5-2007

Discriminative Fields for Modeling Semantic Concepts in Video

Ming-yu Chen
Carnegie Mellon University

Alexander Hauptmann
Carnegie Mellon University

Discriminative Fields for Modeling Semantic Concepts in Video

Ming-yu Chen & Alexander Hauptmann
Carnegie Mellon University
500 Forbes Ave
Pittsburgh, PA 15213, USA
{mychen, alex}@cs.cmu.edu

Abstract

According to some current thinking, a very large number of semantic concepts could provide researcher a novel way to characterize video and be utilized for video retrieval and understanding. These semantic concepts do not isolate to each other and thus exploiting relationships between multiple semantic concepts in video could be a very useful source to enhance the concept detection performance. In this paper we present a discriminative learning framework called Multi-concept Discriminative Random Field (MDRF) for building probabilistic models on video semantic concept detections by incorporating related concepts as well as the observation. The proposed 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 can not only capture the co-occurrence between concepts but also incorporate the data observation in a unified framework. We also present an approximate parameter estimation algorithm and apply it to TRECVID 2005 data. Our experiments show promising results compared to the single concept learning approach for video semantic detection.

1. Introduction

The detection of a large number of semantic concepts (on the order of thousands) has been seen as an intermediate step in enhancing semantic video search and retrieval. Semantic concepts play a role as basic semantic units for users to manipulate the video content. To avoid manually annotating every possible semantic concept, researchers have developed a variety of automatic concept detection techniques on the basis of statistical learning. The most popular approach is to decouple the set of semantic concepts and translate the learning task into multiple binary classification problems with a presence/absence label for each individual concept. Then, for each video shot, its associated video concepts can be detected using multiple unimodal or multimodal classifiers based on visual, audio and text/speech-transcription features.

However, these binary classification approaches assume independence between concepts and ignore the important fact that semantic concepts do not exist in isolation 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" always 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 not rare and it can be expected that mining multi-concept relationships can serve as a useful source of information to improve the concept detection accuracy. Moreover, such a correlated context could also be used to automatically construct a semantic network or ontology tailored to the video collection in a bottom-up manner. This automatic ontology construction is helpful to discover unknown concept relationships that could be complementary to manually specified ontologies.

To automatically exploit benefits from multi-concept relationships, several approaches have been proposed before which build upon advanced pattern recognition techniques within a probabilistic framework. For example, Naphade (Naphade 1998) 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 (Snoek 2004) et al. proposed a semantic value chain architecture including a multi-concept learning layer called the context link. At the top level, it 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 (Hauptmann 2004) et al. fused the multi-concept predictions and captured the inter-concept causation by constructing an additional logistic regression classifier on top the single concept detection results. Amir (Amir 2003) et al. concatenated the concept prediction scores into a long vector called model vectors 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 (Wu 2004) et al. which attempted to model possible influence relations between concepts based on a predefined ontology hierarchy.

One direction that has not been explored, despite the many efforts 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) (Li 2004) is a commonly used model in computer vision to utilize contextual information. 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 purpose, we are more interested in estimating the posterior over labels given the observation than joint probability. Conditional Random Field (CRF) (Lafferty 2001) is 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 depends on the observation sequence, any arbitrary/non-independent features can be derived from the observation sequence, without forcing the model to account the distribution of these dependencies. Therefore, CRF provides a new type of random field models to incorporate the dependency among observations rather than single matches. Discriminative random field (DRF) (Kumar 2003) provides 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.

In this work we present a discriminative learning framework based on the concept of DRF. A multi-concept Discriminative Random Field (MDRF) model is proposed in this paper to achieve our goal in detecting multiple semantic concepts of the video. Our model expands DRF by learning multiple classifications instead of single classifier. Our MDRF brings many new aspects comparing to previous work. First, MDRF is an undirected graph model which does not require prior causation analysis to understand the dependencies between concepts (Naphade 1998)(Seoek 2004)(Amir 2003). Second, MDRF learns both classification and interaction simultaneously. No additional data is required for linkage discovery, unlike (Naphade 1998)(Seoek 2004)(Hauptmann 2004). Third, the interactions are also linked with observation which allows the model to not only capture the co-occurrence between concepts but also learn pair-wise relations in feature space (Yan 2006). In Section 2, we will present the MDRF model with its parameter learning and inference. In Section 3, a generalized MDRF is proposed for more efficient learning and inference purpose. The experimental results based on TRECVID 2005 data will be presented in Section 4. We end with a discussion and future work in Section 5.

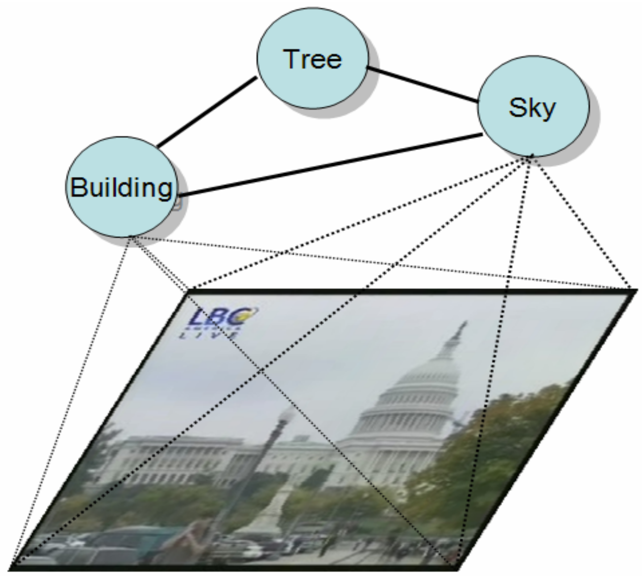


Figure 1: A graph demonstrates the framework of MDRF. There are three semantic concepts in this video shot: building, tree and sky. The top layer shows the concepts relations with each other and constitutes an undirected graph. The edges between each concept can be viewed as interaction potentials in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

2. Multi-concept Discriminative Random Fields

In this section, we present Multi-concept Discriminative Random Fields based on the concept of Discriminative Random Fields. First, we describe our notations which will be used in the whole paper. Y is the vector of multiple concept labels: $Y = (y_1, y_2, \ldots, y_n)$ where y_i denotes the label of *i*th concepts. In this work, each semantic concept detector is considered as a binary classification, i.e. $y_1 = \{-1, 1\}$. X is the observation extracted from the video shot, i.e. $X = R^c$. $\theta = \{W, V\}$ is the parameters in the mode. W is the parameter of the association potential and V is the parameter of the interaction potential.

2.1. Model description

Figure 1 illustrates the framework of MDRF in a video shot. Figure 2 describes the graphical model representation for the proposed 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 jointly optimized in such a unified framework. Thus, based on this graphical model, the conditional probability of the labels Y given the observations X can be written as,

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

where Z is a normalizing constant know as the partition function, A_i is a unary potential function between observations and labels and I_{ij} is a pair-wise potential function related to observations and pair-wise concepts. S in Eq. 1 denotes the concept set and N_i denotes the related concepts to *i*th concept. In this paper, we will describe A_i as an association potential function and I_{ij} as interaction potential function. In general, Z, A_i, and I_{ij} constitute the MDRF model. Z normalizes the exponential value to be a probability and it demonstrates all the possible combinations of association potential and interaction potential. The association potential function basically works like a single concept classifier. In this paper, we use discriminative models in the association potential function to achieve our classification goal. The interaction potential function plays an important role in the MDRF model. This potential expands the model to present pair-wise relationships that cooperate with the observation. It acts like a pair-wise classification to address the dependencies between concepts. In the following discussions, we will discuss the choice of association potentials and interaction potentials in more detail.

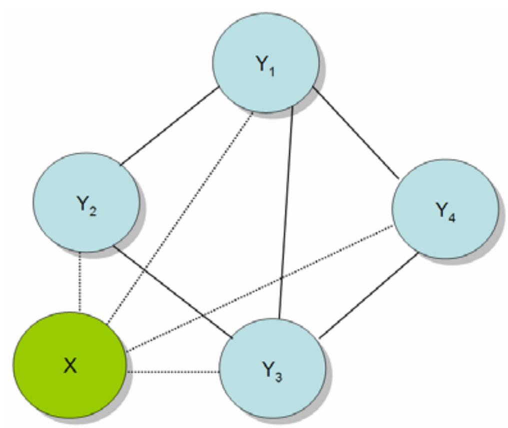


Figure 2: MDRF is a fully connected undirected graphical model. Y nodes denote the semantic concepts. X is the observation extracted from the video. All concepts are dependent on the observation.

2.1.1 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,

$$A_i\left(y_i,W,X\right)=\log\left(p\left(y_i\mid X\right)\right) \quad (2)$$

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,

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

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 performs 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,

$$\begin{aligned} A_i(y_i, W, X) &= \log\left(\mathrm{p}(\mathrm{y}_\mathrm{i} \mid X)\right) \\ &= \log\left(\sigma\left(\mathrm{y}_\mathrm{i} \mathrm{w}_\mathrm{i}^\mathrm{T} h_i(X)\right)\right) \end{aligned} \qquad (4)$$

If the interaction potential function is set up as zero, the whole MDRF is the same as building n individual logistic regressions for each concept. This shows 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 dimensionality reduction. However, in principle, any mapping or transformation function can be used here, for different purposes. In computer vision, researchers try to use a kernel function to utilize contextual information. In the multimedia domain, multi-modality fusion can be performed here to improve the power of the model for complex aspects of the multimedia domain.

2.1.2 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 will emphasize edge features. If a detector detects some blue in the shot, it results in a high possibility for this shot to contain sky. If there are some vertical edges, the shot has high possibility to have buildings within the shot. However, the interaction potential here can learn a model which is contains both color and edge features to predict the co-occurrence of sky and building.

To design the interaction term, we borrow the commonly used form from MRF model, $I = \alpha y_i y_j$, 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,

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

where V is the parameter and $u_{ij}(X)$ is the function to convert an observation into a feature vector. Just as 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 the strongly related concept pairs. Therefore, MDRF is designed to be a fully connected undirected graphical model. We hope this can even capture some useful linkage between concepts that is not obvious to human knowledge but useful in classification. However, MDRF can always be applied after some co-occurrence analysis to cut off weak links between concepts. The full 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 in section 2.1, once we set those pair-wise potential to zeros, the model will act just like a traditional semantic concept detector. In that case, the model won't capture the interactions between concepts and make the assumption that each concept is isolated from each other. Then the 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 MRF 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 and more expendable for the further research.

2.2. Parameter learning

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 the state of the art of our approach. It achieves the goal of training multiple classifications and modeling the relationships between them simultaneously and does not require additional data or further split of the training data to obtain an additional held-out set for the combination step. With the association and interaction potential functions defined by section 2, the MDRF can be written as,

$$P(Y \mid X, \theta) = \frac{1}{Z} \exp\left(\sum_{i \in S} \log\left(\sigma\left(y_i W_i^T h_i(X)\right)\right) + \sum_{i \in S} \sum_{j \in Ni} y_i y_j V_{ij}^T u_{ij}(x) \right) \qquad (6)$$

$$\text{where } \; Z = \sum_{Y'} \exp\left\{ \sum_{i \in S} \log\left(\sigma\left(y_i W_i^T h_i(X)\right)\right) + \sum_{i \in S} \sum_{j \in Ni} y_i y_j V_{ij}^T u_{ij}(x) \right\}$$

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

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

$$P(Y \mid X) = \prod_{m=1}^{M} P(Y^m \mid X^m)$$

$$l(\theta) = \log(P(Y \mid X)) = \sum_{m=1}^{M} \log(P(Y^m \mid X^m))$$

$$= \sum_{m=1}^{M} \left\{ \sum_{i \in S} \log\left(\sigma\left(y_i^m W_i^T h_i(X^m)\right)\right) + \sum_{i \in S} \sum_{j \in Ni} y_i^m y_j^m V_{ij}^T u_{ij}(X^m) - \log(Z^m) \right\} \quad (7)$$

However, this involves the problem of estimating the partition function Z. Partition function Z is to simulate all possible combinations 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 either sampling methods to estimate the partition function, i.e. Markov Chain Monte Carlo (MCMC) (Gilk 1996) sampling, or an approximate Z value. In this work, we tried the pseudo-likelihood approach (Li 2004). 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 means that we assume a certain concept is independent to other non-related concepts. Therefore, applying pseudo-likelihood to the partition function, we can re-write the partition function as,

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

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

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

In the MDRF model, the association potential function estimates the probability given the observation of a certain concept and the interaction potential function captures the pair-wise cooperation between concepts given the observed data. Since the association potential links the concept directly, we want to emphasize the association potential more than the interaction potential to stabilize the model. Therefore, we add a penalization function into the log likelihood in the parameter learning process. The penalized pseudo log likelihood then can be written as,

$$l(\theta) = \log(P(Y \mid X)) - \frac{1}{2\tau^2} V^T V = \sum_{m=1}^{M} \log(P(Y^m \mid X^m)) - \frac{1}{2\tau^2} V^T V$$

$$= \sum_{m=1}^{M} \left\{ \sum_{i \in S} \log\left(\sigma\left(y_i^m W_i^T h_i(X^m)\right)\right) + \sum_{i \in S} \sum_{j \in Ni} y_i^m y_j^m V^T u_{ij}(X^m) - \log(Z^m) \right\} - \frac{1}{2\tau^2} V^T V$$

$$\approx \sum_{m=1}^{M} \sum_{i \in S} \left\{ \log\left(\sigma\left(y_i^m W_i^T h_i(X^m)\right)\right) + \sum_{j \in Ni} y_i^m y_j^m V^T u_{ij}(X^m) - \log(Z_i^m) \right\} - \frac{1}{2\tau^2} V^T V \quad (9)$$

$$where \quad Z_i^m = \sum_{y_i \in \{-1,1\}} \exp\left\{ \log\left(\sigma\left(y_i W_i^T h_i(X^m)\right)\right) - \sum_{j \in Ni} y_i y_j^m V^T u_{ij}(X^m) \right\}$$

where τ is the penalization parameter to V. If τ is given, then (9) is a convex function. The likelihood function can be easily maximized using gradient descent approach and parameter estimation can be achieved by maximizing likelihood. However, τ should be the model parameter and should be estimated in learning process. In equation (9), τ is integrated with parameter V which makes it difficult to learn them together. In our work, we use cross-validation to select τ and set τ as given constant during parameter learning to simplify the task.

For learning the parameters using the gradient descent approach, the derivations are needed to update the likelihood. Before we specify the derivations, we need to re-write the partition function in a simplified format for further presentation in the derivation,

$$Z_i^m = \sum_{y_i \in \{-1,1\}} \exp\left\{ \log\left(\sigma\left(y_i W_i^T h_i\left(X^m\right)\right)\right) - \sum_{j \in Ni} y_i y_j^m V_{ij}^T u_{ij}\left(X^m\right) \right\}$$
$$= \exp\left(zp\left(Y^m, X^m, \theta\right)\right) + \exp\left(zn\left(Y^m, X^m, \theta\right)\right) \quad (10)$$

where zp(Y,X,θ) is the function for positive summation in the partition function and zn(Y,X,θ) is the negative summation term.

The derivation of parameter W is,

$$\frac{dl(\theta)}{dW_i} = \sum_{m=1}^{M} \sum_{i \in S} \left\{ \frac{d \log\left(\sigma\left(y_i^m W_i^T h_i\left(X^m\right)\right)\right)}{dW_i} + 0 + \frac{d \log\left(Z_i^m\right)}{dW_i} \right\}$$
$$= \sum_{m=1}^{M} \sum_{i \in S} \left\{ \left(1 - \sigma\left(y_i^m W_i^T h_i\left(X^m\right)\right)\right)\left(y_i^m h\left(X^m\right)\right) + \frac{d \log\left(Z_i^m\right)}{dW_i} \right\} \quad (11)$$
$$where \quad \frac{d \log\left(Z_i^m\right)}{dW_i} = \frac{d \log\left(\exp\left(zp\left(Y^m, X^m, \theta\right)\right) + \exp\left(zn\left(Y^m, X^m, \theta\right)\right)\right)}{dW_i}$$
$$= \frac{1}{\exp(zp) + \exp(zn)} \left\{ \exp(zp)\left(1 - \sigma\left(W_i^T h_i\left(X^m\right)\right)\right) h\left(X^m\right) - \exp(zn)\left(1 - \sigma\left(-W_i^T h_i\left(X^m\right)\right)\right) h_i\left(X^m\right) \right\}$$

and the derivation of parameter V can be written as,

$$\frac{dl(\theta)}{dV_{ij}} = \sum_{i=1}^{M} \sum_{i \in S} \left\{ 0 + \frac{d\left(\sum_{j \in Ni} y_i^m y_j^m V_{ij}^T u_{ij}\left(X^m\right)\right)}{dV_{ij}} - \frac{d \log\left(Z_i^m\right)}{dV_{ij}} \right\} - \frac{1}{\tau^2} V_{ij}$$
$$= \sum_{i=1}^{M} \sum_{i \in S} \left\{ \sum_{j \in Ni} y_i^m y_j^m u_{ij}\left(X^m\right) - \frac{d \log\left(Z_i^m\right)}{dV} \right\} - \frac{1}{\tau^2} V_{ij} \quad (12)$$
$$where \quad \frac{d \log\left(Z_i^m\right)}{dV_{ij}} = \frac{d \log\left(\exp\left(zp\left(Y^m, X^m, \theta\right)\right) + \exp\left(zn\left(Y^m, X^m, \theta\right)\right)\right)}{dV_{ij}}$$
$$= \frac{1}{\exp(zp) + \exp(zn)} \left\{ \exp(zp) y_j^m u_{ij}\left(X^m\right) - \exp(zn) y_j^m u_{ij}\left(X^m\right) \right\}$$

zp and zn in eq. 11 and 12 denote as zp(Y,X,θ) and zn(Y,X,θ) functions to simplify the representation of the derivations.

2.3. Parameter learning

The inference is to find the optimal label configuration given an observed shot. 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 tires to figure out the configuration which gives the highest probability given the video shot. Exact inference is the way to get the true MAP, however, it's not tractable when n (the number of concepts) is large. A max-flow/min-cut (Boykov 2004) type algorithm is widely used to estimate MAP. However, we use Mean Average Precision (MAP) as the core measurement 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) (Yedidia 2003) provides an efficient method to estimate the MPM solution. BP is a commonly used inference method for MRF models. It was proposed to solve non-loopy graphs first but also shows stable result 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. The update rules of BP for our MDRF are,

$$b_i^{(t)}(y_i) = kA_i(y_i, W, X)\prod_{j\in N_i} m_{ij}^{(t)}(y_i) \tag{13}$$

$$m_{ij}^{(t)}(y_i) = \sum_{y_i} A_i(y_i, W, X)I_{ij}(y_i, y_j, V, X)\prod_{k\in N_i\setminus j} m_{ki}^{(t-1)}(y_i) \tag{14}$$

$b_i()$ function denotes the belief of each node, $m_{ij}()$ function denotes flow from node to node and t is the iteration index. After the propagation converges, the belief of each node is, in fact, the marginal probability after the normalization. Although it can't be proved that BP in loopy graph are guaranteed to converge, empirically, BP works very well even in a fully connected graph.

3. Generalized MDRF

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,

$$P(Y \mid X, W) = \frac{1}{Z}\exp\left(\sum_{i\in S} y_i W_{ii}^T u_{ii}(X) + \sum_{i\in S}\sum_{j\in Ni} y_i y_j W_{ij}^T u_{ij}(X)\right) \tag{15}$$

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

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 in section 2 with easier derivations using a maximum likelihood estimation approach.

However, when we take a deep look at equation 15, 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) (McCullagh 1987). 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,

$$P(Y \mid X, W) \approx \exp\left(\sum_{i\in S} y_i W_{ii}^T u_{ii}(X) + \sum_{i\in S}\sum_{j\in Ni} y_i y_j W_{ij}^T u_{ij}(X)\right) \tag{16}$$

$$= \prod_{i\in S} \exp\left(y_i W_{ii}^T u_{ii}(X)\right)\prod_{i\in S}\prod_{j\in Ni} \exp\left(y_i y_j W_{ij}^T u_{ij}(X)\right)$$

If we approximate the probability without the partition function, we can factorize the probability to exponential values. This makes 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 equation 6.

The parameter estimation of eq. 16 becomes a comparatively easy problem. The learning process can be easily decomposed as 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 still the same as eq. 13 and eq. 14 but change the association and interaction potential in eq. 15.

4. Experimental results

The TREC Video Retrieval Evaluation (TRECVID) (Semeaton 2003) provides an open and rich video collection. Currently, TREVID focuses on the video news domain because it is structured video and contains a broad range of information. The TRECVID 2005 collection contains three different languages, Arabic, Chinese and English. The video collection comes from six different databases, LBC for Arabic news, CCTV4 and NTDTV for Chinese news, CNN, NBC and MSNBC for English news. The development set contains 137 episodes and the total size is about 80 hours. The video collection is decomposed as 74523 shots, which are used as the basic units of the video content. We split this data into two parts. 70% of the collection (55293 shots) is used for training the model and 30% of the data (19230 shots) is used to evaluate the models.

Although the video contains rich information from visual, audio, and text extracted from screen and speech recognition, in our experimental setting, we use only the color and texture features extracted from key-frame images in each shot. The key-frame is the most representative frame in each shot. TRECVID defines the key-frame for each shot. We extract color features from 5x5 fixed grids and Gabor texture feature from the whole frame. The total dimension of our features is 225.

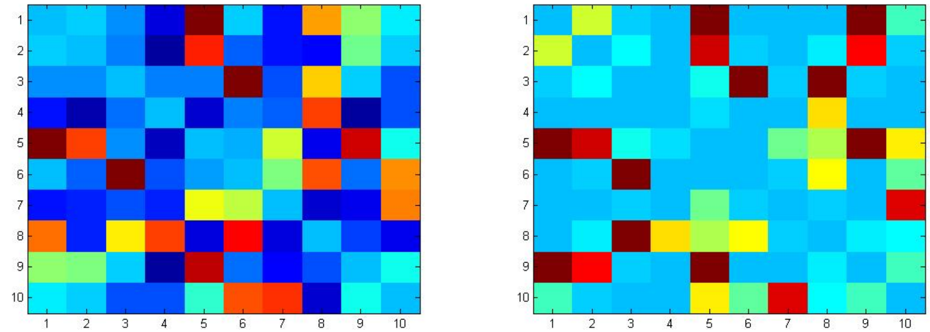
Our semantic concept labels come from Large Scale Concept Ontology for Multimedia Understanding (LSCOM) (Kennedy 2006) workshop. The workshop aims to create a user-driven concept ontology for analysis of video broadcast news. Currently, LSCOM has manually labeled around 500 concepts based on the information needs determined by users, library scientists and knowledge experts.

For comparison, we applied Conditional Random Field (CRF) model which does not model the observation dependency in interaction potential there. The CRF can be written as,

$$P(Y \mid X, W) = \frac{1}{Z} \exp\left(\sum_{i \in S} y_i W_i^T h_i(X) + \sum_{i \in S} \sum_{j \in Ni} y_i y_j V_{ij} \right) \qquad (17)$$

$$Z = \sum_{Y'} \exp\left\{ \sum_{i \in S} y_i W_i^T h_i(X) + \sum_{i \in S} \sum_{j \in Ni} y_i y_j V_{ij} \right\}$$

the $h_i(X)$ function here, we applied PCA with 85% energy which reduces the dimension from 255 to 50. The parameter V captures interactions between concepts independent to observed data. Basically, this CRF model only removes the data dependency term from GMDRF in the interaction potential.



(a) (b)

Figure 3: (a) plots the χ2 analysis result. Each grid denotes the pair-wise χ2 value. The darker shows the higher positive interaction and the lighter shows the lower correlation. (b) plots the parameter learned from MDRF. The brown shows the most positive interaction and deep blue presents the most negative interaction.

4.1. Human knowledge concept set

We chose 10 concepts from the LSCOM concept set. We chose those concepts which we expected to show some relationship to each other. The concepts we choose are **{building, car, face, maps, outdoor, person, sports, studio, urban, walking/running}**. We verified that there were some positive pairs and some negative pairs in this concept set. To illustrate the relationship correlation, we plot the pair-wise Chi-Square (χ2) analysis in Figure 3(a). In Figure 3(a), the darker (brown) shows the higher χ2 score and the lighter shows the lower χ2 score. In this graphic, we can figure some positive pairs like {building, outdoor}, {building, urban}, {face, person}, {walking/running, sports} and some negative pairs like {building, maps}, {maps, person}, {urban, sports}. This graph shows there is a large number of concepts which are strongly related or co-occur together. In Figure 3(b), we plot the parameter V from MDRF. In figure 3(b), red/brown shows the positive interaction and deep blue denotes negative interaction. Figure 3(b) demonstrates that MDRF captures co-occurrence of concepts. Compare to χ2 analysis, MDRF demonstrates more variety in estimating the interaction due to the comprehensive model. This also shows that the MDRF model can actually capture the interactions between concepts.

The proposed MDRF model was applied to the task of detecting 10 concepts in TRCVID 2005 video collection. The aim is to label each shot with its corresponding concepts. For each shot, the key-frame represents the content of the shot. 225-dim color and texture features are extracted from the key-frame as observed data. For the association potential, the transformed feature vector $h_i(X)$ is a 50-dim vector with its dimensionality reduced by PCA at 85% energy. For the interaction potential, $u_{ij}(X)$ is applied the same as $h_i(X)$. The MDRF model is a full connected graph. We make an assumption that every concept has a certain interaction with any other concept in the set. Figure 3 shows we can learn the interaction fairly well even if the model is fully connected. For MDRF, τ is chosen as 0.01 by cross-validation. Logistic regression (LG) is first applied to the concept set as the baseline. We construct 10 individual logistic regressions as classifiers for each concept. This baseline is comparable since the local discriminative functions in the proposed MDRF are logistic functions. CRF removes the data dependency in the interaction potential. GMDRF is the generalized MDRF described in section 4. In MDRF/GMDRF, parameter W is a 510-dim vector, 10 concept * 51-dim (additional dimension comes from constant) from feature vector, and parameter V is a 2295-dim vector, 45 concept pairs * 51-dim. We use mean average precision as the performance measurement because it is the standard measurement in TRECVID. Table 1 shows the mean average precision result on the evaluation set. We can observe both MDRF and GMDRF obtain around 5% improvement over the baseline logistic regression. Furthermore, only 1 concept (maps) in the MDRF result has lower performance than the baseline. This provides the evidence that the interaction potential can actually exploit additional information when constructing semantic concept detections. Comparing CRF with MDRFs shows that using the data dependency provides a better result; however, the difference is quite minimal. Generally, we can discover some interesting properties in the performance. For most strong detectors, {face, outdoor, person, studio}, which have a MAP higher than 0.5, the improvement isn't so obvious. However, for weak detectors, {building, car, maps, sports, urban, walking/running}, which have MAP lower than 0.5, MDRF/GMDRF improves the results up to 15%. Table 2 shows the performance analysis between strong detectors and weak detectors. We also found CRF only improves limited amount to baseline.

Concept	**Mean Average Precision (MAP)**			
	LG	**CRF**	**MDRF**	**GMDRF**
Building	0.2557	0.2671	0.2899	0.2987
Car	0.2165	0.2008	0.2553	0.2536
Face	0.6846	0.6879	0.6840	0.6895
Maps	0.3732	0.3930	0.3712	0.3699
Outdoor	0.6485	0.6531	0.6632	0.6537
Person	0.7896	0.7897	0.7871	0.7957
Sports	0.1609	0.2031	0.2213	0.2011
Studio	0.5997	0.6043	0.6010	0.5953
Urban	0.1122	0.1213	0.1649	0.1536
Walking/running	0.1796	0.1901	0.1931	0.1905
Avg	0.4020	0.4110	0.4231	0.4202

Table 1: Performance comparison of different MDRF models in 10 concept set; MDRF$^-$ denotes the MDRF model without data dependency in interaction potential. MDRF is the model we proposed. GMDRF is the generalized MDRF model.

Concept	**Mean Average Precision (MAP)**			
	LG	**CRF**	**MDRF**	**GMDRF**
Strong detectors	0.6806	0.6838 (0.4%)	0.6838 (0.4%)	0.6836 (0.4%)
Weak detectors	0.2163	0.2292 (5.9%)	0.2493 (15.3%)	0.2446 (13.1%)

Table 2: Performance analysis between strong detectors and weak detectors

4.2. Random concept set

Human knowledge provides a very solid understanding about concept relationships. However, one goal of multiple semantic concept learning is to discover the ontology between concepts or, in the other word, the relationship behind human knowledge. Therefore, we constructed another concept set without human knowledge. We used concepts with 3000-5000 positive examples in our LSCOM concept set. The concepts we choose were **{anchor, government leader, interview on location, meeting, road, text, urban, Asian people, car, interview sequences, politicians, real trees, simple female person}**.

The performance comparison is shown in table 3 and the strong/weak detector results are represented in table 2. Strong detectors here are {anchor, text} and the remaining classes are all weak detectors which have MAP lower than 0.5. From the performance results, CRF performs worse compared to the baseline but MDRF/GMDRF still shows around a 9% improvement. The strong/weak detector performance result shows the multiple-concept learning still improves week classifiers more than strong classifiers. This concept sets illustrates the data dependency stability problems of the MDRF model. The relationships between concepts are not as strong as in the first data set and most classifiers here are weak classifiers. Therefore, CRF only models the interaction in the co-occurrence of labels and discard the observed data. This makes the model emphasize the interaction relations which are not strong evidence to detect the concept. We can discover this in table 4 where the performance of weak detectors degrades a lot in CRF. MDRF/GMDRF has the data dependent interaction potential. In weak classifiers, it is hard to find the right decision boundary to classify the concept. It turns out the interaction potential has same difficulty in figuring out the correct decision boundary. This makes the MDRF/GMDRF not overestimate the interaction potential and stabilizes the performance.

Concept	Mean Average Precision (MAP)			
	LG	CRF	MDRF	GMDRF
Anchor	0.6840	0.6690	0.6953	0.6967
Government Leader	0.1783	0.1305	0.1542	0.1803
Interview on Location	0.1446	0.1317	0.1981	0.1810
Meeting	0.1204	0.1039	0.1429	0.1452
Road	0.1429	0.1141	0.1622	0.1615
Text	0.6119	0.6041	0.6101	0.6213
Urban	0.1122	0.0913	0.1312	0.1234
Asian People	0.2104	0.2020	0.2983	0.2930
Car	0.2165	0.1918	0.2173	0.2365
Interview Sequences	0.2594	0.2513	0.3122	0.2787
Microphones	0.0783	0.0646	0.1020	0.0957
Politicians	0.1946	0.1652	0.1731	0.1952
Real trees	0.2355	0.2242	0.2542	0.2642
Single female person	0.2186	0.1607	0.2570	0.2479
Avg	0.2434	0.2217	0.2649	0.2658

Table 3: Performance comparison of different MDRF models in 14 concepts set.

Concept	Mean Average Precision (MAP)			
	LG	**CRF**	**MDRF**	**GMDRF**
Strong detectors	0.6480	0.6366 (-1.7%)	0.6953 (7.3%)	0.6967 (7.5%)
Weak detectors	0.1760	0.1526 (-13.3%)	0.2002 (13.8%)	0.2002 (13.8%)

Table 4: Performance analysis between strong detectors and weak detectors

4.3. Discussion

In the experimental results, we can discover if there are strong/weak pairs, it normally helps the weak classification, e.g. {Anchor, Interview Sequences}, {Anchor, Single female person}. Unfortunately, there are too few strong detectors in our concept set. This also limits the power of the current MDRF model in this data set. The interaction potential captures the agreement and disagreement of concept pairs. This is similar to the idea of $\chi 2$ analysis to model the co-occurrence. However, most of the semantic concept classifications in video database are rare classifier problems. This means the positive data only has very small numbers in the data set. Even we choose 3000-5000 positive example concepts in our randomly selected concept set, there is only around 5-7% positive data in the whole data set. Therefore, the negative-negative pairs dominate the agreement part which we consider less significant than positive-positive pairs in the interaction potential. On the other hand, we want to emphasize the positive examples more.

5. Conclusion and Future Work

In this paper, we proposed Multi-concept Discriminative Random Fields (MDRF) to explode the multiple concepts learning idea in multimedia domain. MDRF combines the local discriminative models with adaptive data-dependent interactions to learn and explode semantic content of the video. The experimental results show MDRF improves the classification results especially for weak concept detectors significantly.

5.1. Conclusion

This is the very initial work for us to bring undirected models into video analysis domain. The experimental results show it's a promising direction to work with. There are some benefits to use undirected model to achieve multiple concepts learning. First, it achieves the learning and relationship modeling processes by one step. It does not require additional held out set for further linkage analysis. Second, the model doest not require prior human knowledge to construct semantic structure. From our experimental results, even you choose the concept randomly without any human knowledge or statistic analysis, MDRF model still achieve the improvement to the classification tasks. This provides researchers a novel approach to discover video ontology automatically. Third, MDRF gives the evidence that observation dependency provides useful information when modeling the interaction. The experimental result shows without data dependency CRF works worse in our corpus. Finally, MDRF itself still has a lot of flexibilities for different area. Researchers can design the association potential and interaction potential for specific domains. For an example, in text categorization, it also exists the topic that multiple categories for a certain document. MDRF is a suitable model for this research topic. Our experimental results support the ideal of MDRF to be able to achieve multiple concept learning.

5.2. Future work

There are many different aspects for future work we are exploring now. The most important part is to design a new interaction potential which has the ability to solve the problems we mentioned in the experimental result discussion section. One possible interaction potential is a function which is able to model positive-negative, positive-positive and negative-negative interactions rather than agreement/disagreement.

There are many further research topics around the association potential. Currently, we use logistic regression as our association potential. However, Support Vector Machine (SVM) is current state of the art in classification. Mapping SVM kernel function as association potential will be a next step for association potential.

Another interesting direction for future works is in transformation functions on the association potential and interaction potential ($h_i(X)$ and $u_{ij}(X)$). In fact, there are two further interesting topics in video classification; one is the multi-modality issue and the other is temporal information. Multi-modality is a key feature of the multimedia domain because multimedia data contains several features from multiple sources. How to fuse and utilize them is still an ongoing research topic. Temporal information plays a key feature in video source. The connected shots provide some information by their connections. We think the transformation function gives us the flexibility to adopt those into the model.

Other future work is on graphical model learning problems. There are many different parameter estimation and inference methods which are suitable in different situation and domain. There are many alternative ways for parameter learning and inference, like Contrastive Divergence (CD) and BP in parameter learning.

Acknowledgements

This material is based on work supported by the Advanced Research and Development Activity (ARDA) under contract number H98230-04-C-0406 and NBCHC040037.

References

Naphade, M. R., Kristjansson, T., Frey, B., and Huang, T.S. (1998). Probabilistic multimedia objects (multijects): A novel approach to video indexing and retrieval in multimedia systems. In Proc. of IEEE Internatioanl Conference on Image Processing (ICIP)

Snoek, C.G.M., Worring, M., Geusebroek, J.M., Koelma, D.C., and Seinstra, F.J. (2004). The mediamill TRECVID 2004 semantic video search engine. In Proc. of TRECVID

Hauptmann, A., Chen, M.-Y., Christel, M., Huang, C., Lin, W.-H., Ng, T., Papernick, N., Velivelli, A., Yang, J., Yan, R., Yang, H., and Wactlar, H.D. (2004). Confounded expectations: Informedia at trecvid 2004. In Proc. of TRECVID

Amir, A., Hsu, W., Iyengar, G., Lin, C.-Y., Naphade, M., Natsev, A., Neti, C., Nock, H.J., Smith, J.R., Tseng, B.L., Wu, Y., and Zhang, D. (2003). IBM research TRECVID-2003 video retrieval system. In Proc. of TRECVID

Wu, U., Tseng, B.L., and Smith, J.R. (2004). Ontology-based multi-classification learning for video concept dtection. In Proc. of IEEE International Conference on Multimedia and Expo (ICME)

Li., S. Z. (2004). Markov Random Field Modeling in Image Analysis. Springer-Verlag, Tokyo

Lafferty, J. McCallum, A., and Pereira, F. (2001). Conditioanl random fields: Probabilistic models for segmenting and labeling sequence data. In Proc. of 18th International Conference on Machine Learning

Kumar, S., and Hebert, M. (2003). Discriminative Random Fields: A Discriminative Framework for Contextual Interaction in Classification. In Proc. of IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)

Yan, R., Chen, M.-Y., and Hauptman, A. (2006). A. Mining relationship between video concepts using probabilistic graphical models, In Proc. of IEEE International Conference on Multimedia and Expo (ICME)

W.R., Gilks, Richardson, S., and Speigelhalter, D.J. (1996). Markov Chain Monte Carlo in Practice, Chapman and Hall, London

Boykov, Y., and Kolmogorov, V. (2004). An Experimental Comparison of Min-Cut/Max-Flow Algorithms for Energy Minimization in Vision. In IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 26, no. 9, pp. 1124-1137

Yedidia, J.S., Freeman, W.T., and Weiss, Y. (2003). Understanding belief propagation and its generalizations. Exploring articifical intelligence in the new millennium, Morgan Kaufmann Publishers Inc. San Francisco, pp. 239-269

McCullagh, P. and Nelder, J.A. (1987). Generalised Linear Models. Chapman and Hall, London

Smeaton, A.F. and Over, P. (2003). TRECVID: Benchmarking the effectiveness of information retrieval tasks on digital video. In Proc. of the International Conference on Image and Video Retrieval

Kennedy L., Hauptmann, A., Naphade, M., Smith, J.R. and Chang, S.-F. (2006). LSCOM Lexicon Definitions and Annotations Version 1.0, DTO Challenge Workshop on Large Scale Concept Ontology for Multimedia. Columbia University ADVENT Technical Report #217-2006-3