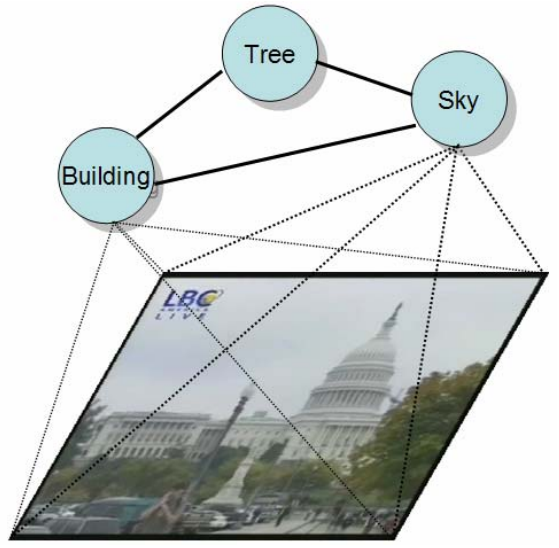

Figure 1: A graph demonstrates the framework of MDRF. There are three semantic concepts in this video shot: building, tree and sky. The top layer shows the concepts relations with each other and constitutes an undirected graph. The edges between each concept can be viewed as interaction potentials in the MDRF formula. The dotted lines from concepts to the video shot illustrate the classifications of each concept which act as association potentials in the MDRF model. In the MDRF model, concepts are denoted as variable y and a video shot is denoted as observation X.

In Figure 2, we can interpret MDRF as a fully connected undirected graphical model. There are 3 concepts as Y_1, Y_2 *and* Y_3 that are linked to each other as well as to the observation X. The linkages between concepts encode the interaction potentials in MDRF and the linkages to observation encode the association potentials.

We predict the validation set and test set by the models built from our training set. For shots in the validation set and testing set, predictions become our observations for this shot. To be clearer, we have 39 different concepts and every concept has 4 different modalities. The observation is a 156-dimension vector (39x6). We adopt logistic function as association potential:

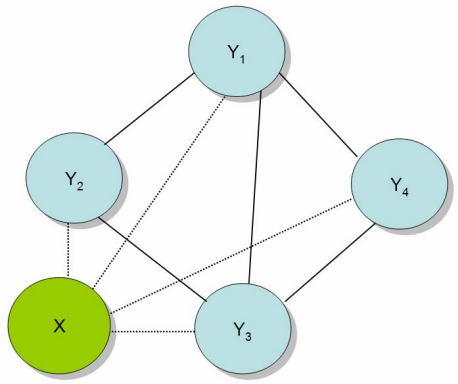

Figure 2: MDRF is a fully connected undirected graphical model. Y nodes denote the semantic concepts. X is the observation extracted from the video. All concepts are dependent on the observation.

$$A_i(y_i, W, X) = \log(p(y_i \mid X)) = \log(\sigma(y_i w_i^T h_i(X))) \quad (2)$$

$$I(y_i, y_j, V, X) = y_i y_j V_{ij}^T u_{ij}(X) \quad (3)$$

From Eq.(2), we know the association potential works like a logistic regression classifier which outputs the probability of label given the observation. Eq.(3) shows the interaction potential function. $u_{ij}(X)$ can be any function to deal with the observation. V is the parameter of interaction potential, which emphasizes the agreement between two concepts and searches the observation that supports the agreement.

Table 4: MDRF for semantic concept extraction

Runs	MAP
SVM, multi-modal feature (baseline)	0.146
MDRF with chi-square feature selection	0.148
MDRF without chi-square feature selection	0.114

Table 4 shows the performance of MDRF in TRECVID 2006 submission in comparison with a SVM approach that does not consider inter-concept relationships. We use feature selection method based on chi-square statistics to filter out some concept pairs which are not related in order to remove noises from the model. We discovered that even when the threshold of chi-square statistics is set as small as 0.05, very few concepts in 39 concept corpus connected to each other. Not many concepts are related to each other in TRECVID 2006 set, and we so didn't obtain a significant improvement by considering the multi-concept relationships. We also found that chi-square feature selection is critical since without it the performance was much worse.

4.4. Multi-modal Feature Combination

Multiple types of low-level features need to be combined in an effective way to provide better performance than any single type of features.

Monolingual text features are a bag-of-words representation of words spoken in a shot of dimensions of VE, where VE is the vocabulary size of English. Multilingual text features, on the other hand, contain both native languages and translations (e.g., Chinese and English translation), and is of the dimension VE + VC + VA, where VC and VA are the vocabulary sizes of Chinese and Arabic, respectively. We built text classifiers on this multi-lingual feature using SVM with a linear kernel. We evaluated the proposed multilingual text features on the development set of TRECVID'06. Experimental results showed that multilingual text features were remarkably more effective than monolingual text features (i.e., English only). Multilingual run improved the mean average precision (MAP) of the 39 concepts from 0.134 to 0.175 (30% improvement) on the held-out development-test set. Contrasting two runs in our officially evaluated submission also shows multilingual text features consistently perform better than monolingual text features (see Table 6 for the official results). In addition to the ASR transcripts and translations by provided by NIST, text features were also obtained using the SAIL Labs [www.sail-technology.com] speech recognition engine for English and Arabic speech recognition. The Arabic transcripts were further translated into English using Google translation [www.google.com/translate_t] through automated scripts.

To fuse results from different classifiers using different techniques, we adopted a mixture of the early fusion and late fusion strategy. To color and texture features are stacked into a large feature vector of 273 dimensions (i.e., early fusion) due to their low dimensionality and close relationships. In contrast, we use late fusion strategy to combine this color-texture feature with the local feature and the textual feature. Specifically, we train SVM classifier for each concept based on each type of feature, and apply the trained classifiers to predict the label of each shot in the testing set. Therefore, for any shot, there will be predictions based on color-texture feature, local feature, and text feature, respectively. We train meta-level classifiers using logistic regression or SVM, which take the component prediction scores as input and output an overall prediction. Table 5 shows the comparison between the two meta-level classifiers with different low-level features. Clearly, logistic regression outperforms SVM as a meta-level classifier in this

data. We thus choose logistic regression to fuse predictions from multi-modal features, shown in Table 5. Due to space considerations, we do not list the individual concept results. However, for some concepts the multi-language linguistic features provided a substantial benefit, while for many others there was no improvement, and for some the results were worse.

5. Conclusions

Unfortunately, the experiments presented here do not lend themselves to one simple conclusion. The unfortunate fact is that there is no one approach that consistently outperforms others on all concepts and data sets. In fact, it is likely that our quest for the one cure-all approach is doomed to failure. However, this does not mean we should stop trying. Each of the successful comparisons points to some technique or trick that can play a role for some concept in some dataset. The research, as results suggested, should be focused on uncovering as many techniques as possible, and to leave it as an engineering exercise to determine which combinations of techniques appears to work, based on empirical evidence for a given set of concepts and the specific collection characteristics. This is the approach of the Pathfinder system [26] and others. [1] who try different approaches and select the specific approaches on a concept by concept basis. Some of the proposed methods, such as MDRF, or SVMs with complex kernels, large feature vectors and large training data sets, are very computationally intensive at the model building step, while others are quite cheap to apply.

Table 5: Comparison of logistic regression and SVM for multi-modality fusion

Multi-modality Runs	MAP
Logistic regression (color-texture + local +monolingual text)	0.146
SVM (color-texture + local +monolingual text)	0.121
Logistic regression (color-texture + local +multilingual text)	0.153
SVM (color-texture + local + multilingual text)	0.126

A long-term research goal is to devise methods for predicting for a particular concept and data combination which combination approaches are most likely to yield the best results, without empirically trying all possible methods. This grand scientific goal would then also result in an explanation why some methods work for a specific concept and some don't.

We have presented ideas of techniques that contribute to improved detection performance. It is our hope that by establishing the synergy between them substantial progress is possible. Current detection rates are still low for many concepts, but there is hope [27] that even this limited detection accuracy with large numbers of concepts will be sufficient for substantial help with concept-based video retrieval.

9. Acknowledgements

This material is based in part on work supported by the National Science Foundation Grant IIS-0205219.

10. References

1. J. Cao, et al. *Intelligent Multimedia Group of Tsinghua University at TRECVID 2006*. in *TRECVID Video Retrieval Evaluation*. 2006. Gaithersburg, MD, USA: NIST.
2. Smeulders, A.W.M., et al., *Content based Image Retrieval at the End of the Early Years.* IEEE Trans. Pattern Analysis and Machine Intelligence, 2000. **22**(12): p. 1349-1380.
3. Smith, J.R., Basu, S., Lin, C-Y., Naphade, M. and Tseng, B. *Interactive Content-based Retrieval of Video*. in *IEEE International Conference on Image Processing (ICIP)*. 2002. Rochester, NY.
4. Rodden, K., et al. *Does organisation by similarity assist image browsing?* in *CHI '01: Proceedings of the SIGCHI conference on Human factors in computing systems*. 2001. New York, NY, USA,: ACM Press.
5. Markkula, M. and E. Sormunen, *End-User Searching Challenges: Indexing Practices in the Digital Newspaper Photo Archive.* Information Retrieval, 2000. **1**(4): p. 259-285.
6. Hauptmann, A., R. Yan, and W.-H. Lin. *How many high-level concepts will fill the semantic gap in news video retrieval?* in *International Conference on Image and Video Retrieval (CIVR)*. 2007. Amsterdam, The Netherlands.
7. Chang, S.F., R. Manmatha, and T.S. Chua. *Combining text and audio-visual features in video indexing.* in *IEEE ICASSP'05*. 2005.
8. Barnard, K., et al., *Matching words and pictures.* Journal of Machine Learning Research, 2002. **3**.
9. Naphade, M.R.a.H., T.S. *Semantic Video Indexing using a Probabilistic Framework*. in *I.E.E.E. International Conference on Image Processing*. 1998. Chicago, Il.
10. Lin, C.-Y., B.L. Tseng, and M. Naphade. *VideoAL: A Novel End-to-End MPEG-7 Automatic Labeling System*. in *IEEE Intl. Conf. on Image Processing*. 2003. Barcelona.
11. Lin, W. and A. Hauptmann. *News Video Classification Using SVM-based Multimodal Classifiers and Combination Strategies*. in *ACM Multimedia 2002*. 2002. Juan-les-Pins, France.
12. Jeon, J., V. Lavrenko, and R. Manmatha. *Automatic image annotation and retrieval using cross-media relevance models*. in *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in information retrieval*. 2003.
13. Wu, Y., et al. *Optimal multimodal fusion for multimedia data analysis*. in *Proceedings of the 12th annual ACM international conference on Multimedia*. 2004.

14. Hauptmann, A., et al. *Informedia at TRECVID 2003: Analyzing and Searching Broadcast News Video*. in *Proceedings of (VIDEO) TREC 2003 (Twelfth Text Retrieval Conference)*. 2003. Gaithersburg, MD.
15. Natsev, A.P., M.R. Naphade, and J. Te¡si´c. *Learning the semantics of multimedia queries and concepts from a small number of examples*. in *Proceedings of the 13th ACM International Conference on Multimedia,*. 2005.
16. Over, P. *TRECVID: TREC Video Retrieval Evaluation*. 2007. http://www-nlpir.nist.gov/projects/t01v/.
17. Smeaton, A.F. and P. Over, *The TREC-2002 Video Track Report*. 2002.
18. Over, P., et al. *TRECVID 2006 - An Overview*. in *TRECVID'06 Video Retrieval Evaluation*. 2006. Gaithersburg, MD: NIST.
19. Kennedy, L. and A. Hauptmann, *LSCOM Lexicon Definitions and Annotations Version 1.0, DTO Challenge Workshop on Large Scale Concept Ontology for Multimedia."* Columbia University: New York.
20. Chang, S.-F., et al. *Columbia University TRECVID-2006 Video Search and High-Level Feature Extraction*. in *TRECVID Video Retrieval Evaluation*. Gaithersburg, MD: NIST.
21. Yanagawa, A., W. Hsu, and S.-F. Chang, *Brief Descriptions of Visual Features for Baseline TRECVID Concept Detectors*. 2006, Columbia University: New York, NY.
22. W. Zhao, Y.G. Jiang, and C.W. Ngo. *Keyframe retrieval by keypoints: Can point-to-point matching help?* in *International Conf. on Image and Video Retrieval*. 2006.
23. Lowe, D.G., *Distinctive image features from scale-invariant keypoints.* International Journal of Computer Vision, 2004. **60**(2): p. 91-110.
24. Sivic, J. and A. Zisserman. *Video Google: A Text Retrieval Approach to Object Matching in Videos*. in *Ninth International Conference on Computer Vision (ICCV'03)*. 2003. Nice, France.
25. D. Pelleg and A. Moore. *X-means: Extending k-means with efficient estimation of the number of clusters*. in *Proc. of the 7th Int'l Conf. on Machine Learning (ICML)*. 2000. San Francisco: Morgan Kaufmann.
26. Snoek, C., M. Worring, and A.G. Hauptmann, *Learning rich semantics from news video archives by style analysis.* TOMCCAP, 2006. **2**(2): p. 91-108.
27. Hauptmann, A., et al., *Can high-level concepts fill the semantic gap in video retrieval? A case study with broadcast news.* IEEE Transactions on Multimedia Journal, 2007.

Table 6: **The high-level semantic concept extraction as evaluated by NIST**

Method	Features	Official Result Mean Average Precision
SVM, multi-modality (early fusion)	color, texture	0.099
SVM, multi-modality (late fusion)	color, texture, local feature, monolingual text	0.146
MDRF without χ^2 selection	color, texture, local feature, monolingual text	0.114
MDRF with χ^2 selection	color, texture, local feature, monolingual text	0.148
SVM, multi-modality (late fusion)	color, texture, local feature, multilingual text	0.153
Borda voting	color, texture, local feature, multilingual & monolingual text	0.159