A Discrete Direct Retrieval Model for Image and Video Retrieval

Shaolei Feng *
Integrated Data Systems Department
Siemens Corporate Research, Inc.
Princeton, NJ, 08540
slfeng@cs.umass.edu

R. Manmatha
Multimedia Indexing and Retrieval Group
Center for Intelligent Information Retrieval
Department of Computer Science
University of Massachusetts
Amherst, MA, 01003
manmatha@cs.umass.edu

ABSTRACT

This paper proposes a formal framework for image and video retrieval using discrete Markov random fields (MRF). The training dataset consists of images with keywords (regions are not labeled). The model is built using a discrete vocabulary of vector quantized region or point features generated from the training images. Since performance is dependent on the size of the vocabulary, a large vocabulary of a couple of million visterms is used. Such large vocabularies cannot be generated by conventional clustering algorithms so hierarchical k-means is used to generate it. Unlike many previous techniques, our MRF based model doesn't require an explicit annotation step for retrieval. The model directly ranks all test images according to the posterior probability of an image given a query. Traditionally, most models are trained by maximizing likelihood - instead this model is trained by maximizing average precision. Image and video retrieval experiments are performed on two standard datasets (a Corel dataset and a TRECVID3 dataset) which consist of 4,500 images and about 44,100 keyframes respectively. The results show that based on a large visual vocabulary the model runs extremely fast on even very large datasets while having comparable retrieval performance to the best performing (continuous feature) models.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Retrieval Models; I.5.1 [**Models**]: Statistical; I.4.9 [**Image Processing and Computer Vision**]: Applications

General Terms

Algorithms, Experimentation, Performance

*This work was done while the author was at the University of Massachusetts.

CIVR'08, July 7–9, 2008, Niagara Falls, Ontario, Canada.

Keywords

Image Annotation, Image Retrieval, Video Retrieval, Markov Random Field, Discrete Models, Large Visual Vocabulary

1. INTRODUCTION

In this paper we propose a discrete Markov Random Field (MRF) model for directly retrieving un-annotated images using text queries. The model is learned using a training set of images labeled with annotations. Discrete implies that image features are mapped to a discrete set of visterms (or visual words). The advantage of discrete visterms (as opposed to continuous features) is that both training and testing are very fast - by several orders of magnitude. While previous annotation based techniques have used discrete visterms [12, 13] they have performed much more poorly than versions based on continuous features [15, 8, 6]. It is shown in this paper that by using a very large visual vocabulary one can improve the retrieval precision to a level close to that of the best continuous models while at the same time creating a very fast algorithm. Our vocabularies have a couple of million words. Traditional clustering algorithms cannot create such large vocabularies. Recently, however, approaches have been proposed for creating large vocabularies in the image matching literature [22, 23]. In this paper hierarchical k-means [22] is used to create a large vocabulary. Indexing and querying are both speeded up using this algorithm. It is also shown that performance is correlated with the size of vocabulary. This demonstrates the need for large training sets.

A larger vocabulary requires using both larger numbers of training images and also using a larger number of features per image. Typically, an optimal vocabulary size is some fraction (say around 1/3) of the total number of features in the training data. Using a large number of features per image ensures that a model can learn relationships between features rather than relying on an explicit segmentation which is hard to do. However, one cannot use every pixel in the image since there is often a minimum size required to required to reliably compute the features. Using more features per image also constrains the models that can be used. For example, relevance models [12, 15, 8] fail when the number of features per image or alternatively the number of visterms per image is large [1]. These models compute a joint proba-

[1]This problem also occurs with relevance models in text

bility over all visterms (or features) in an image. When the number of these visterms (or features) is large, this product cannot be estimated well. The Markov random field model proposed here does not suffer from this limitation.

The model proposed here directly retrieves images without an intermediate annotation step. Specifically, we do this by directly optimizing mean average precision unlike annotation techniques in the literature [7, 12, 2] which optimize annotation accuracy by optimizing likelihood. Annotation techniques do not optimize retrieval. Annotation accuracy is usually measured in terms of how well the annotation words are predicted for an image. Retrieval accuracy maximizes mean average precision for a query word. An example will make this clearer. Assume we have a database of images where grass or water is very frequent while pictures depicting animals such as deer and bears are less frequent. Good annotation performance requires that many pictures will be annotated with grass and water. Labeling every image with grass and water may in fact improve annotation performance. However, obtaining good retrieval performance requires that one be able to retrieve images corresponding to a number of different queries including "rare word queries". In fact one can do very well at retrieval even if the grass and water queries fail completely. In this respect, good retrieval performance requires that the model do well with many different kinds of "word" queries while annotation requires one to do well at finding as many keywords as possible for images. This issue has also been observed in text retrieval where it has been shown that the likelihood surface is unlikely to correlate with the retrieval metric surface [21, 19].

We propose a Markov random field (MRF) model for direct retrieval which is analogous to one proposed by Metzler and Croft [19] to capture dependencies in text retrieval. MRF models have been used before in image processing and computer vision for a wide variety of tasks ranging from the low-level (edge detection and image segmentation) to the high-level (matching) [25, 4]. The MRF model proposed here is quite different from them in terms of the problems tackled, topological structure and training strategy. Most MRF models in computer vision are used to label image sites with word labels. We propose a model to directly rank images in response to a text query. The goal of our proposed MRF model is to compute the joint probabilities of images and queries. The graph consists of a set of region nodes representing an image and a set of word nodes representing a query, with edges determining the dependency among these nodes. Unlike labeling problems, our proposed model neither requires that every training image is labeled region by region nor outputs annotations at the region level. Instead, it calculates the joint probability of a query word with the entire image in order to rank images. Finally, we maximize mean average precision rather than the likelihood to optimize the retrieval performance. This leads to a linear form for the model simplifying computation and allowing us to handle large numbers of visterms per image.

Experiments demonstrate that the model performs very well with a mean average precision (MAP) of 0.28 on a standard 5K Corel set. For comparison the best published retrieval results we know of on the same data set are 0.31 (SML [6]) and 0.30 (NCRM [14]). On a TRECVID3 dataset [11] the corresponding numbers are 0.152 and 0.158 for the discrete MRF model and the NCRM model respectively. The discrete MRF takes 90s for all queries while NCRM takes 6.8 hrs. Clearly, the discrete MRF model has comparable precision with the best models while being very fast. Further improvements may be possible to push up the precision even further.

where they do not work well for long queries

1.1 Related Work

There is a fair amount of literature on image annotation models in the past few years. Barnard et al. [2, 1] discuss a number of models for image annotation and labeling including machine translation, probabilistic latent semantic indexing and latent Dirichlet allocation. Models were based on both discrete visterms and continuous ones which modeled features as mixtures of Gaussians. Carbonetto et al. [5] proposed a shrinkage model which essentially allows for continuous features in a translation model unlike the discrete translation model used in [1]. Blei and Jordan [3] proposed several hierarchical probabilistic models based on latent Dirichlet allocation, which assumes a low-dimensional topology with about 200 "latent aspects". Their model again used a mixture of Gaussians to generate the features. Carneiro et al. [6] used multiple instance learning and hierarchical Gaussian mixture models for annotation and retrieval and claimed that their model worked best closely followed by the MBRM in [8]. Jeon et al. [12] proposed a cross-media relevance model which viewed the image annotation problem as analogous to cross-lingual information retrieval. The model used discrete visterms. Other discrete models include one using maximum entropy [13] and inference nets [20]. A number of models have been proposed which use hidden Markov models [26, 16, 9]. The model in [9] generates the features using a mixture of Gaussians. Magahães and Rüger [18] proposed using logistic regression to learn image semantics from generic codebooks and key words. Shi et al. [24] proposed a Bayesian hierarchical multinomial mixture model for image annotation, which utilized the prior knowledge of concept dependencies. They again modeled features using a mixture of Gaussians.

Our direct retrieval model based on Markov random field (MRF) is analogous to the Markov random field framework proposed by Metzler and Croft [19] for text retrieval, which explored different query term dependencies when retrieving text documents. Compared with annotation based retrieval models, our MRF model doesn't involve an explicit annotation step and is trained through directly maximizing mean average precision.

Little previous work has been done on direct image retrieval based on text queries. Jeon *et al.* [12] directly ranked images according to the K-L divergence of visterm distributions of the query model and the document model. Their approach assumed that the query and a relevant image have similar visterm distributions. However, in the real world these distributions may be very different. Inspired by the cross-language latent semantic indexing (LSI) in text retrieval, Hare *et al.* [10] proposed a singular value decomposition (SVD) based approach to learn the semantic structure of the visterms and annotation words from the training set and retrieve images according to the positions of the text images in the sematic space. Like the LSI techniques in text retrieval, their approach assumed latent concepts linking the visual terms and annotation words.

Our direct retrieval framework doesn't make such assump-

tions. It estimates an underlying joint distribution of queries and images $P_\Lambda(Q, I)$ through modeling the dependency of annotation words and each image region among a test image, and formulates this as a Markov random field with a set of parameters Λ. Images are ranked according to the posterior $P_\Lambda(I|Q)$. Given a set of query and image pairs, the parameters Λ are estimated by directly maximizing the mean average precision rather than the likelihood of the training data.

Our discrete MRF is based on a very large visual vocabulary. Recent literature on object or image matching [22, 23] using discrete features has shown that [2] the size of visual vocabulary can substantially affect the matching performance and good performance requires large visual vocabularies. Nister and Stewenius [22] constructed a vocabulary tree to index SIFT descriptors of images. Images of the same object from a different viewpoint or under different lighting conditions are retrieved based on *tf-idf* measures of the visual words. Philbin *et al.* [23] presented fast image matching approaches using large visual vocabularies constructed by an approximate k-means clustering method over SIFT descriptors.

2. MARKOV RANDOM FIELD FOR IMAGE RETRIEVAL

Markov random fields (MRFs) have been widely used to model the joint distribution of a set of random variables. In the computer vision domain MRFs have been applied to image restoration, edge detection, texture analysis, image segmentation and image matching [17]. In this section we describe a Markov random field based model for text query based image and video retrieval. Our MRF models the joint distribution $P_\Lambda(Q, I)$ over text queries Q and images I, parameterized by Λ. Based on the joint distribution, images are ranked according to the posterior probability of $P_\Lambda(I|Q)$ without an explicit annotation step.

2.1 Framework Overview

A Markov Random Field is an undirected graph where the nodes represent random variables and the edges encodes the conditional independencies among those random variables. Given its clique set (a neighborhood set), a random variable in a MRF is conditionally independent of all other random variables (this is the Markov property). We propose a model similar to one proposed by Metzler and Croft [19] for text retrieval. In our MRF G for image retrieval based on text queries, the random variables are the key words $\{q_i\}$ in a query Q and the image I which is represented by a set of visterms $\{v_i\}$. The joint distribution $P(Q, I)$ of a query Q and image I is given by:

$$P(Q, I) = \frac{1}{Z_\Lambda} \prod_{c \in C(G)} \phi(c; \Lambda) \tag{1}$$

where $Q = q_1 \dots q_n$, the image I is represented by a set of visterms $\{v_i\}$, and $\phi_{c;\Lambda}$ is a set of non-negative potential functions parameterized by Λ, one for each clique c in graph G. The normalizing constant Z_Λ is:

$$Z_\Lambda = \sum_{Q,I} \prod_{c \in C(G)} \phi(c; \Lambda) \tag{2}$$

[2]Jeon and Manmatha also noticed this in unpublished work

This quantity is expensive to compute due to the exponential number of summations. We shall show later that in the case where we are only interested in the ranking and not actual probability values, we do not have to compute it.

For textual query based image retrieval, we rank images according to the posteriors $P(I|Q)$ of each image I given the query Q:

$$\begin{aligned} P(I|Q) &= \frac{P_\Lambda(Q, I)}{P_\Lambda(Q)} \\ &\stackrel{rank}{=} \log P_\Lambda(Q, I) - \log P_\Lambda(Q) \\ &\stackrel{rank}{=} \sum_{c \in C(G)} \log \phi(c; \Lambda) \end{aligned} \tag{3}$$

where $\stackrel{rank}{=}$ implies rank equivalence. Note that the normalizing factor has been eliminated and the result is a linear sum over the logarithm of the potential functions.

The potential function is non-negative and often assumed to be an exponential:

$$\phi(c; \Lambda) = \exp[\lambda_c f(c)] \tag{4}$$

where $f(c)$ is some feature function over clique c, and λ_c is the weight of this particular feature function. The ranking function may, therefore, be re-written as:

$$P_\Lambda(I|Q) \stackrel{rank}{=} \sum_{c \in C(G)} \lambda_c f(c) \tag{5}$$

This is linear in the feature functions. Later we show that one feature function is computed for each visterm-word combination. This means that there is no term involving a large number of products. Hence, this probability can be computed for a large number of visterms or features in each image.

λ_c may depend on the kind of feature function. For example, cliques containing one query term and one visterm may be weighted differently than cliques containing one query term and two visterms. In this paper we will only consider cliques containing a query term and a visterm. Further, λ_c will be assumed the same for all cliques containing a particular visterm. Specifically, λ_c will be set equal to the inverse document frequency (idf) of the visterm.

3. VISTERM GENERATION

In this paper, an image I is represented as a set of visterms $\{v_1, v_2, \dots, v_m\}$. Each visterm v_i corresponds to a region r_i either obtained through superimposing a rectangular grid or automatic segmentation. Real-valued visual features (color, texture) are first extracted from each region. Then a visual vocabulary is built over those features and each region is quantized into a discrete visterm (visual word). Note that instead of regions we may use local interest descriptors or augment regions with local interest descriptors. The model proposed in this paper has no specific requirements on image representation.

Since a discrete Markov random field is built on discrete image features, the first step is to quantize image features into discrete visterms. Unsupervised clustering methods are usually employed for this purpose, e.g. K-means clustering or hierarchical clustering. Most clustering methods require that one pre-defines either the number of clusters or some threshold controlling the number of clusters. Each cluster is

a visterm, so the size of the vocabulary is equal to the number of clusters. Using discrete features for image matching [22, 23] has shown that the size of the visual vocabulary can substantially affect the matching performance and good performance requires large visual vocabularies. This is reasonable since large visual vocabularies are better at distinguishing different visual features. However, too large a visual vocabulary can also segregate features originating from the same objects. So selecting the appropriate size of the visual vocabulary is very important, but usually very difficult without any domain knowledge. To test on different size of visual vocabularies, one requires a fast clustering approach which can deal with large-scale features in high dimension space. A flat K-means clustering or a hierarchical agglomerative clustering (also called single linkage clustering) doesn't meet this requirement. For this reason, we adopt the hierarchical k-means for clustering in our work [22, 23].

Hierarchical k-means applies a tree structure for representation of the clustering results over a set of training features, where k defines the branch factor of the tree rather than the total number of the categories. Initially, the k-means algorithm partitions the training features into k clusters, each of which forms a node in the tree consisting of feature vectors closest to a particular cluster center. The same k-means algorithm is then recursively applied to each node and splits each of them into k finer clusters. This process is recursively performed until the depth of the tree reaches a pre-defined level. So if the depth of the tree is d, the number of clusters at the leaf level will be k^d. The computational cost of the hierarchical k-means is logarithmic in the number of leaf nodes, which is much smaller than that of non-hierarchical clustering methods.

The visual vocabulary tree is constructed by clustering all the feature vectors in the training set using the hierarchical k-means. Then the feature vectors of the test set will be clustered through an efficient search procedure, which propagates the vector down the tree till the leaf level by comparing the vector with the k candidates cluster centers at each level and selecting the closest one. This lookup only takes $\mathbf{O}(log(n))$ compared with the complexity $\mathbf{O}(n)$ of a flat K-means for the same task, where n is the size of the visual vocabulary. In the case of an extremely large training set, the visual vocabulary tree may be constructed using a portion of the training feature vectors sampled from the whole training set. Then the corresponding visual words of the rest of training vectors are obtained through searching over the tree as for the test vectors. So finally, an image is represented as a set of visual words (visterms) each of which corresponds to one image region, noted as $\{v_1, \ldots, v_m\}$, where m is the number of the regions.

4. DISCRETE MARKOV RANDOM FIELDS

The MRF can model various dependencies among the variables involved. This paper assumes that all image regions are independent of each other given some query Q. This assumption is also made by many annotation or retrieval models for images and videos, e.g. relevance models and machine translation models. Under this assumption, the likelihood of one image region is independent of others given the query. It is straightforward to generalize the MRF to model higher order dependencies but the computational load becomes higher. Preliminary experiments show that estimating bigrams for a large vocabulary is difficult (for a vocabulary of size n there are potentially $O(n^2)$ bigrams - in our case this is potentially $O(10^{12})$ bigrams). Bigram estimation may require much larger training sets or more powerful estimation techniques. So higher order dependencies are not explored in this paper. One can also easily model other kinds of features such as point features. Figure 1 illustrates the configurations of our MRF.

4.1 Clique Potentials

Equation 5 shows that using MRF for image retrieval may be reduced to a linear combination problem. Under this framework we do not need to calculate the exact joint probability $P(Q, I)$. Instead, by choosing proper potential functions we try to approximate the joint distribution in a generalized exponential form. For example, the potential function should give a higher value for a clique including the query word "zebra" and an image region with white and black strips than a clique of the same query word and a plain red image region. The proposed MRF explicitly models the context information since each query word node is connected to all image regions or region pairs. As an example, given an image of a zebra in shrubbery and the query word "zebra", each visual word (no matter from a zebra region or shrubbery region) in this image will contribute to the energy function with a non-negative value depending on how compatible that region is with the query word. This is quite different from standard MRF based annotation or recognition methods which make hard decisions for each region by labelling it with one word. For any clique which doesn't involve an image region node, the potential is assumed to be 1 since it does not have an impact on ranking. In our model the simplest clique is a 2-clique consisting of a query node w and an image region node r.

The potential function of the MRF model is defined over an image region represented as a discrete visterm v and one query word w. Since equation 4 expresses potential functions in terms of sums of weighted feature functions $\lambda_c f(c)$, all we need to do is to estimate $f(c)$ and λ_c. Formally,

$$f(c) = f(w, v) = P(w|v)P(v|I) \tag{6}$$

It is not desirable to have a separate weight λ_c for each word and visterm combination since the estimation would require estimating *size of word vocabulary* $\times$ *size of visterm vocabulary* weights. In text retrieval, the inverse document frequency plays an important part in deciding how important a word is for indexing purposes. Very frequent words like "the" have no indexing value. Rare words are very useful for retrieval. This analogy and intuition may be used to guide the choice of the weight. Common visterms should have smaller weight while rarer visterms are a better indication of the relevance of the document. Thus, for any clique containing a visterm v and a particular word w, λ_c is directly calculated as the *idf (inverse document frequency)* of the visterm. That is, all cliques containing the visterm v have the same weight. To show this express dependence we can write $\lambda_c = \lambda_v = \log \frac{|D|}{\sharp(d_j \ v \in d_j)}$ which measures the general importance of the discrete visterm v. Of course we could have made other choices for λ_c. $P(v|I)$ is the probability of a visual word v observed in the test image I and $P(w|v)$ is the posterior probability of a query word w given a visual word v. So here the potential function basically represents the possibility of predicting query word w from

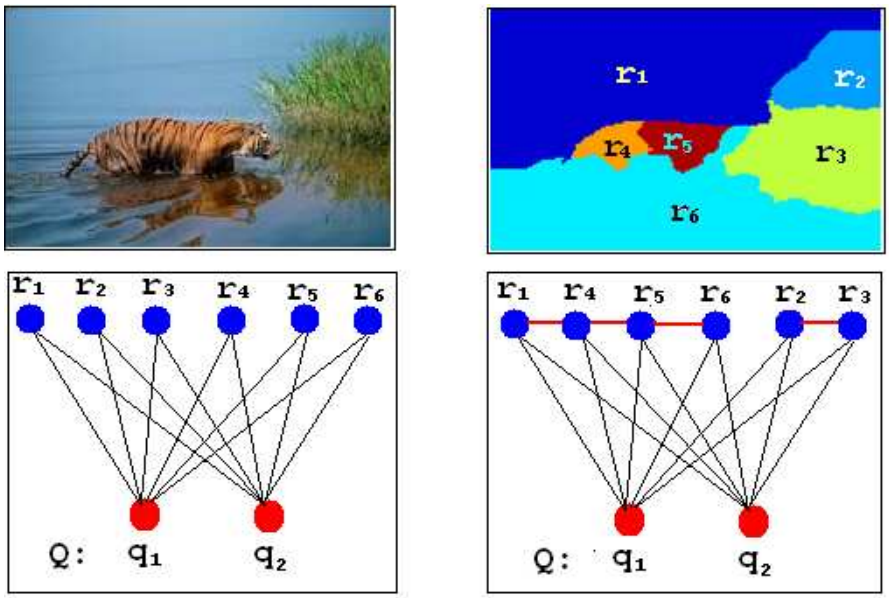

Figure 1: The configuration of MRF models for image retrieval. The top line shows the original image and its regional representation. The bottom line (left) shows the MRF model used here where all image regions are independent of each other. This figure shows segmentation-based partitions. In experiments we use a rectangular grid to generate a large number of regions per image based on which we can build large visual vocabularies. The version on the right shows a bigram version of the MRF model where edges in red are determined by nearest neighboring region pairs. This second model is not explored in this paper due to estimation difficulties

the occurrences of visual word v in the test image I.

Estimating the probabilities $P(w|v)$ and $P(v|I)$ depends on the distributions of the words and the visterms. Previous work [8, 14] has demonstrated that the normalized multinomial or multiple Bernoulli model is more suitable than a multinomial for annotation word distribution. So here, we utilize a normalized multinomial distribution for word probability estimation - this is essentially a multinomial distribution over the words after the annotations for each training set image have been padded to a constant length by using null words if necessary. Without any prior knowledge of the discrete visterm distribution, we investigate both multinomial and multiple Bernoulli models for $P(v|I)$ (not that it is not obvious that the distribution for words and visterms should be the same).

4.2 Probability Estimation

The posterior probability $P(w|v)$ is calculated under the Bayesian framework:

$$P(w|v) = \frac{P(w,v)}{P(v)} = \frac{P(w,v)}{\sum_w P(w,v)} \tag{7}$$

The joint probability of $P(w,v)$ is calculated through an expectation over all training images (alternatively, one may view this as a mixture over the training set of images):

$$P(w,v) = \sum_{J \in \tau} P(J)P(w|J)P(v|J) \tag{8}$$

where τ is the training set and J is an image in the training set. The word probability $P(w|J)$ is estimated based on the relative frequency of the word w in the annotation of image J which has been padded to a fixed length with a special "null" word. Following the word probability estimation in [14], $P(w|J)$ is estimated using a normalized multinomial distribution:

$$P(w|J) = \lambda \frac{N_{w,J}}{N_J} + (1-\lambda)\frac{N_w}{N} \tag{9}$$

where N_J is the fixed length of annotations of training images, $N_{w,J}$ the number of occurrence of word w in image J, N_w the number of w in the whole training set and N the total number of annotation words in the training set. λ is a tuned weight to smooth the word probability given an image using the prior probability estimated from the whole training set. The value of λ is tuned over a validation set.

We investigated two different distributions – multinomial distribution and multiple Bernoulli distribution – to estimate the probability of a visterm v given a training image J or a test image I.

4.2.1 Multinomial Visterm Model

Based on a multinomial distribution assumption of the visterms from images, the probability of $P(v|I)$ is calculated as the frequency of the visterm v in image I:

$$P(v|I) = \frac{\sharp(v,I)}{\sum_v \sharp(v,I)} \tag{10}$$

where $\sharp(v,I)$ is the number of occurrences of the visterm v in image I.

With a multinomial distribution the visterm probability $P(v|J)$ of a visterm v generated by a training image J is estimated similarly.

4.2.2 Multiple Bernoulli Visterm Model

A multiple Bernoulli visterm model only considers if a particular discrete visterm occurs in the image or not and ignores the number of occurrences of that visterm if it does exist in the image. Correspondingly, the probability of $P(v|I)$ is estimated as a discrete Kronecker delta function:

$$P(v|I) = \delta_{v,I} \tag{11}$$

where $\delta_{v,I} = 1$ if the word v occurs in the annotation of image I and zero otherwise.The Bernoulli visterm model basically emphasizes the presence or absence of a particular visterm rather than its frequency in an image, which implies that a repeating visterm in an image may not be more important than a unique one. Taking an image of a zebra and trees as an example, the visterms from the tree regions may have higher frequencies than those from the zebra region in that image. But it may not be reasonable for a visterm of tree to contribute more to the probabilities of both word "tree" and word "zebra" than a visterm of zebra, without knowing which visterm corresponds to which word.

As for the multinomial case, the probability $P(w, v)$ is calculated using Equation 8 but $P(v|J)$ is computed based on the multiple Bernoulli distribution: $P(v|J) = \delta_{v,J}$.

5. EXPERIMENTAL RESULTS

We use two different datasets in our experiments for comparison between our model and other models. The first one is a standard Corel image set which contains 5000 images widely used for comparing results. The second one is the large scale data set consisting of the entire TRECVID 2003 development dataset and feature set used by [11].

The 5K Corel dataset has been used in many papers [7, 15, 8, 6] for image annotation and retrieval. This dataset consists of 5000 images from 50 Corel Stock Photo CD's. [3] Each CD includes 100 images on the same topic, and each image is also associated with 1-5 keywords. Overall there are 371 keywords in the dataset. In experiments, we divided this dataset into 3 parts: a training set of 4000 images, a validation set of 500 images and a test set of 500 images. The validation set is used to find model parameters. After finding the parameters, we merged the 4000 training set and 500 validation set to form a new training set. This corresponds to the training set of 4500 images and the test set of 500 images used by Duygulu *et al.* [7].

Other tests were run on NIST's entire TRECVID3 development dataset containing 58 mpegs of ABC World News Tonight and 57 mpegs files of CNN Headline News as in [11]. The set is divided into 45 hrs of training data and 15 hrs of test data. The mpeg files are segmented into video shots, each of which is represented by a key frame. So by retrieving key frames, one can retrieve corresponding video shots. In total there are about 44100 key frames. Each key frame in the TRECVID3 development dataset has been manually annotated with key words from about 100 semantic concepts, from which about 75 concepts are selected in our experiments to guarantee that each of them has more than 20 training examples in the development set. In the final test of our algorithm, the training set and the test set are also separated in time, with 34,880 key frames for training and 9,220 key frames for test.

[3]We thank Kobus Barnard for making the Corel dataset available at
http://www.cs.arizona.edu/people/kobus/research/data/eccv_2002

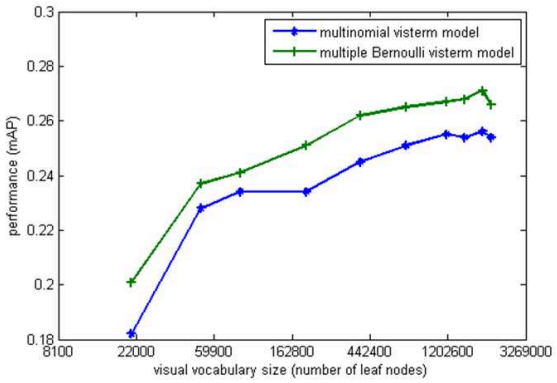

Figure 2: Curves of performance vs visual vocabulary size for multinomial visterm model and multiple Bernoulli visterm model.

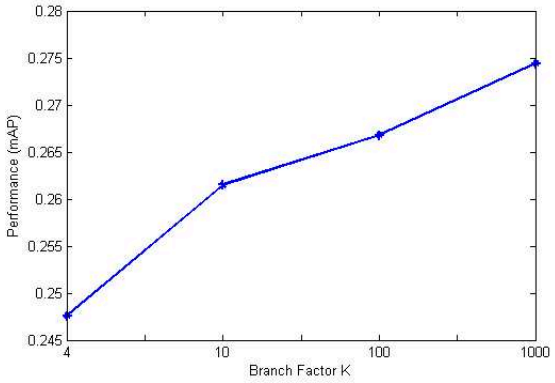

Figure 3: Performance vs branch factors with 1M leaf nodes

For both the 5K Corel set and the TRECVID dataset, every image is partitioned using a rectangular grid and a feature vector is then calculated for every grid region. The features used include 18 color features and 12 texture features. The color features for an image region include the average, the standard deviation and the skewness of the pixel values for each channel of the RGB color space and L^*a^*b space. The texture features consist of Gabor energy computed over 3 scales and 4 orientations.

5.1 Retrieval Results

In our experiments on discrete MRFs, images are first partitioned into rectangular regions and the discrete visterms are built from those regions.

On the 5k Corel image set, we tested the effects of the size of the visual vocabulary on the retrieval performance and compared the multinomial visterm model and the multiple Bernoulli visterm model. Figure 2 shows that the mean average precision dramatically increases with vocabulary size and then flattens out. We can also see that the multiple Bernoulli visterm model works better than the multinomial visterm model by a small margin.

We also observed that with the same visual vocabulary

	CMRM	CRM	N-CRM	SML [6]	Discrete MRF
mAP	0.14	0.26	0.30	0.31	0.28
RunningTime(secs)	10	660	660	-	16

Table 1: Retrieval performance comparison between discrete MRF and other models on the 5k Corel set. The running time is measured for all the 371 words in the vocabulary.

	N-CRM	Discrete MRF
mAP	0.158	0.152
P@10	0.319	0.335
RunningTime	6.8(hrs)	90(secs)

Table 2: Retrieval performance comparison between discrete MRF and N-CRM on the TRECVID03 set. The running time is measured for all the 75 query words.

size, the performance increases with the branching factor (see Figure 3). We believe this is because of the property of the hierarchical k-means clustering. Since the visual vocabulary is constructed from a hierarchical k-means tree, then any errors in clustering made at the higher level will be propagated to the lower level and cannot be corrected. A tree with a larger branch factor reduces the chances of propagation of the clustering error. However, there is a tradeoff since larger branching factors will slow down the algorithm.

In our experiments we also tested the discrete visterm sets constructed from the multiple-scale space of the original image via the image Gaussian pyramid. However, the results did not show any apparent improvement over the single scale setting.

The best results for the discrete MRF is obtained using the multiple Bernoulli visterm model with a 2085136 (38^4) vocabulary size over 16x16 overlapping rectangular partitions (resulting in 308 rectangular regions per image), which achieves a mean average precision (mAP) 0.28 and precision@10 of 0.198. Table 1 shows the comparison with other models, from which we can see that our discrete MRF achieves better results than the cross-media relevance models(CMRM) and the continuous relevance model(CRM) and is comparable to normalized-continuous relevance models and Carneiro's hierarchical Gaussian mixture model [6]. However our discrete MRF is much more efficient than continuous models in terms of running time. The implementation used sparse matrix techniques to accelerate the probability calculation.

Finally, our discrete MRF model was tested on the TRECVID03 dataset. Each keyframe in this dataset is partitioned into 32x32 overlapping rectangular regions, which results in 300 partitions per keyframe. And then color and texture features are extracted from each region. From the training features we construct a visual vocabulary of size 2085136 (38^4). Then all test features are determined by look up using the tree to obtained their corresponding discrete visterms. Our results are shown in table 2, from which we can see that the mean average precision of our discrete models on this dataset is very close to the NCRM and the precision@10 is slightly better. Compared to the 6.8 hours of running time measured for the NCRM, our discrete model only takes 1.5 minutes to complete the whole procedure once the discrete visterms are computed.

Figure 4 shows the 5 top ranked images in the returned rank list of the discrete MRF in response to the query word "birds" over the 5k Corel set. Note the third image does contains a bird (seagull) although it is quite small and the ground-truth annotation of that image does have the word "birds". Although the ground-truth annotation of the first image doesn't contain the word "birds" (instead it has "albatross"), our model correctly associates word "birds" with it. Figure 5 shows a retrieval example for the discrete MRF in response to the query word "sport_event" over the TRECVID03 dataset.

6. CONCLUSIONS

This paper proposes a discrete MRF based models for image and video retrieval. Unlike other automatic annotation based retrieval models, our MRF models directly retrieve images without involving an explicit annotation step and are directly trained through maximizing the retrieval performance - mean average precision. In our discrete MRF model, large visual vocabularies are obtained by using hierarchical K-means to cluster image features. We demonstrated that the discrete MRF model runs much faster while having comparable retrieval performance with the continuous models. Our future work will investigate incorporating different features and the use of feature dependencies.

7. ACKNOWLEDGMENTS

This work was supported in part by the Center for Intelligent Information Retrieval and in part by grant #NSF CNS-0619337. Any opinions, findings and conclusions or recommendations expressed in this material are the author(s) and do not necessarily reflect those of the sponsor.

8. REFERENCES

[1] K. Barnard, P. Duygulu, N. de Freitas, D. Forsyth, D. Blei, , and M. I. Jordan. Matching words and pictures. *Journal of Machine Learning Research*, pages 1107–1135, 2003.

[2] K. Barnard and D. Forsyth. Learning the semantics of words and pictures. In *Proc. ICCV*, volume 2, pages 408–415, 2001.

[3] D. M. Blei and M. I. Jordan. Modeling annotated data. In *Proc. of the 26th Annual Int'l ACM SIGIR Conf.*, pages 127–134, Toronto, Canada, July 28-August 1 2003.

[4] Carbonetto, N. de Freitas, and K. Barnard. A statistical model for general contextual object recognition. In *Proc. ECCV*, 2004.

[5] Carbonetto, N. de Freitas, P. Gustafson, and N. Thompson. Bayesian feature weighting for unsupervised learning, with application to object recognition. In *Proceedings of the 9th International Workshop on Artificial Intelligence and Statistics*, 2003.

Figure 4: 5 top ranked images of the discrete MRF in the test set of the 5k Corel set in response to the query word "birds"

Figure 5: 5 top ranked images of the discrete MRF in the test set of the TRECVID03 set in response to the query word "sport_event"

[6] G. Carneiro, A. B. Chan, P. J. Moreno, and N. Vasconcelos. Supervised learning of semantic classes for image annotation and retrieval. *IEEE Trans. Pattern Anal. Mach. Intell*, 29(3):394–410, 2007.

[7] P. Duygulu, K. Barnard, N. de Freitas, and D. Forsyth. Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary. In *Proceedings of the 7th European Conference on Computer Vision.*, pages 97–112, 2002.

[8] S. L. Feng, R. Manmatha, and V. Lavrenko. Multiple bernoulli relevance models for image and video annotation. In *Proc. CVPR*, pages 1002–1009, 2004.

[9] A. Ghoshal, P. Ircing, and S. Khudanpur. Hidden markov models for automatic annotation and content-based retrieval of images and video. In *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 544–551, 2005.

[10] J. S. Hare, P. H. Lewis, P. Enser, and C. J. Sandom. A linear-algebraic technique with an application in semantic image retrieval. In *In CIVR06*, 2006.

[11] G. Iyengar, P. Duygulu, S. Feng, P. Ircing, S. Khudanpur, D. Klakow, M. Krause, R. Manmatha, H. Nock, D. Petkova, B. Pytlik, and P. Virga. Joint visual-text modeling for automatic retrieval of multimedia documents. In *Proc. ACM Multimedia*, 2005.

[12] J. Jeon, V. Lavrenko, and R. Manmatha. Automatic image annotation and retrieval using cross-media relevance models. In *Proc. of the 26th Annual Int'l ACM SIGIR Conf.*, pages 119–126, Toronto, Canada, July 28-August 1 2003.

[13] J. Jeon and R. Manmatha. Using maximum entropy for automatic image annotation. In *Proceedings of the 3rd International Conference on Image and Video Retrieval*, pages 24–32, 2004.

[14] V. Lavrenko, S. L. Feng, and R. Manmatha. Statistical models for automatic video annotation and retrieval. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, pages 1044–1047, 2004.

[15] V. Lavrenko, R. Manmatha, and J. Jeon. A model for learning the semantics of pictures. In *Proceedings of Advances in Neural Information Processing Systems 16, NIPS 2003.*, 2003.

[16] J. Li and J. Z. Wang. Automatic linguistic indexing of pictures by a statistical modeling approach. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25:1075–1088, 2003.

[17] S. Li. *Markov Random Field Modeling in Image Analysis*. Springer-Verlag Telos, 2001. 2Rev Ed edition.

[18] J. Magalhães and S. M. Rüger. Logistic regression of generic codebooks for semantic image retrieval. In *CIVR*, pages 41–50, 2006.

[19] D. Metzler and W. B. Croft. A markov random field model for term dependencies. In *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR '05)*, pages 472–479, 2005.

[20] D. Metzler and R. Manmatha. An inference network approach to image retrieval. In *Proceedings of the 3rd International Conference on Image and Video Retrieval*, pages 42–50, 2004.

[21] W. Morgan, W. Greiff, and J. Henderson. Direct maximization of average precision by hill-climbing with a comparison to a maximum entropy approach. *Technical report, MITRE*, 2004.

[22] D. Nister and H. Stewenius. Scalable recognition with a vocabulary tree. In *Proc. of CVPR*, 2006.

[23] J. Philbin1, O. Chum1, M. Isard2, J. Sivic1, and A. Zisserman. Object retrieval with large vocabularies and fast spatial matching. In *Proc. of CVPR*, 2007.

[24] R. Shi, T.-S. Chua, C.-H. Lee, and S. Gao. Bayesian

learning of hierarchical multinomial mixture models of concepts for automatic image annotation. In *CIVR*, pages 102–112, 2006.

[25] S.Z.Li. *Markov Random Field Modeling in Image Analysis*. Springer-Verlag Telos; 2Rev Ed edition, 2001.

[26] L. Xie, L. Kennedy, S.-F. Chang, A. Divakaran, H. Sun, and C.-Y. Lin. Discovering meaningful multimedia patterns with audio-visual concepts and associated text. In *IEEE International Conference on Image Processing, October*, 2004.