J Sign Process Syst
DOI 10.1007/s11265-009-0382-z

A Multi-Pronged Approach to Improving Semantic Extraction of News Video

A. G. Hauptmann · M.-Y. Chen · M. Christel · W.-H. Lin · J. Yang

Received: 26 May 2009 / Revised: 26 May 2009 / Accepted: 26 May 2009

Abstract In this paper we describe a multi-strategy 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. We present a discriminative learning framework called Multi-concept Discriminative Random Field (MDRF) for building probabilistic models of video semantic concept detectors by incorporating related concepts as well as the low-level observations. The model exploits the power of discriminative graphical models to simultaneously capture the associations of concept with observed data and the interactions between related concepts. Compared with previous methods, this model not only captures the co-occurrence between concepts but also incorporates the raw data observations into a unified framework. We also describe an approximate parameter estimation algorithm and present results obtained from the TRECVID 2006 data. No single approach, however, provides a consistently better result for all concept detection tasks, which suggests that extracting video semantics should exploit multiple resources and techniques rather than naively relying on a single approach

Keywords Video analysis · Discriminative random fields · Semantic concept detection · Undirected graphical models

A. G. Hauptmann (✉) · M.-Y. Chen · M. Christel · W.-H. Lin · J. Yang
School of Computer Science, Carnegie Mellon University, Pittsburgh, PA, USA
e-mail: alex+@cs.cmu.edu

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

Published online: 29 July 2009

improving video concept extraction. The 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 multi-facetted approach, unlike previous work that focus on only developing better features or fusing techniques, may provide a more holistic answer to the question how we can achieve the most improvement for extracting video semantics? We tested the described multi-strategy approach on a well-established testbed, TRECVID [16] (Section 2) using common low-level features (Section 3). The findings and future direction are summarized in Section 5.

2 The TRECVID Semantic Concept Detection Task

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 [17][Smeaton06]. In 2001, NIST started the TREC Video Track (now referred to as TRECVID [18]) 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 [19]. 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, 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) [20] 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 [21] 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. [22]
- **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 [23] provided by City University of Hong Kong, which are detected using the DoG detector and depicted by SIFT descriptors [24]. 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.

4 A Multi-Pronged Approach

4.1 Importance of Classifier Tuning

Our basic approach is to use the training set to train our baseline classifiers for all concepts based on various combinations of low-level features. Support vector machines (SVM) with radial basis kernel function (RBF) are used in the training of baseline 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. As shown in Table 1, using the optimal parameter setting achieves an average of 27% improvement (0.2633 to 0.3352) over the default setting in terms of the mean average precision (MAP) metric on the 39 concepts in the cross-validation experiment. 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.

Table 1 Comparison between default SVM parameters and the 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
Corporate-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
Government-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-Marching	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
Walking/Running	0.1635	0.2491
Waterscape/Waterfront	0.3171	0.4421
Weather	0.5887	0.6869
Average	0.2633	0.3351

4.2 Using Multiple Feature Sets

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 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 [23, 25], 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 [26], 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 2^{128} 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.

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

Table 2 The MAP of concept classification using region-based visual-term features computed at various spatial partitions.

Vocabulary Size	Spatial Partitioning			
	1×1	2×2	3×3	4×4
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 SVM)	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%)

The percentage in the parenthesis shows the relative improvement over the performance at 1x1 partition.

Table 3 MAP for concept classification of various global features, local features based on 3x3 image grid, and their combinations.

		Visual Words (3x3 Partition)					
		Vocab	200	1,000	5,000	20,000	80,000
		Baseline MAP	0.267	0.291	0.285	0.289	0.290
Global Feature	Color	0.250	0.295	0.301	0.321	0.334	0.334
	Gabor	0.182	0.273	0.293	0.286	0.291	0.315
	Color+Gabor	0.292	0.318	0.329	0.339	0.349	0.349

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

The most common approaches to concept detection translate the concept learning task into a set of binary one-versus-all classification problems with a presence/absence label for each individual concept, thereby decoupling any connection between semantic concepts. Then, for each video shot, its associated video concepts can be detected using unimodal or multimodal classifiers based on visual, audio and speech-transcript features. These binary classification approaches thus assume independence between concepts and ignore the important fact that semantic concepts do not exist in isolation to each other.

However, when looking at the data, it becomes obvious that the semantic concepts to be detected are not independent to each other. They are interrelated and connected by their semantic interpretations and hence exhibit a certain co-occurrence pattern in the video collection. For example, the concept "sky" usually co-occurs in a video shot with the concept "outdoor" while the concept "studio" is not likely to appear together with "sky". Such kinds of concept relationships are quite frequent and mining these multi-concept relationships could provide a useful source of information to improve concept detection accuracy. Moreover, such a correlated context could also be used to automatically construct a semantic network or ontology derived from the video collection in a bottom-up manner. This automatic ontology construction may be helpful to discover unknown concept relationships that could be complementary to manually specified ontologies.

To automatically exploit multi-concept relationships, several approaches have been proposed, which build upon advanced pattern recognition techniques within a probabilistic framework. For example, Naphade [9] et al. explicitly modeled the linkage between various semantic concepts via a Bayesian network that offers an ontology semantics underlying of the video collection. Snoek et al. [27, 28] proposed a semantic value chain architecture including a multi-concept learning layer called the context link. At the top level, this tries to merge the results of detection output from different concept detectors. Two configurations were explored: one was based on a stacked classifier on top of a context vector and the other was based on an ontology with certain common sense rules. Hauptmann et al. [29] fused the multi-concept predictions and captured the inter-concept causation by constructing an additional logistic regression classifier on top of the single concept detection results. Amir et al. [30] concatenated the concept prediction scores into a long vector called a model vector and stacked a support vector machine on top to learn a binary classification for each concept. An ontology-based multi-classification algorithm was proposed by Wu et al. [13] which attempted to model possible influence relations between concepts based on a predefined ontology hierarchy.

One direction that has not been explored much, despite the much work on learning multiple concept detectors of video, are undirected probabilistic graphical models. These probabilistic graphic models provide an alternate, elegant approach to handle the semantic concept detection problem. In computer vision, researchers have started to utilize contextual information to enhance pattern recognition performance. This is similar to the idea of using related concepts to boost multi-concept detection performance. Markov Random Field (MRF) [31] is a commonly used model in computer vision to utilize contextual information.

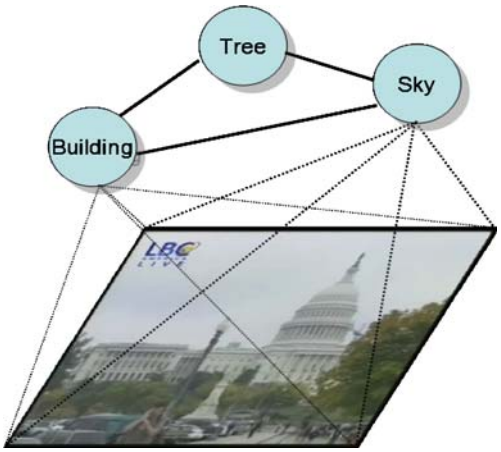


Figure 1 A graph demonstrates the framework of MDRF. There are three semantic concepts in this video shot: building, tree and sky. The top layer shows the concepts relations with each other and constitutes an undirected graph. The edges between each concept can be viewed as interaction potentials in 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.

MRFs are generally used in a probabilistic generative framework that models the joint probability of the observed data and the corresponding labels. However, for classification purposes, we are more interested in estimating the posterior probability over labels given the observation rather than the joint probability. Conditional Random Fields (CRF) [32] are a conditional probabilistic graphical model for segmenting and labeling sequence data. It specifies the probabilities of the possible label sequences given an observation sequence. Because the conditional probabilities of the label sequence depend on the observation sequence, any arbitrary/non-independent features can be derived from the observation sequence, without forcing the model to account for the distribution of these dependencies. Therefore, CRF provide a new type of random field model that incorporate the dependency among observations rather than single matches. Discriminative random fields (DRF) [33] provide a more advanced model to jump from a 1-D sequence dependency to a 2-D spatial dependency. DRF was first proposed for structure detection in natural images. The model has two major building blocks: an association term that consists of local discriminative models to capture the association between observations and labels for each individual node and an interaction term that exploits pair-wise co-classification within nearby nodes. DRF uses local discriminative models to model interactions in both the observed data and the labels in a principled manner. This means the classification result will derive from not only the observation for a certain node but also the context nearby.

Thus, exploiting relationships between multiple semantic concepts in video could be an effective approach to enhancing the concept detection performance. Using the TRECVID 2006 evaluation data, we developed and tested a multi-concept fusion technique called Multiple Discriminative Random Field (MDRF) to detect multiple semantic concepts of the video. Our model expands DRF by jointly learning multiple classifications instead of a single classifier at a time. Our MDRF introduces several new aspects compared to previous work. First, MDRF is an undirected graph model which does not require prior causation analysis to understand the dependencies between concepts [9, 27, 30]. Second, MDRF learns both classification and interaction simultaneously. No additional training data is required for concept linkage discovery, unlike [9, 27, 29]. Third, the concept interactions are also linked to direct raw data observations which allows the model to not only capture the co-occurrence between concepts but also learn pair-wise relations in the low-level feature space [34]. In the following paragraphs, we will present the MDRF model with its parameter learning and inference, followed by evaluation results on the TRECVID 2006 test.

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 assume that low-level features describe the video shot in its entirety, without spatial or temporal segmentation. 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. In this model, we use a set of pair-wise linkages between concept nodes to model the inter-concept relationship and associate the observed low-level features with every single concept node, so that the conceptual relationship and low-feature modeling can be

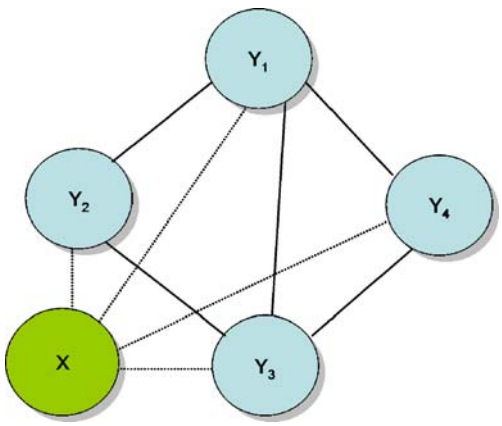


Figure 2 MDRF is a fully connected undirected graphical model. Y nodes denote the semantic concepts. X is the observation extracted from the video. All concepts are dependent on the observation.

jointly optimized in such a unified framework. Based on this graphical model, the conditional probability of the labels Y given the observations X can be written as:

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

where $Y=(y_1, y_2, ..., y_n)$ 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, the 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 the training set that have both the "*sky*" and "*tree"* concept, the bluish and greenish color feature (which are typical for the two concepts) might 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.

Now we can take a closer look at Fig. 2. In Fig. 2, we could 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 potential in MDRF and the linkages to the observations encode the association potential.

4.3.1 The Association Potential

The association potential function works like a single concept classifier. In this paper, we use discriminative models in the association potential function to achieve our classification goal. Theoretically, $A_i(y_i, W, X)$ can be any model which functions as a classification. This provides the flexibility so people can choose specific models for certain domain problems. In MDRF, we try to model the classification using local discriminative models which output a label y_i given the association with observation X. Thus the association potential function can be written in a conditional probability format,

$$A_i(y_i, W, X) = \log(p(y_i|X)) \quad (2)$$

where the log function here maps the probability value to a real number.

In our work, a logistic model then serves as our basic classification model, since it is a well-known and efficient discriminative model for estimating the posterior probability given the observation. Therefore, for each individual concept, the posterior probability can be written with the logistic formula,

$$p(y_i = 1|X) = \frac{1}{1+e^{-\left(w_{i0}+w_{i1}^T h_i(X)\right)}} = \sigma\left(w_{i0} + w_{i1}^T h_i(X)\right) \quad (3)$$

where $w_i=\{w_{i0}, w_{i1}\}$ is the parameter for the i^{th} semantic concept classification and the $h_i(X)$ function maps the observation to the feature space as a vector. The mapping function also provides the flexibility to allow dimensionality reduction and other transforms to enhance the representation of the observations, for example, if the h function is a nonlinear function, this will extend the logistic model to model a nonlinear decision boundary in the feature space. W_{i0} here works as a constant term to stabilize the logistic function. y_i is a binary label as $\{-1, 1\}$. Therefore, we can add an additional constant term into the transformed vector and then re-write association potential function by combining (2) and (3) as,

$$\begin{aligned} A_i(y_i, W, X) &= \log(p(y_i|X)) \\ &= \log\left(\sigma\left(y_i w_i^T h_i(X)\right)\right) \end{aligned} \quad (4)$$

If the interaction potential function is defined as zero, MDRF is equivalent to building n individual logistic regressions for each concept. This illustrates how the association potential function plays the role of capturing the association between observations and labels which originally define the classification.

In our work, we apply Principle Component Analysis (PCA) as the $h_i(X)$ function mainly for the purpose of dimensionality reduction. However, in principle, any mapping or transformation function can be used here. In computer vision, researchers often use a kernel function to utilize contextual information. In the multimedia domain, multi-modality fusion can be performed at this point to improve the power of the model to capture more complex aspects.

4.3.2 The Interaction Potential

The interaction potential function plays an important role in the MDRF model. This potential expands the model to

utilize pair-wise relationships that cooperate with the observation. It can be seen as a measure of how concepts *i* and *j* which are related should interact with each other given the observed video shot. As an example, if our concept set contains *sky* and *building*, the sky detector might emphasize the color features while a building detector could emphasize edge features. If a detector detects some blue colors in a shot, there is a high possibility for this shot to contain a view of the sky. If there are some vertical edges, the shot has high possibility to include buildings. However, the interaction potential here can learn a model which contains both color and edge features to predict a co-occurrence of sky and building.

To design the interaction term, we borrow the commonly used form from the MRF model, $I=\alpha y_i yj$, which is a smoothing term that penalizes every dissimilar pair of labels. However, in the MRF framework, this does not permit the use of observations and the interaction term turns out to model the co-occurrence only. Therefore, in MDRF, we define the interaction potential function as,

$$I(y_i, y_j, V, X) = y_i y_j V_{ij}^T u_{ij}(X) \tag{5}$$

with parameter V and function $u_{ij}(X)$ converting an observation into a feature vector. Similar to the $h_i(X)$ function in the association potential, $u_{ij}(X)$ can be designed for specific usage in different domains. In our work, we use the same function as $h_i(X)$ in the association potential, i.e., the PCA function, to reduce the dimensionality.

This form of interaction potential models the agreement or disagreement between related concepts. The potential function tries to capture the observations which support agreement between two concepts and learns the model of this pair-wise incorporation. Ideally, if there are enough training data, the parameters should emphasize strongly related concept pairs. Therefore, MDRF is designed as a fully connected, undirected graphical model. We hope this type of model captures useful linkages between concepts that may not be obvious to a human but useful in classification. However, MDRF can always be applied after co-occurrence analysis from another source to eliminate very weak linkages between concepts. The fully connected graph has an exponential number of interactions based on the number of concepts; therefore, the flexibility to choose related concepts makes the model more efficient. As mentioned earlier, once we set those pair-wise potentials to zeros, the model will act just like a traditional semantic concept detector. In that case, the model is not able to capture the interactions between concepts and instead makes the assumption that each concept is independent. In that case, MDRF acts as a logistic classifier which calculates the conditional probability of each concept given the observation. The interaction potential function comes from the MRF model which is a generalization of the MDRF model. In the MRF model, this term works as a smoothing term to absorb the errors if the classification terms. Although our interaction potential becomes observation dependent, it can still perform as a smoothing term to absorb errors of the association potential. Moreover, we can also set up a penalization for parameter V if we expected the association potential to have better classification power than the interaction potential. This will make the model emphasize the association potential more and decrease the effect of the interaction potential. It also makes the model more flexible for application in different domains.

4.3.3 Parameter Estimation

In our MDRF model, we have two parameters, *V* and *W*, to be estimated. The parameter learning process of MDRF will learn both parameters simultaneously, which is a very attractive feature of our approach. MDRF is able to achieve the goal of training multiple classifications and modeling the relationships between them simultaneously and does not require additional data or further splitting of the training data to obtain an additional held-out set for the combination step. With the association and interaction potential functions defined as above, the MDRF can be written as,

$$P(Y|X,\theta) = \tfrac{1}{Z}\exp\left(\sum_{i\in S}\log\left(\sigma\left(y_i W_i^T h_i(X)\right)\right) + \sum_{i\in S}\sum_{j\in Ni} y_i y_j V_{ij}^T u_{ij}(x)\right)$$
$$\text{where } Z = \sum_{Y'}\exp\left\{\sum_{i\in S}\log\left(\sigma\left(y_i W_i^T h_i(X)\right)\right) + \sum_{i\in S}\sum_{j\in Ni} y_i y_j V_{ij}^T u_{ij}(x)\right\} \tag{6}$$

Y' here is denoted as all possible combination of vector Y which means 2^n combinations of possible concept sets, *n* is the number of the concepts. *Z* is the partition function which normalizes the exponential value to a probability.

Maximum likelihood estimation with a gradient descent approach is commonly used in undirected graphical model learning process. Given that we have *M* shots in our video collection, the log-likelihood of MDRF model is written as,

$$P(Y|X) = \prod_{m=1}^{M} P(Y^m|X^m)$$
$$l(\theta) = \log(P(Y|X)) = \sum_{m=1}^{M}\log(P(Y^m|X^m))$$
$$= \sum_{m=1}^{M}\left\{\begin{array}{l}\sum_{i\in S}\log\left(\sigma\left(y_i^m W_i^T h_i(X^m)\right)\right) \\ + \sum_{i\in S}\sum_{j\in Ni} y_i^m y_j^m V_{ij}^T u_{ij}(X^m) - \log(Z^m)\end{array}\right\} \tag{7}$$

However, this involves estimating the partition function *Z*. Partition function *Z* is to simulate all possible combina-

tions though vector Y which is the set of concepts. To get the exact value of Z is an NP-hard problem. Therefore, we can use various sampling methods to estimate the partition function, i.e. Markov Chain Monte Carlo (MCMC) [35] sampling, or an approximate Z value. In this work, we applied the pseudo-likelihood approach [31]. Pseudo-likelihood is a simple but solid approach to approximate Z. It gives consistent results when the vector set of Y is large. In pseudo-likelihood, we factorize the probability $p(Y|X) \approx \prod p(y_i|y_{Ni},X)$. This assumes a concept is independent to other non-related concepts. Therefore, applying a pseudo-likelihood to the partition function, we can re-write the partition function as,

$$Z = \prod_{i \in S} Z_i$$

$$Z_i = \sum_{y_i \in \{-1,1\}} \exp\left\{ \log\left(\sigma\left(y_i W_i^T h_i(X)\right)\right) + \sum_{j \in Ni} y_i y_j V_{ij}^T u_{ij}(X) \right\} \tag{8}$$

With this formula, we assume the concepts are independent in the partition function.

The inference step involves finding the optimal label configuration given an observed shot [36]. The optimal label configuration represents our estimate of the content of that shot. These are the semantic concept predictions from the model. There are two popular approaches for this inference; Maximum A Posteriori (MAP) and Maximum Posterior Marginal (MPM). MAP is widely used to estimate the prediction for binary classifiers. It tries to estimate the configuration which yields the highest probability given the video shot. Exact inference would result in obtaining the true MAP, however, this is not tractable when the number of concepts n is large. A max-flow/min-cut [37] type algorithm is frequently used to estimate MAP. However, we use Mean Average Precision (MAP) as the core metric for detection accuracy in TRECVID and this requires us to compute the marginal probability for each concept. Therefore, we apply MPM type inference in this work. MPM tries to marginalize each concept variable which is another NP-hard problem. Belief Propagation (BP) [38] provides an efficient method to estimate the MPM solution. BP is a commonly used inference method for MRF models. It was first proposed to solve non-loopy graphs but also shows stable results when applied to loopy graphs. BP simulates flows between nodes within the graph and estimates the marginal probabilities for each node when the flows are stable. To reduce the computational effort in this parameter estimation process, we proposed a generalized multi-concept discriminative random field model (GMDRF) which unifies the computation of the interaction and association potential using the same method.

4.3.4 Generalized MDRF (GMDRF)

Our MDRF model is constructed from local discriminative models combined with pair-wise interactions. However, we realized it is not necessary to have both log and logistic functions inside the exponent which makes the derivation complicated and takes more computation for the partition function. Therefore, we propose a revised version of MDRF called generalized MDRF. The generalized MDRF can be written as,

$$P(Y|X,W) = \frac{1}{Z} \exp\left(\sum_{i \in S} y_i W_{ii}^T u_{ii}(X) + \sum_{i \in S} \sum_{j \in Ni} y_i y_j W_{ij}^T u_{ij}(X) \right)$$

$$Z = \sum_{Y'} \exp\left\{ \sum_{i \in S} y_i W_{ii}^T u_{ii}(X) + \sum_{i \in S} \sum_{j \in Ni} y_i y_j W_{ij}^T u_{ij}(X) \right\} \tag{9}$$

Basically, we eliminated the log and logistic function in the association potential and made association potential similar to the interaction potential. If we take the interaction potentials to be zero, the generalized MDRF will merely be the product of logistic classifiers, which still matches the discriminative framework. In the generalized form, there is only one parameter W. The parameter estimation can be done much more easily than what we described earlier with easier derivations using a maximum likelihood estimation approach.

However, when we take a deep look at eq. 9, we discover some interesting properties. If we extend our label set as $\{ y_1, y_2, \ldots, y_n, y_1y_2, y_1y_3, \ldots, y_iy_j, \ldots\}$, the generalized MDRF can be consider as a family of Generalized Linear Model (GLM) [39]. Therefore, the value of the parameters can be obtained by maximum likelihood estimation, which requires iterative computational procedures. Moreover, we can approximate the probability by,

$$P(Y|X,W) \approx \exp\left(\sum_{i \in S} y_i W_{ii}^T u_{ii}(X) + \sum_{i \in S} \sum_{j \in Ni} y_i y_j W_{ij}^T u_{ij}(X) \right)$$
$$= \prod_{i \in S} \exp\left(y_i W_{ii}^T u_{ii}(X)\right) \prod_{i \in S} \prod_{j \in Ni} \exp\left(y_i y_j W_{ij}^T u_{ij}(X)\right) \tag{10}$$

Table 4 Semantic concept extraction with MDRF.

Runs	Average precision
SVM, multi-modal feature (baseline)	0.146
GMDRF with chi-square feature selection	0.148
GMDRF without chi-square feature selection	0.114

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

Multi-modality Comparisons	Average precision
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

If we approximate the probability without the partition function, we can factorize the probability to exponential values. This transforms the whole generalized MDRF into an independent logistic function given the labels as $\{ y_1, y_2, \ldots, y_n, y_1y_2, y_1y_3, \ldots, y_iy_j, \ldots\}$. In other words, we decompose the model into several logistic models. The first n terms denote the conditional probabilities for each semantic concept given the observation and the remaining terms work as the pair-wise logistic classifiers which depend on the data. Although this decomposition assumes the independence property for each concept given the pair-wise dependent, in our experimental results, it provides consistent result with eq. 6.

The parameter estimation of eq. 10 is a comparatively easy problem. The learning process can be easily decomposed into several logistic regression training steps. Since we factorize the model into multiple logistic models, we also decompose parameters into lower dimensional parameter sets. In each logistic regression, the sub-model tries to fit the training data to its own parameters and the overall training time is much faster than optimizing the whole parameters globally. This makes the generalized MDRF model more tractable if the concept set is large. Belief propagation is still used to estimate the marginal probabilities for each concept. The BP inference is the same as for MDRF but follows the modified definition of the association and interaction potential from eq. 9.

4.3.5 Experimental Results for GMDRF

We predict the validation set and test set using models built from our TRECVID 20005 training set and applied to the TRECVID 2006 test data. For shots in the validation set and testing set, predictions become our observations for this shot. To be more specific, we have 39 different concepts and every concept has 4 different modalities. The observation thus is a 156-dimensional vector (39x6). We adopt a logistic function as association potential.

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 shows the performance of MDRF in the TRECVID 2006 evaluation 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 (Table 5).

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
GMDRF without x^2 selection	color, texture, local feature, monolingual text	0.114
GMDRF with x^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

Monolingual text features are a bag-of-words representation of words spoken in a shot of dimensions of V_E, where V_E 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 $V_E+V_C+V_A$, where V_C and V_A 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 trained an 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 the meta-level classifier in this corpus. We thus choose logistic regression to fuse the predictions based on multi-modal features, as seen in Table 5.

5 Conclusions

Unfortunately, these experiments, as 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 has been the approach of the Pathfinder system [40] and others. [1], who explore different approaches and carefully select the combination of approaches on a concept by concept basis.

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 some ideas of techniques that may 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 [41] that even this limited detection accuracy with large numbers of concepts will be sufficient for substantial help with concept-based video retrieval.

Acknowledgements This material is based in part on work supported by the National Science Foundation under Grant No. IIS-0205219. Details about Informedia research project team can be found at http://www.informedia.cs.cmu.edu

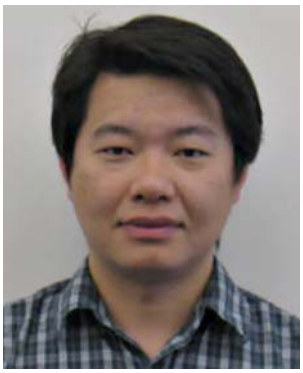
References

1. Cao, J., et al. (2006). *Intelligent Multimedia Group of Tsinghua University at TRECVID 2006. in TRECVID Video Retrieval Evaluation*. Gaithersburg: NIST.
2. Smeulders, A. W. M., et al. (2000). Content based image retrieval at the end of the early years. *IEEE Trans. Pattern Analysis and Machine Intelligence, 22*(12), 1349–1380.
3. Smith, J. R., Basu, S., Lin, C.-Y., Naphade, M. & Tseng, B. (2002). *Interactive Content-based Retrieval of Video*. in *IEEE International Conference on Image Processing (ICIP)*. Rochester, NY.
4. Rodden, K., et al. (2001). *Does organisation by similarity assist image browsing? in CHI '01: Proceedings of the SIGCHI conference on Human factors in computing systems*. New York: ACM Press.
5. Markkula, M., & Sormunen, E. (2000). End-user searching challenges: indexing practices in the digital newspaper photo archive. *Information Retrieval, 1*(4), 259–285.
6. Hauptmann, A., Yan, R., & Lin, W.-H. (2007). *How many high-level concepts will fill the semantic gap in news video retrieval?* in *International Conference on Image and Video Retrieval (CIVR)*. Amsterdam, The Netherlands.

7. Chang, S. F., Manmatha, R., & Chua, T. S. (2005). *Combining text and audio-visual features in video indexing.* in *IEEE ICASSP'05*.
8. Barnard, K., et al. (2002). Matching words and pictures. *Journal of Machine Learning Research*, **3**.
9. Naphade, M. R.,& Huang, T. S. (1998). *Semantic Video Indexing using a Probabilistic Framework*. in *I.E.E.E. International Conference on Image Processing*. Chicago, Il.
10. Lin, C.-Y., Tseng, B. L., & Naphade, M. (2003). *VideoAL: A Novel End-to-End MPEG-7 Automatic Labeling System*. in *IEEE Intl. Conf. on Image Processing*. Barcelona.
11. Lin, W., & Hauptmann, A. (2002). *News Video Classification Using SVM-based Multimodal Classifiers and Combination Strategies*. in *ACM Multimedia 2002*. Juan-les-Pins, France.
12. Jeon, J., Lavrenko, V. & Manmatha, R.. *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*.
13. Wu, Y., et al. (2004). *Optimal multimodal fusion for multimedia data analysis*. in *Proceedings of the 12th annual ACM international conference on Multimedia*.
14. Hauptmann, A., et al. (2003). *Informedia at TRECVID 2003: Analyzing and Searching Broadcast News Video*. in *Proceedings of (VIDEO) TREC 2003 (Twelfth Text Retrieval Conference)*. Gaithersburg, MD.
15. Natsev, A. P., Naphade, M. R., & Te¡si´c, J. (2005). *Learning the semantics of multimedia queries and concepts from a small number of examples*. in *Proceedings of the 13th ACM International Conference on Multimedia*.
16. Smeaton, A. F., Over, P., & Kraaij, W. (2006). *Evaluation campaigns and TRECVid. in Proceedings of the 8th ACM International Workshop on Multimedia Information Retrieval (MIR '06)*. Santa Barbara: ACM.
17. Over, P. (2007). *TRECVID: TREC Video Retrieval Evaluation*. http://www-nlpir.nist.gov/projects/t01v/.
18. Smeaton, A. F. & Over, P. (2002). *The TREC-2002 Video Track Report*.
19. Over, P., et al. (2006). *TRECVID 2006 - An Overview. in TRECVID'06 video retrieval evaluation*. Gaithersburg: NIST.
20. Kennedy, L. & A. Hauptmann, *LSCOM Lexicon Definitions and Annotations Version 1.0, DTO Challenge Workshop on Large Scale Concept Ontology for Multimedia."* Columbia University: New York.
21. Chang, S.-F., et al. *Columbia University TRECVID-2006 Video Search and High-Level Feature Extraction*. in *TRECVID Video Retrieval Evaluation*. Gaithersburg, MD: NIST.
22. Yanagawa, A., Hsu, W., & Chang, S.-F. (2006). *Brief descriptions of visual features for baseline TRECVID concept detectors*. New York: Columbia University.
23. Zhao, W., Jiang, Y. G. & Ngo, C. W. (2006). *Keyframe retrieval by keypoints: Can point-to-point matching help?* in *International Conf. on Image and Video Retrieval*.
24. Lowe, D. G. (2004). Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision, 60*(2), 91–110.
25. Sivic, J. and Zisserman, A. (2003). *Video Google: A Text Retrieval Approach to Object Matching in Videos*. in *Ninth International Conference on Computer Vision (ICCV'03)*. Nice, France.
26. Pelleg, D., & Moore, A. (2000). *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)*. San Francisco: Morgan Kaufmann.
27. Snoek, C. G. M., et al. (2004). *The MediaMill TRECVID 2004 Semantic Video Search Engine*. in *Proceedings of TRECVID 2004*. Gaithersburg, MD.
28. Snoek, C. G., et al. (2006). The semantic pathfinder: using an authoring metaphor for generic multimedia indexing. *IEEE Trans. Pattern Anal. Mach. Intell., 28*(10), 1678–1689.
29. Hauptmann, A., et al. (2004). *Confounded Expectations: Informedia at TRECVID 2004*. in *Proceedings of the TREC Video Retrieval Evaluation 2004*. Gaithersburg, MD.
30. Amir, A., et al. (2003). *IBM research TRECVID-2003 video retrieval system. in NIST TRECVID-2003.*. Gaithersburg: NIST.
31. Li, S. Z. (2004). *Markov random field modeling in image analysis*. Tokyo: Springer-Verlag.
32. Lafferty, J., McCallum, A., & Pereira, F. (2001). *Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data*. in *18th International Conference on Machine Learning (ICML)*. Morgan Kaufmann, San Francisco, CA.
33. Kumar, S. and Hebert, M. (2003). *Discriminative Random Fields: A Discriminative Framework for Contextual Interaction in Classification*. in *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*.
34. Yan, R., Chen, M.-Y. & Hauptmann, A. G. (2006). *Mining Relationship between Video Concepts using Probabilistic Graphical Models*. in *IEEE International Conference On Multimedia and Expo (ICME)*. Toronto, CA.
35. Gilks, W. R., Richardson, S., & Spiegelhalter, D. J. eds. (1996). *Markov Chain Monte Carlo in Practice*. Chapman and Hall: London.
36. Chen, M.-Y. & Hauptmann, A. (2007). *Discriminative Fields for Modeling Semantic Concepts in Video*. in *RIAO Large-Scale Semantic Access to Content*. Pittsburgh.
37. Boykov, Y., & Kolmogorov, V. (2004). An experimental comparison of Min-Cut/Max-Flow algorithms for energy minimization in vision. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 26*(9), 1124–1137.
38. Yedidia, J. S., Freeman, W. T., & Weiss, Y. (2003). Understanding belief propagation and its generalizations (chapter 8). In G. Lakemeyer & B. Nebel (Eds.), *Exploring artificial intelligence in the New Millennium*. San Francisco: Morgan-Kaufmann.
39. McCullagh, P., & Nelder, J. A. (1987). *Generalised Linear Models*. London: Chapman and Hall.
40. Snoek, C., Worring, M., & Hauptmann, A. G. (2006). Learning rich semantics from news video archives by style analysis. *TOMCCAP, 2*(2), 91–108.
41. Hauptmann, A., et al. (2007). *Can high-level concepts fill the semantic gap in video retrieval? A case study with broadcast news.* IEEE Transactions on Multimedia Journal.

Alexander G. Hauptmann is a Senior Systems Scientist in the CMU Computer Science Department and also a faculty member with CMU's Language Technologies Institute. He received his B.A. and M.A. in

Psychology from Johns Hopkins University, studied Computer Science at the Technische Universität Berlin from 1982–1984, and received his Ph.D. in Computer Science from CMU in 1991. His research interests have led him to pursue and combine several different areas: man-machine communication, natural language processing, speech understanding and synthesis, machine learning. He worked on speech and machine translation at CMU from 1984–94, when he joined the Informedia project for digital video analysis and retrieval and led the development and evaluation of the News-on-Demand applications.



Ming-yu Chen received his computer science and information engineering degree from the National Taiwan University, Taiwan, in 1999. He is now pursuing the Ph.D. degree at the Carnegie Mellon University. His research interest lies in the area of video retrieval, video understanding and human action analysis in large video achieves.

Michael Christel has worked at Carnegie Mellon University, Pittsburgh, PA, since 1987, first with the Software Engineering Institute and since 1997 as a Senior Systems Scientist in the School of Computer Science. In 2008 he joined the Entertainment Technology Center faculty as a research professor. He is a founding member of the Informedia research team at Carnegie Mellon University designing, deploying, and evaluating video analysis and retrieval systems for use in education, health care, humanities research, and situation analysis. His research interests focus on the convergence of multimedia processing, information visualization, and digital library research. He has worked with digital video since its inception in 1987 and received his Ph.D. from the Georgia Institute of Technology, Atlanta, GA, in 1991, with his thesis examining digital video interfaces for software engineering training. He received the B.S. degree in mathematics and computer science from Canisius College, Buffalo, NY, in 1983.

Wei-Hao Lin received the B.E. degree from National Taiwan University, Taipei, Taiwan, in 1999, and received the Ph.D. degree from Language Technologies Institute, School of Computer Science, Carnegie Mellon University, Pittsburgh, PA in 2009. His research interest includes multimedia content understanding, image/video retrieval, and natural language processing. His Ph.D. thesis focuses on identifying ideological perspectives (e.g., political views) expressed in text and video. He has contributed to CMU's highlevel feature extraction submissions to TRECVID, many of which performed best in 2004, 2005, and 2006.

Dr. Jun Yang is currently a Software Engineer at Google Inc. Dr. Yang received his Ph.D. (2008) degree from School of Computer Science, Carnegie Mellon University, and M.S. (2003) and B.S. (2000) degree form Department of Computer Science, Zhejiang University. His research interests include multimedia information retrieval, video analysis, and machine learning. He has authored over 50 refereed conference and journal papers, and has won Best Paper Award at Conf. on Image and Video Retrieval (2006) and SIAM Conf. on Data Mining (2007).