MINING RELATIONSHIP BETWEEN VIDEO CONCEPTS USING PROBABILISTIC GRAPHICAL MODELS

Rong Yan, Ming-yu Chen, Alexander Hauptmann *

School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213
{yanrong, mychen, alex+}@cs.cmu.edu

Abstract

For large scale automatic semantic video characterization, it is necessary to learn and model a large number of semantic concepts. These semantic concepts do not exist in isolation to each other and exploiting this relationship between multiple video concepts could be a useful source to improve the concept detection accuracy. In this paper, we describe various multi-concept relational learning approaches via a unified probabilistic graphical model representation and propose using numerous graphical models to mine the relationship between video concepts that have not been applied before. Their performances in video semantic concept detection are evaluated and compared on two TRECVID'05 video collections.

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, 2]. These semantic concepts cover a wide range of topics that can be roughly categorized as objects, sites, events, and specific personalities and named entities. Researchers have developed a large number of automatic concept detection techniques and the most popular approach is to translate the learning task into multiple binary classification problems with the presence/absence label of each individual concept. Then for each concept, its associated video concepts can be detected via multiple unimodal classifiers based on visual, audio and text features.

However, these binary classification approaches ignore an 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 certain co-occurrence patterns in video collections. For example, the concept "car" always co-occurs in a video shot with the concept "road" while the concept "office" is not likely to appear with "road". Such kinds of concept relationships are not rare and it can be expected that mining multi-concept relationship 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 are helpful to discover unknown concept relationship that is complementary to human prior knowledge.

*This research is partially supported by Advanced Research and Development Activity (ARDA) under contract number H98230-04-C-0406 and NBCHC040037.

To automatically exploit benefits from the multi-concept relationship, several approaches have been proposed before which are built upon the advanced pattern recognition techniques within a probabilistic framework. For example, Naphade et al. [3] explicitly modeled the linkages between various semantic concepts via a Bayesian network that implicitly offered ontology semantics underlying the video collection. Cees et al. [4] proposed an semantic value chain architecture including a multi-concept learning layer called context link. At the top level, it aims at merging the results of content outputs from concept detectors. Two configurations were explored where one was based on a stacked classifier upon a context vector and the other was based on ontology with some common sense rules. Alex et al. [5] fused the multi-concept predictions and captured the inter-concept causations by constructing an additional logistic regression classifier atop the uni-concept detection results. Amir et al. [6] 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. A ontology-based multi-classification algorithm was proposed by Wu et al. [7] which attempted to model the possible influence relations between concepts based on a predefined ontology hierarchy.

Although such a huge effort is made to learn the multi-concept relationship, a number of well established and successful probabilistic graphical models (especially the undirected graphical models) such as restricted Boltzman machines(RBM) and conditional random field(CRF) have never been applied in the domain of video annotation. Moreover, since extant methods have been evaluated in different testbeds, the advantages and disadvantages between these approaches remains unclear. In this paper, we describe several multi-concept relational learning approaches via a unified probabilistic graphical model representation and propose using numerous graphical models to mine the relationship between video concepts that have not been applied before. Their effectiveness in video semantic concept detection is evaluated and compared on two TRECVID'05 video collections.

2. GRAPHICAL MODEL REPRESENTATIONS FOR VIDEO CONCEPTS

Many multi-concept learning approaches can be concisely represented in form of probabilistic graphical models that express dependencies among random variables by a graph in which each random variable is a node. There are two types of graphical models including directed graphical models (a.k.a. Bayesian network)

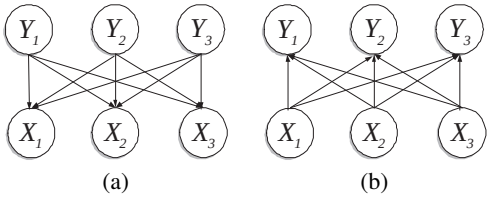

Fig. 1. Two directed graphical models (a.k.a. Bayesian network) for multi-concept learning including (a) a generative model assuming concepts generate detection outputs, and (b) a discriminative model directly modeling the conditional probability of concepts given detection outputs.

which represents a factorization of the joint probability of all random variables and undirected graphical models in which graph separation encodes conditional independencies between variables. In this paper, we consider building graphical models between concepts and predictions generated from existing uni-modal semantic detectors rather than the low-level video features, because we want to reduce the feature dimension and computational efforts in the learning process. Formally for a specific video shot, let $X_i \in \mathcal{R}$ denote the observations of i^{th} uni-modal semantic detector, $Y_j \in \{0,1\}$ denote the presence/absence labels of j^{th} concept. $\mathbf{X}, \mathbf{Y}$ represent the vectors of $\{X_i\}$, $\{Y_j\}$. For the purpose of parameter estimation, we assume there are D training data with truth annotations $\{\mathbf{X}_d, \mathbf{Y}_d\}$. In this setting, the purpose of concept detection is to predict the hidden concept labels from visible observations provided by uni-modal classifiers, i.e., estimate the conditional probabilities of $P(Y_j|\mathbf{X})$ under a given model representation. In the following discussions, we discuss some existing models and propose several new models for mining the multi-concept relationship using both the directed graphical models and the undirected graphical models.

2.1. Directed Graphical Models

Most previous approaches belong to the category of directed graphical models. Essentially, all these models can be understood as a two-layer directed graphical model with one layer of hidden units and one layer of input units connected by fully-linked edges. Figure 1 shows several examples of directed graphical models for video concept mining. Among them, Figure 1(a) corresponds to a generative model(**BNG**) that assumes the detection outputs are generated by concept variables. One such example is the Bayesian network version of the multi-net model proposed by Naphade [3]. In this model, the hidden layer can be taken as a representation of the "latent concept aspects" and the input layer corresponds to the observed predictions of uni-modal semantic detectors. This graph naturally implies the conditional independence of predictions X_i given the concepts $\mathbf{Y}$. Typically, the prior distribution of Y_j is modeled as a Bernoulli distribution $Bernoulli(p_j)$ with a draw probability p_j and the conditional probability of $X_i|Y_j$ is modeled as a Gaussian distribution with mean $\sum_j w_{ij} y_j$ and variance σ_i^2. The model parameters can be learned based on the maximum likelihood estimation(MLE),

$$\begin{aligned}(W,\Sigma,P) &= \arg\max \sum_d \log P(x_{d1},...,x_{dM},y_{d1},...,y_{dN}) \\ &= \arg\max \sum_d \sum_i \log P(x_{di}|\mathbf{Y}) + \sum_d \sum_j \log P(y_{dj})\end{aligned}$$

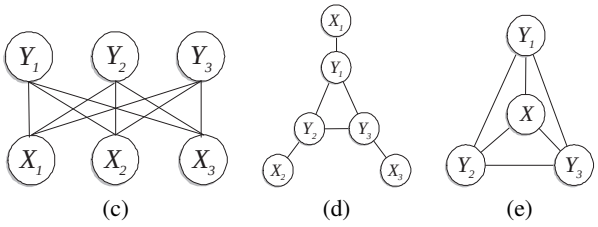

Fig. 2. Three undirected graphical models for multi-concept learning including (c) restricted Boltzmann machines, (d) Markov random fields and (e) conditional random fields.

By setting the derivatives of likelihood function with respect to each parameter to be zero, we can derive the maximum likelihood estimators in an analytical form where the estimated parameters are shown as follows,

$$p_j^* = \frac{\sum_d I(y_{dj}=1)}{D} \quad (1)$$

$$w_{ij}^* = \arg\max_w \sum_d (x_i - \sum_j w_{ij} y_j)^2, j = 1...N \quad (2)$$

$$\sigma_i = std(x_i) \quad (3)$$

Note that the estimation of parameter w_{ij} is equivalent to a linear regression on X_i with underlying variables $Y_1,...,Y_N$. Although the parameter estimation process is quite simple, it is usually intractable to infer the conditional probability of $\mathbf{Y}$ given $\mathbf{X}$ due to lack of conditional independence between $\mathbf{Y}$. Therefore, we adopt a popular approximate inference technique called Gibbs sampling, which is applicable when the joint distribution is not known explicitly but the conditional distribution of each variable is able to be computed.[1] According to Gibbs sampling, we can repeatedly sample the following conditional probability to approximate the joint distribution and then compute the expectation of labels $\mathbf{Y}$,

$$Y_j \sim P(Y_j = 1|\mathbf{X}, \mathbf{Y} \setminus Y_j) = \frac{P(Y_j=1,\mathbf{X},\mathbf{Y})}{P(Y_j=0,\mathbf{X},\mathbf{Y}) + P(Y_j=1,\mathbf{X},\mathbf{Y})}$$

The model in Figure 3(b) corresponds to another type of directed graphical models which directly model the conditional probability of $\mathbf{Y}$ given $\mathbf{X}$, or called discriminative models(**BND**). It can used to describe the approaches proposed in [5, 4]. Unlike the previous models, this graph reversely implies the conditional independence of predictions Y_i given the observations $\mathbf{X}$ and thus results in an fast inference process. In practice, the labels Y are usually modeled by a logistic regression based on observation variables X where,

$$P(\mathbf{Y}|\mathbf{X}) \propto \exp\left[\sum_i (\alpha_i + \sum_j w_{ij} x_j) y_i\right]$$

With an additional validation set, the parameters can be estimated by using any gradient descent methods such as iterative reweighted least squares(IRLS) algorithm [8].

[1] The Gibbs sampling algorithm is to generate an instance from the distribution of each variable in turn, conditional on the current values of the other variables. It can be shown that the sequence of samples comprises a Markov chain, and the stationary distribution of that Markov chain is just the sought-after joint distribution.

2.2. Undirected Graphical Models

The Bayesian network formalism offers clear causal semantics and manipulability from a modeling point of view. However, as pointed out by [9], inference of the latent concepts in such models can be prohibitively expensive due to the conditional dependencies between all hidden variables. This drawback could seriously affect the model performance in real-time prediction tasks and in EM-based learning. Moreover, directed models have to explicitly retain the casuality between different observed/hidden variables and thus it can lead to a sophisticated network structure if we want to incorporate additional dependency between concepts.

As alternatives of directed graphical models, undirected graphical models could be a better formalism for handling the relation between concepts without explicitly imposing the concept casuality. However to our surprise, the options of applying undirected models to video annotation have seldom been explored before. One example of undirected models is shown in the Figure 2(c) called the restricted Boltzmann machine(**RBM**)(a.k.a. harmoniums) which can be viewed as an undirected counterpart of the aforementioned directed concept models with the arrows of the edges removed. In a RBM model, observations X are fully connected with the concept presence Y in form of a bipartite graph. According to the undirected model semantics, there is no *marginal* independence for either input or hidden variables. However, it enjoys the advantages of *conditional* independence between hidden variables given observed variables, which is generally violated in the directed models. This property can greatly reduce inference cost although it comes at a price of a more difficult learning process due to the presence of a global partition function. Formally, if we want to generate the conditional probabilities as follows,

$$P(x_i|\mathbf{Y}) = \mathcal{N}\left[\sigma_i^2(\beta_i + \sum_j w_{ij}y_j), \sigma_i^2\right], P(y_j|\mathbf{X}) = \mathcal{S}\left[\alpha_j + \sum_i w_{ij}x_i\right]$$

where $N(\mu,\sigma)$ is a normal pdf function with mean μ and variance σ, and $\mathcal{S}(x)$ is a logistic function $1/(1+e^{-x})$, then we have to define the the joint probability of $\mathbf{X},\mathbf{Y}$ to be,

$$P(\mathbf{X},\mathbf{Y}) \propto \exp\left[-\frac{1}{2}\sum_i \frac{x_i^2}{\sigma_i^2} + \sum_i \beta_i x_i + \sum_j (\alpha_j + \sum_i w_{ij}x_i)y_j\right]. \quad (4)$$

The gradient-descent learning rules can be obtained by taking derivatives of the log-likelihood defined by Eq. 4 with respect to the model parameters. It can be found that the gradient of each parameter are equivalent to the difference between expectation of its corresponding potential under empirical distribution and that under model distribution [9]. However, the expectations under the model distribution are usually difficult to compute because of the intractable normalization factor. Therefore, we have to utilize some approximate inference approaches such as loopy belief propagation, contrastive divergence(CD) and variational methods. In practice, we adopt the contrastive divergence as the basic inference method [9] which approximates the intractable model distribution using a single or a few iterations of Gibbs sampling, and is therefore highly efficient.

Until now, we have only discussed two-layer bipartite graphical models with the nodes in each layer are fully connected with the nodes in the other layer. Rather than modeling the concept relationship in such an indirect way, the flexibility of undirected models allow us to impose the links directly on the concept nodes which cannot be easily achieved by a directed model. We considered two possibilities of such kinds of model designs in the following discussions. Figure 2(d) corresponds to a Ising-model like Markov random field(**MRF**) where the concept nodes are fully linked and the observations only interact with their corresponding concepts. Based on the model semantics and similar potential definitions as before, the joint probability of the observations and labels can be represented as follows,

$$P(\mathbf{X},\mathbf{Y}) \propto \exp\left[-\frac{1}{2}\sum_i \frac{x_i^2}{\sigma_i^2} + \sum_i \beta_i x_i + \sum_i (\alpha_i + w_i x_i + \sum_j u_{ij}y_j)y_i\right].$$

The major difference between above equation and Eq. 4 lies in the pairwise interaction terms of $u_{ij}y_jy_i$ which directly captures the concept co-occurrence patterns. The maximum likelihood estimation of this model can also be achieved by contrastive divergence. Note that an advanced version of the multi-net model based on the factor graph model [3] can be viewed as a variant of above model with slight differences in the model presentation. Figure 2(e) plots a more recently developed graphical model called the conditional random field(**CRF**) which is a random field globally conditioned on the observations X. It means the observations Y, when conditioned on X, obeys the Markov property with respect to the undirected graph in Figure(e). According to the Hammersley Clifford theorem and assuming only the pairwise clique potential are nonzero, we can define the joint probability as,

$$P(\mathbf{Y}|\mathbf{X}) \propto \exp\left[\sum_i (\alpha_i + \sum_j w_{ij}x_j)y_i + \sum_i \sum_k u_{ik}y_iy_k\right]$$

The conditional random field takes the advantage of modeling the conditional probability of concepts given observations and thus avoid the problem of learning complex class density. Due to the space limit, please refer to [10] for the details of learning and inference in the conditional random fields.

3. EXPERIMENTAL RESULTS

We evaluated all five multi-concept detection models using the TRECVID'05 [1] development data. The development data are split into three parts, where 70% as the training set to generate the concept detection outputs, 15% as the validation set to learn the multi-concept relationship, and remaining 15% as the testing set to evaluate the detection performance. As mentioned before, one application of multi-concept learning models is to automatically discover the co-occurrence patterns in a specific video collection. To illustrate this, we plotted the Figure 3 which shows the relationships between 17 concepts found by CRF with each grid indicating a pair of concepts. As can be seen, there are a considerable amount of strong correlations between semantic concepts in the video collection, including both positive interactions (two concepts are positively correlated with each other) and negative interactions ((two concepts are negatively correlated). We believe the concept detection task should be able to benefit from capturing this semantic context. In more detail, some of the strongest positive/negative concept pairs are listed below,

Positive Pairs: (outdoor, building), (urban, building), (person, face), (studio, maps), (car, road), (urban, road), (text, sports)

Negative Pairs: (sports, building), (outdoor, computer tv screen), (outdoor, maps), (commercial, studio), (waterfront, urban)

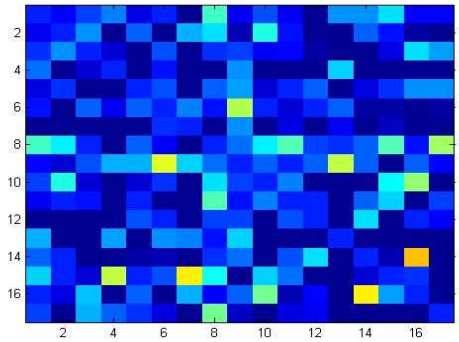

Fig. 3. Illustration of inter-concept relationships. Each grid indicates a pair of concepts. Lighter colors stand for stronger positive relations and darker colors stand for stronger negative relations.

However, we also notice that not every concept can exhibit co-occurrence patterns with other concepts due to the limited number of training data. It would be beneficial to remove those isolated concepts in the training data before the learning process. By conducting the χ^2 test between every pair of concepts, we eliminated the concepts that do not have any χ^2 scores exceeding certain thresholds and thus not strongly correlated to others. Finally, we constructed a five-concept collection and an eleven-concept collection that include the sets of concepts as follows,

5-concept: car, face, person, text, walking/running

11-concept: building, car, face, maps, outdoor, person, sports, studio, text, urban, walking/running

Table 1 shows the mean average precision on the testing set of the five graphical models discussed before and the baseline obtained without taking any conceptual relations into account. We can observe that the best multi-concept modeling approaches can usually bring an additional 2-3% improvement over the baseline performance in terms of mean average precision. Table 1 also lists the number of concepts that have better and worse performances than baseline. It can be found that the detection accuracy of each concept is more likely to be improved with aid of multi-concept relational modeling, which shows the effectiveness of incorporating contexts into the detection results. The undirected graphical models (i.e., RBM, MRF and CRF) demonstrate their promising potentials in the task of concept detection given the high MAP on both datasets. But on the other hand, our experimental results also show the inconsistency of the performance of various models. For example, in the 5-concept dataset, BND and RBM are among the best models with similar MAPs around 60%. But in contrast, in the 11-concept dataset MRF and CRF provide the best performance around 52%. After an in-depth analysis on this dataset, we found that the inferior performances of BND and RBM mainly come from some significant degradations on one or two concepts even they can improve on the others, which might indicate their instabilities in handling a large amount of concepts. However, it is worth pointing out that so far the differences between models and baseline are not statistically significant yet. Further evaluations are suggested to provide more insights on their comparison.

Method	5-concept collection			11-concept collection		
	Better	Worse	MAP	Better	Worse	MAP
Base	-	-	0.5715	-	-	0.4994
BNG	2	3	0.5742	5	6	0.5187
BND	4	0	0.6036	6	4	0.4996
RBM	4	0	0.6024	5	5	0.4822
MRF	3	1	0.5714	6	4	0.5210
CRF	3	1	0.5882	7	3	0.5211

Table 1. Performance w.r.t. multi-concept relational learning models and their baseline. Better/Worse means how many concepts have a better/worse performance than baseline. MAP means the mean average precision of the learning methods.

4. CONCLUSIONS AND FUTURE DIRECTIONS

In this paper, we describe several approaches for mining the relationship between video concepts via a unified probabilistic graphical model representation and propose using numerous graphical models that have not been applied to this task before. Two types of graphical models have been discussed including two directed models and three undirected models. While most previous work can be generalized by the direct graphical models semantics, few attentions have been paid to the undirected models which do not need to impose casuality between concept nodes and have gained their success in the field of machine learning. Our experiments provide a fair comparison between all these approaches on two video collections. They show the effectiveness and potentials of using undirected models to learn the concepts relations. In future, we would like to conduct more studies to validate our conclusions.

5. REFERENCES

[1] A.F. Smeaton and P. Over, "TRECVID: Benchmarking the effectiveness of information retrieval tasks on digital video.," in *Proc. of the Intl. Conf. on Image and Video Retrieval*, 2003.

[2] A. G. Hauptmann, R.V. Baron, M.-Y. Chen, M. Christel, P. Duygulu, C. Huang, R. Jin, W.-H. Lin, T. Ng, N. Moraveji, N. Papernick, C.G.M. Snoek, G. Tzanetakis, J. Yang, R. Yan, , and H.D. Wactlar, "Informedia at TRECVID 2003: Analyzing and searching broadcast news video," in *Proc. of TRECVID*, 2003.

[3] M. R. Naphade, T. Kristjansson, B. Frey, and T.S. Huang, "Probabilistic multimedia objects (multijects): A novel approach to video indexing and retrieval in multimedia systems," in *Proc. of ICIP*, 1998.

[4] C.G.M. Snoek, M. Worring, J.M. Geusebroek, D.C. Koelma, and F.J. Seinstra, "The mediamill TRECVID 2004 semantic viedo search engine," in *Proc. of TRECVID*, 2004.

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

[6] A. Amir, W. Hsu, G. Iyengar, C.-Y.Lin, M. Naphade, A. Natsev, C. Neti, H. J. Nock, J. R. Smith, B. L. Tseng, Y. Wu, and D. Zhang, "IBM research TRECVID-2003 video retrieval system," in *NIST TRECVID-2003*, Nov 2003.

[7] Y. Wu, B. L. Tseng, and J. R. Smith, "Ontology-based multi-classification learning for video concept detection," in *IEEE International Conference on Multimedia and Expo (ICME)*, 2004.

[8] M. I. Jordan and R. A. Jacobs, "Hierarchical mixtures of experts and the EM algorithm," *Neural Computation*, vol. 6, pp. 181–214, 1994.

[9] E. P. Xing, R. Yan, and A. G. Hauptmann, "Mining associated text and images using dual-wing harmoniums," in *Uncertainty in Artifical Intelligence (UAI)'05*, 2005.

[10] J. Lafferty, A. McCallum, and F. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," in *Proc. 18th Intl. Conf. on Machine Learning*. 2001, pp. 282–289, Morgan Kaufmann, San Francisco, CA.