

Collective Media Annotation using Random Field Models

Matthew Cooper
FX Palo Alto Laboratory
Palo Alto, CA USA
cooper@fxpal.com

Abstract

We present methods for semantic annotation of multimedia data. The goal is to detect semantic attributes (also referred to as concepts) in clips of video via analysis of a single keyframe or set of frames. The proposed methods integrate high performance discriminative single concept detectors in a random field model for collective multiple concept detection. Furthermore, we describe a generic framework for semantic media classification capable of capturing arbitrary complex dependencies between the semantic concepts. Finally, we present initial experimental results comparing the proposed approach to existing methods.

1. Introduction

Substantial current multimedia analysis research focuses on information retrieval for video content. As media collections move onto the internet, web search companies are extending their text-based search capabilities to video data. These systems typically rely on the link structure and text on the web pages containing the videos to index content. Video search and retrieval has also been the focus of the highly successful TRECVID workshops [10]. Although the use of visual information is emphasized in the TRECVID evaluations, extracting semantics from visual data in the absence of textual descriptors remains a major challenge.

Recent work to address this semantic gap has concentrated on ontology-based approaches to semantic feature extraction [3, 13]. In this work, a “basis” set of binary classifiers are built to determine if a video shot exhibits a specific semantic feature. These classification outputs are combined statistically to provide higher-level analysis or enhance indexing and retrieval. Many of these approaches operate at the shot-level following an initial segmentation. This is desirable for computational efficiency, dynamic analysis of local sets of frames, and for extraction of semantic features that exhibit some temporal duration.

Simultaneously, manual tags are now proliferating on

various shared video and image sites such as Flickr and YouTube. While this information is of tremendous potential value for video indexing, such as for refining and training automatic systems, it also presents challenges. Lengthy videos can have tags that apply only to a small (unidentified) portion of the video. Also, the classic problems of polysemy and synonymy in text categorization (e.g. [1]) are inherited in aggregating tag data for multimedia categorization.

To supply consistent and reliable annotations for indexing or metadata creation, a semi-automatic approach is required. Manual annotation is not feasible on the scale of legacy assets, let alone for the ever increasing amounts of newly produced content. Automatic techniques that scale can accommodate the quantity of the data, but the quality of automatic annotations is not sufficient. We can identify several requirements for the analysis components of an ideal media annotation system:

- The ability to integrate heterogeneous modalities in a common framework.
- A generic architecture suitable for a wide array of types of annotations which vary considerably in their sparseness in data sets.
- The ability to supply a confidence measure or ranking associated with annotations to support manually revising results as needed. It should be possible to provide such measures in time-varying forms at various temporal resolutions.
- A fully automatic mode for annotating either archival data, or data for which manually supplied tag information, text transcripts, or web page link structure is absent.

We review related work and propose a framework below to work towards these aims. While we focus on annotating video, the system is broadly applicable to digital media collections of various modalities.

2. Independent concept detection

The problem addressed here is the automatic annotation of temporally segmented video using a set of binary attributes. We refer to these attributes as concepts, and the general problem as concept detection. We describe a system to jointly detect the presence/absence of a set of concepts by exploiting concept interdependence. General sets of such attributes include those used in [15, 9]. We now describe the integration of discriminative single label classifiers in a framework for collective multimedia annotation in several computational phases.

2.1. Feature extraction

In the first processing step, low-level feature data must be extracted. In our case, the source video is segmented according to shots, and keyframes are identified within each shot. The keyframes are processed to extract low-level feature data. This data may include visual features such as color histograms, texture or edge features, motion analysis or face detection output. If time-aligned text data is present, we can also include standard low-level text features such as word counts or tf/idf features [12]. The specific set of features used is not critical. In the multi-modal case, early or late fusion approaches can be used to construct the single concept classifiers [14]. We assume their availability for the construction of the probabilistic model detailed below.

2.2. Classification

In the second step, we train a discriminative classifier for each concept using a labeled training set of low-level features. For several years, support vector machines (SVMs) have been the preferred classifier at TRECVID, and we use them here. The output of each SVM is transformed to a probability using a logistic mapping to provide a system for independent concept detection. In the experiments below, we use the publicly available low-level features (described in [16]) and SVM outputs provided by the MediaMill team [15]. These SVM outputs are *highly* optimized using a computationally intensive grid search for classification parameters on a per concept basis. These outputs represent an extremely competitive baseline annotation system.

The graphical model corresponding to this approach appears in Figure 1(a). Probabilistically, if the i^{th} concept in the concept set \mathcal{C} is denoted by the binary random variable Y_i , then

$$P(\mathbf{Y}|\mathbf{X}) = \prod_{i \in \mathcal{C}} P(Y_i|\mathbf{X}) . \quad (1)$$

for low level feature data X .

3. Collective concept detection

We now build collective annotation models using the single concept discriminative models described in the previous section. Our first approach builds on the discriminative random field (DRF) model of Kumar and Hebert [6]. This model combines discriminative single concept classifiers with pairwise concept co-occurrence features representing contextual information. Their goal is to perform collective binary classification of pixel blocks in images as either “natural” or “man-made.” For per-block classification, they use logistic regression. The random field model incorporates spatial dependencies. More specifically, they model the probability of the (collective) vector of binary labels \mathbf{Y} given the low-level image block data \mathbf{X} as

$$P(\mathbf{Y}|\mathbf{X}) = \frac{1}{Z(\mathbf{X})} \exp \left(\sum_{i \in \mathcal{S}} A_i(Y_i, \mathbf{X}) + \sum_{i \in \mathcal{S}} \sum_{j \in \mathcal{N}_i} I_{ij}(Y_i, Y_j, \mathbf{X}) \right) . \quad (2)$$

The terms A_i are the association potentials (or unary or node potentials) which are local discriminative classifiers for each spatial location i in the set \mathcal{S} . I_{ij} is the term representing the interaction between the spatial locations i, j . \mathcal{N}_i is the set of neighbors of node i in the graph. [7] details learning and inference methods for the broad class of conditional random fields, including DRFs.

We adapt this approach to collective semantic concept detection in a video clip. Specifically, we use the SVM classifiers of Section 2 for the association terms in (2). For this, we map each SVM output to the corresponding probability $P_D(Y_i = 1|\mathbf{X})$ following [11]. The subscript D denotes the single concept (discriminative) classifiers output probabilities¹. We then set the association term for concept Y_i to be:

$$A_i(Y_i, \mathbf{X}) = \log(P_D(Y_i = 1|\mathbf{X})) \quad (3)$$

so that in the absence of interaction terms ($I_{ij} = 0$), (2) reverts to the independent per-concept SVM models.

Next we must identify which concepts are related, i.e. which concepts are connected by an edge in our graph. For this, [17] performs a chi-squared test using the ground truth labeling of the training set. They connect each concept to the other concepts to which it has the most statistically significant relationships. The resulting graph defines the neighborhoods \mathcal{N}_i for each concept Y_i . This is surely not optimal. However, learning optimal unstructured graphs in the general case is NP-hard. Nonetheless, we feel that evaluating other heuristics and approximation schemes for

¹ $P_D(Y_i|\mathbf{X})$ does not equal the marginal probability $P(Y_i|\mathbf{X})$ computed from $P(\mathbf{Y}|\mathbf{X})$ from (2).

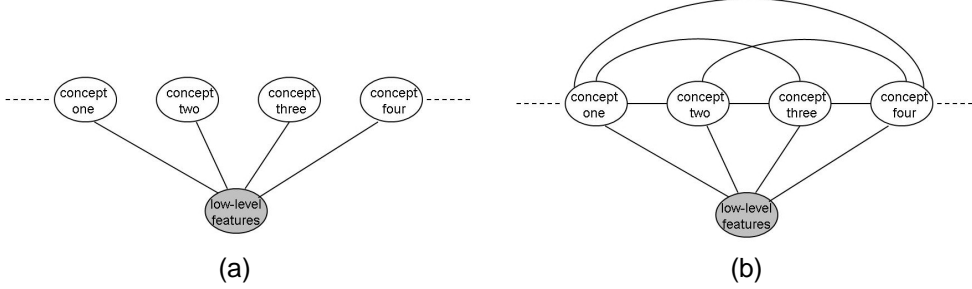


Figure 1. An example of graphical models used for detecting four concepts. Panel (a) shows the model for independent classification, while panel (b) depicts collective classification.

graph induction is both critical to the success of these methods and a promising area for future work.

The interaction potentials in [6] were inspired by long-standing work in image analysis using Markov random fields. Here, we define interaction potentials building on recent work in text categorization [2]. First, we rewrite (2) to clarify our notation:

$$P(\mathbf{Y}|\mathbf{X}) = \frac{1}{Z(\mathbf{X})} \exp \left(\sum_{i \in \mathcal{C}} \log(P_D(Y_i|\mathbf{X})) + \sum_{i \in \mathcal{C}} \sum_{j \in \mathcal{N}_i} \sum_{k \in \mathcal{K}} \lambda_{ij}^{(k)} f_{ij}^{(k)}(Y_i, Y_j, \mathbf{X}) \right). \quad (4)$$

We use the notation P_D to distinguish the probability mapped from the single concept SVM output from the corresponding marginal probability of the joint model. Comparing this equation to (2), we see that our interaction term is the linear form:

$$\begin{aligned} I(Y_i, Y_j, \mathbf{X}) &= \sum_{k \in \mathcal{K}} \lambda_{ij}^{(k)} f_{ij}^{(k)}(Y_i, Y_j, \mathbf{X}) \\ &= \Lambda_{ij}^T \mathbf{F}_{ij}(Y_i, Y_j, \mathbf{X}). \end{aligned} \quad (5)$$

We finally note that we can rewrite (4) compactly as:

$$P(\mathbf{Y}|\mathbf{X}) = \frac{1}{Z(\mathbf{X})} \exp \left(\mathbf{P}_D(\mathbf{X}) + \sum_{i \in \mathcal{C}} \sum_{j \in \mathcal{N}_i} \Lambda_{ij}^T \mathbf{F}_{ij}(Y_i, Y_j, \mathbf{X}) \right). \quad (6)$$

Here the i^{th} element of $\mathbf{P}_D(\mathbf{X})$ is $\log(P(Y_i|\mathbf{X}))$. Likewise, the k^{th} elements of Λ_{ij} and $\mathbf{F}_{ij}(Y_i, Y_j, \mathbf{X})$ are $\lambda_{ij}^{(k)}$ and $f_{ij}^{(k)}(Y_i, Y_j, \mathbf{X})$, respectively. This form shows that the random field model is simply a log-linear classifier. For maximum likelihood model training, the gradients of the log-likelihood thus take a standard form (e.g. [6]).

3.1. The CML+I model

We now detail the interaction potential functions for two models. The first model, denoted collective multi-label interaction (CML+I), captures inter-concept co-occurrence. These features are defined for each pair of concepts Y_i, Y_j that are connected in our graph (i.e. not for all pairs). Thus, we have the indexed family of interaction potential functions:

$$\begin{aligned} f_{ij}^{(0)}(Y_i, Y_j, \mathbf{X}) &= \begin{cases} 1 & Y_i = Y_j = 0 \\ 0 & \text{otherwise} \end{cases} \\ f_{ij}^{(1)}(Y_i, Y_j, \mathbf{X}) &= \begin{cases} 1 & Y_i = 1, Y_j = 0 \\ 0 & \text{otherwise} \end{cases} \\ f_{ij}^{(2)}(Y_i, Y_j, \mathbf{X}) &= \begin{cases} 1 & Y_i = 0, Y_j = 1 \\ 0 & \text{otherwise} \end{cases} \\ f_{ij}^{(3)}(Y_i, Y_j, \mathbf{X}) &= \begin{cases} 1 & Y_i = Y_j = 1 \\ 0 & \text{otherwise} \end{cases} \end{aligned} \quad (7)$$

By modeling the four possible combinations separately, we hope to capture all types of pairwise co-occurrence within the model. For example, the concept “urban” can be expected to occur when “outdoor” is one (true), however, “outdoor” may occur when “urban” is zero (false). The above features allow the model to distinguish the three cases in which either “urban” or “outdoor” or both are false. For this model the index set for the interaction potentials is simply $\mathcal{K} = \{0, 1, 2, 3\}$. This interaction model was proposed in [2] for text categorization without discriminatively trained association potentials.

3.2. The CMLF+I model

We define a second model to capture concept-feature co-occurrence, combining ideas from [2] and [4] to define

the collective multi-label-feature interaction (CMLF+I). We first quantize the low level visual features \mathbf{X} across the training set using k-means (or any other vector quantization technique). Denote this discrete representation for the low-level features as $\mathcal{Q}(\mathbf{X}) \in \{0, \dots, Q\}$. For this model, we define interaction potentials:

$$\begin{aligned} f_{ij}^{(q,0)}(Y_i, Y_j, \mathbf{X}) &= \begin{cases} 1 & Y_i = Y_j = 0, \mathcal{Q}(\mathbf{X}) = q \\ 0 & \text{otherwise} \end{cases} \\ f_{ij}^{(q,1)}(Y_i, Y_j, \mathbf{X}) &= \begin{cases} 1 & Y_i = 1, Y_j = 0, \mathcal{Q}(\mathbf{X}) = q \\ 0 & \text{otherwise} \end{cases} \\ f_{ij}^{(q,2)}(Y_i, Y_j, \mathbf{X}) &= \begin{cases} 1 & Y_i = 0, Y_j = 1, \mathcal{Q}(\mathbf{X}) = q \\ 0 & \text{otherwise} \end{cases} \\ f_{ij}^{(q,3)}(Y_i, Y_j, \mathbf{X}) &= \begin{cases} 1 & Y_i = Y_j = 1, \mathcal{Q}(\mathbf{X}) = q \\ 0 & \text{otherwise} \end{cases} \quad (8) \end{aligned}$$

The index set for the interaction potentials is $\mathcal{K} = \{(q, i) : 0 \leq i \leq 3, 0 \leq q \leq Q\}$. Like CML+I, this model distinguishes among the four possible combinations of each pair of labels. It extends CML+I to distinguish among each label combination in conjunction with a low-level feature observation. For example, if we observe the quantized low-level feature value q , which represents most observations with large green regions in the frame (i.e. vegetation), we may expect the model weights for features with the ‘‘outdoor’’ concept one (true) and the ‘‘urban’’ concept to be zero (false). In this manner, content can be used to more finely model the inter-concept relationships.

4. Experiments

Here we summarize initial experiments comparing various proposed systems using conditional random fields using the TRECVID 2005 development data for the high-level concept detection task. The training and test sets each include more than 6000 video shots from various broadcast news sources collected in 2004. We also use the graphs in [17] for direct comparison:

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

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

The experimental results are summarized in Figure 2 for the 5 concept graph (panel (a)) and the 11 concept graph (panel (b)). The performance measure is mean average precision (MAP) which averages precision at each level of recall for each concept, and then computes the mean (of the average precisions) over the set of concepts [10].

4.1. Model training and inference

We train the systems to maximize the likelihood of the training set to fit parameters. The log-likelihood is

$$\mathcal{L}(\mathcal{D}) = \sum_{d \in \mathcal{D}} \left(\sum_{i \in \mathcal{C}} \log(P_D(Y_i^{(d)} | \mathbf{X}^{(d)})) + \sum_{i \in \mathcal{C}} \sum_{j \in \mathcal{N}_i} \sum_{k \in \mathcal{K}} \lambda_{ij}^{(k)} f_{ij}^{(k)}(\mathbf{Y}^{(d)}, \mathbf{X}^{(d)}) - \log(Z(\mathbf{X}^{(d)})) \right). \quad (9)$$

This gives the gradient equations:

$$\frac{d\mathcal{L}}{d\lambda_{ij}^{(k)}} = \sum_{d \in \mathcal{D}} \left(f_{ij}^{(k)}(Y_i^{(d)}, Y_j^{(d)}, \mathbf{X}^{(d)}) - \sum_{Y_i, Y_j} f_{ij}^{(k)}(Y_i, Y_j, \mathbf{X}^{(d)}) P(Y_i, Y_j | \mathbf{X}^{(d)}) \right). \quad (10)$$

Here we use $\mathbf{Y}^{(d)}$ to denote the ground truth for training sample $\mathbf{X}^{(d)}$. Also, $Y_i^{(d)}, Y_j^{(d)}$ denote the true values for concepts $i, j \in \mathcal{C}$ for sample $\mathbf{X}^{(d)}$ while Y_i, Y_j denote binary variables of integration. As demonstrated in [8] limited memory conjugate gradient methods greatly accelerate training maximum entropy and log-linear classifiers. We employ the Broyden-Fletcher-Goldfarb-Shanno (BFGS) minimization routine for maximum likelihood model training.

We presently use exhaustive inference, which entails evaluating $P(\mathbf{Y} | \mathbf{X})$ for each *observed* $\mathbf{Y}^{(d)}, d \in \mathcal{D}$ in training, and marginalizing to compute $P(Y_i | \mathbf{X}) \forall i \in \mathcal{C}$. This is generally prohibitive due to the exponential growth in concept combinations with the number of concepts. In the present context, the number of observed combinations is much smaller. For example, we observe fewer than 200 combinations in the experiments using the 11 concept graph (as opposed to the 2048 possible combinations).

4.2. Additional system descriptions

We have two experimental baselines. The first, denoted CMU, shows the results from [17] which proposed the following conditional random field model:

$$P(\mathbf{Y} | \mathbf{X}) = \frac{1}{Z(\mathbf{X})} \exp \left(\sum_{i \in \mathcal{C}} (\alpha_i + \mathbf{V}_i^T \mathbf{X}) Y_i + \sum_{i \in \mathcal{C}} \sum_{j \in \mathcal{N}_i} \lambda_{ij} Y_i Y_j \right). \quad (11)$$

The key differences between our approach and theirs are

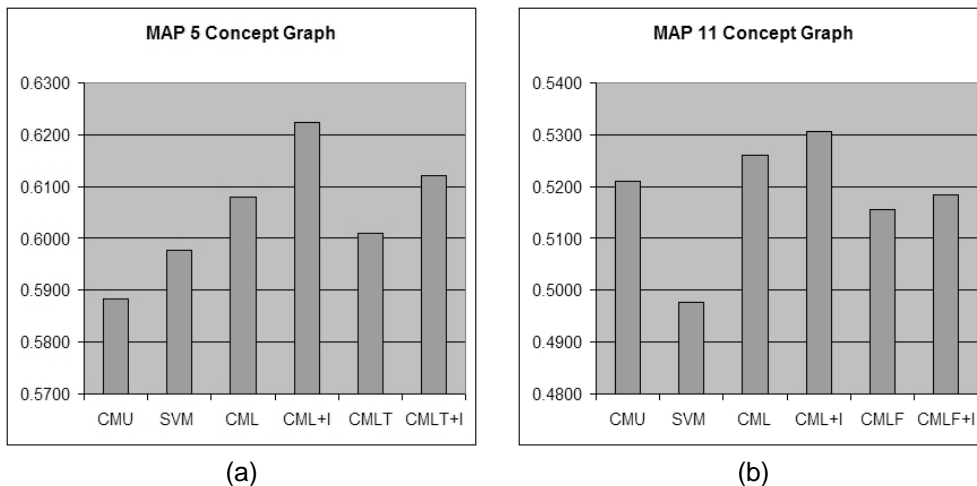


Figure 2. Experimental results comparing various system variants. Panel (a) shows results for the 5 concept graph, while panel (b) shows results for the 11 concept graph.

1. We use a probabilistic mapping of an SVM output in place of their linear association term.
2. Their interaction term does not distinguish between the three cases in which $Y_i \cdot Y_j = 0$.
3. Their interaction term does not account for feature-concept interaction as in the CMLF+I systems.

It should be noted we did not use the same training/test data split as [17]. The second, denoted SVM, shows the results of using the discriminative SVM outputs for independent concept detection as provided by [15].

To isolate the contributions of the discriminative per-label classifiers, we use the following interaction potential as used in the CMU system, to build the collective multi-label (CML) system with (4):

$$f_{ij}(Y_i, Y_j, \mathbf{X}) = \begin{cases} 1 & Y_i = Y_j = 1 \\ 0 & \text{otherwise} \end{cases} \quad (12)$$

For this case there is a single feature for each edge (Y_i, Y_j) , and the summation over k degenerates to a single term (so we drop the index k above). This system differs from the CMU system only in its use of the SVMs for the discriminative classifier terms. The CML system (0.6089) does better than SVM system (0.5977) and the CMU system (0.5882) for the 5 concept graph. It also outperforms both systems on the 11 concept graph (CMU=0.5211, SVM=0.4975, CML=0.5262).

Next, we add the more complete interaction model of (7) to (4) to form the system denoted CML+I. This system accounts for different types of inter-concept relationships and

performs at the highest level of all systems in both cases, with MAP of 0.6228 and 0.5307 for the 5 concept and 11 concept graphs, respectively.

[2] includes another closely related model that we include in our experimental comparison. Their system is denoted CMLF, and is similar to CMLF+I. It uses feature-concept interaction potentials of the form:

$$f_{ij}^{(q)}(Y_i, Y_j, \mathbf{X}) = \begin{cases} 1 & Y_i = Y_j = 1, \mathcal{Q}(\mathbf{X}) = q \\ 0 & \text{otherwise} \end{cases} \quad (13)$$

As included in our experiments, we use the SVMs to provide the per-concept association potentials. [2] used log-linear association potentials. The main differences between CMLF and CMLF+I are:

1. We quantize the low-level features extracted from the video shot or keyframe, so that the resulting interaction potentials are indexed by the concept pair (i, j) and the quantized value $\mathcal{Q}(\mathbf{X})$. Their context is text categorization, so the potentials are indexed by the concept pair (i, j) and the term (word) w corresponding to the element in the term vector $\mathbf{X}(w)$. So there is no quantization step, although it's straightforward to use term clustering to effect quantization in the text domain.
2. We consider features corresponding to all four possible combinations for Y_i and Y_j .

The CMLF outperforms the SVM baseline for both graphs, but the CMU baseline does better for the 11 concept graph. The CMLF system has MAP of 0.6009 and 0.5156 for the 5 and 11 concept graphs, respectively. The

Table 1. The table shows per-concept average precision results using the CML and CML+I models on the 5 concept graph.

| Per concept results 5-concept graph | | | |
|-------------------------------------|----------|----------|----------|
| | SVM | CML | CML+I |
| CAR | 0.252123 | 0.229518 | 0.260092 |
| WALKING/RUNNING | 0.352651 | 0.34861 | 0.355762 |
| PEOPLE | 0.830886 | 0.927372 | 0.935306 |
| TEXT | 0.669119 | 0.650824 | 0.670135 |
| FACE | 0.894924 | 0.888202 | 0.89243 |
| MAP | 0.599941 | 0.608905 | 0.622765 |

CMLF+I system uses the more complete interaction potential functions of (8). This system outperforms CMLF, but in both cases is worse than CML+I. The CMLF+I system has MAP of 0.6122 and 0.5184 for the 5 and 11 concept graphs, respectively. We believe that this is due to the coarse quantization and relatively uninformative low-level features used (k-means to 126 classes), and expect that more sophisticated quantization will yield further improvements.

Table 1 shows the average precision results for each concept in the 5 concept graph. “People” shows the biggest increase in both the CML and CML+I cases, we speculate, due to joint inference with the “face” concept, and possibly “walking/running” as well. The reverse effect is not evident, as “people” can be observed without viewing a “face” (or “walking/running”). This fact could cause the slightly worse performance for “face” and “walking/running” in the CML case, relative to the independent SVMs. However for each concept, CML+I performs essentially as well or better than the SVMs. We feel these limited results demonstrate the potential performance gains associated with collective annotation using full pairwise concept interaction models and a properly constructed graph.

5. Related work

There is substantial amount of related work, the most closely related in both technique and application domain is [17]. As noted above, our random field model has several components differing from that system that we feel enhance the model. However, these components are inspired by several other related systems. The basic CML and CMLF models were proposed for multi-category text document classification in [2]. The main difference here is the use of independently trained discriminative classifiers for the association terms. Also, they did not use quantization to build the features analogous to (13). The use of quantization to extend maximum entropy text categorization to image categorization was proposed by [4]. However, this

work concerned independent per-category rather than collective categorization. Again, we mention the original work on the DRF model in [6, 5], which integrated discriminative local classifiers for collective classification. However, this work was targeted at classification of multiple regions within a single still image, and used spatially-based interaction terms to define the graphical model. It also did not include concept-feature interaction. Finally, the original work on conditional random fields in [7] lays much of the theoretical groundwork for the various extensions in [6, 17, 2] and the work herein.

6. Conclusion

In this paper, we have presented early work on further extending conditional random field models for generic media annotation and semantic concept detection. The goal of these models is to incorporate contextual information by modeling complex pairwise concept interaction, and pairwise concept and feature interactions. Our initial experiments have shown moderately positive results, but the results using the CMLF+I model suggest that we need to further explore the use of quantization to determine its impact on performance.

Significant future research remains to be performed. The biggest single unresolved issue is graph induction. Performance will depend critically on a graph that includes important inter-concept dependencies. At the same time, the number of edges in the graph directly impacts computational complexity, indicating an important tradeoff. Including extraneous edges can also be expected to degrade performance. Additionally, scalability of the approach will need to be a major focus. In particular, our current reliance on exact inference can’t scale to hundreds or thousands of concepts, even if their co-occurrence is highly sparse. There is a burgeoning literature on approximate inference techniques for graphical models. [2] suggests two approaches: binary pruned inference and supported inference, which are

based on label co-occurrence in the training set. Belief propagation is also becoming a fairly standard approximate inference technique [18]. [5] also includes extensive experiments with approximate inference techniques for DRFs which are also applicable to the models described here.

References

- [1] M. W. Berry, S. T. Dumais, and G. W. O'Brien. Using linear algebra for intelligent information retrieval. *SIAM Rev.*, 37(4):573–595, 1995.
- [2] N. Ghamrawi and A. McCallum. Collective multi-label classification. In *CIKM '05: Proceedings of the 14th ACM international conference on Information and knowledge management*, pages 195–200, New York, NY, USA, 2005. ACM Press.
- [3] L. Hollink and M. Worring. Building a visual ontology for video retrieval. In *MULTIMEDIA '05: Proceedings of the 13th annual ACM international conference on Multimedia*, pages 479–482, New York, NY, USA, 2005. ACM Press.
- [4] J. Jeon, V. Lavrenko, and R. Manmatha. Automatic image annotation and retrieval using cross-media relevance models. In *SIGIR '03: Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, pages 119–126, New York, NY, USA, 2003. ACM Press.
- [5] S. Kumar. *Models for learning spatial interactions in natural images for context-based classification*. PhD thesis, Carnegie Mellon Univeristy, Pittsburgh, PA, USA, 2005. Chair-Martial Hebert.
- [6] S. Kumar and M. Hebert. Discriminative random fields. *Int. J. Comput. Vision*, 68(2):179–201, 2006.
- [7] J. D. Lafferty, A. McCallum, and F. C. N. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *ICML '01: Proceedings of the Eighteenth International Conference on Machine Learning*, pages 282–289, San Francisco, CA, USA, 2001. Morgan Kaufmann Publishers Inc.
- [8] R. Malouf. A comparison of algorithms for maximum entropy parameter estimation. In *COLING-02: proceeding of the 6th conference on Natural language learning*, pages 1–7, Morristown, NJ, USA, 2002. Association for Computational Linguistics.
- [9] M. Naphade, J. R. Smith, J. Tesic, S.-F. Chang, W. Hsu, L. Kennedy, A. Hauptmann, and J. Curtis. Large-scale concept ontology for multimedia. *IEEE Multimedia Magazine*, 13(3), 2006.
- [10] P. Over, T. Ianeva, W. Kraaij, and A. Smeaton. Trecvid 2006 an overview. In *Proceedings of the TRECVID 2006 Workshop*, Nov. 2006.
- [11] J. Platt. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In *Advances in Large Margin Classifiers*. MIT Press, 1999.
- [12] S. E. Robertson and K. Sparck-Jones. Simple, proven approaches to text retrieval. Technical Report TR356, University of Cambridge, Computer Laboratory, 1997.
- [13] C. G. Snoek, M. Worring, D. C. Koelma, and A. W. Smeulders. A learned lexicon-driven paradigm for interactive video retrieval. *IEEE Transactions on Multimedia*, 9(2):280–292, 2007.
- [14] C. G. M. Snoek, M. Worring, and A. W. M. Smeulders. Early versus late fusion in semantic video analysis. In *MULTIMEDIA '05: Proceedings of the 13th annual ACM international conference on Multimedia*, pages 399–402, New York, NY, USA, 2005. ACM Press.
- [15] C. G. M. Snoek, M. Worring, J. C. van Gemert, J.-M. Geusebroek, and A. W. M. Smeulders. The challenge problem for automated detection of 101 semantic concepts in multimedia. In *MULTIMEDIA '06: Proceedings of the 14th annual ACM international conference on Multimedia*, pages 421–430, New York, NY, USA, 2006. ACM Press.
- [16] J. C. van Gemert, J.-M. Geusebroek, C. J. Veenman, C. G. M. Snoek, and A. W. M. Smeulders. Robust scene categorization by learning image statistics in context. In *CVPRW '06: Proceedings of the 2006 Conference on Computer Vision and Pattern Recognition Workshop*, page 105, Washington, DC, USA, 2006. IEEE Computer Society.
- [17] R. Yan, M.-Y. Chen, and A. Hauptmann. Mining relationships between concepts using probabalistic graphical models. In *Proc. IEEE ICME*, 2006.
- [18] J. S. Yedidia, W. T. Freeman, and Y. Weiss. Constructing free-energy approximations and generalized belief propagation algorithms. *IEEE Transactions on Information Theory*, 51(7):2282–2312, 2005.