Exploiting Spatial Context Constraints for Automatic Image Region Annotation

Jinhui Yuan, Jianmin Li and Bo Zhang
State Key Laboratory of Intelligent Technology and Systems
Department of Computer Science and Technology
Tsinghua University, Beijing, 100084, P. R. China
yuan-jh03@mails.tsinghua.edu.cn
{lijianmin, dcszb}@mail.tsinghua.edu.cn

ABSTRACT

In this paper we conduct a relatively complete study on how to exploit spatial context constraints for automated image region annotation. We present a straightforward method to regularize the segmented regions into 2D lattice layout, so that simple grid-structure graphical models can be employed to characterize the spatial dependencies. We show how to represent the spatial context constraints in various graphical models and also present the related learning and inference algorithms. Different from most of the existing work, we specifically investigate how to combine the classification performance of discriminative learning and the representation capability of graphical models. To reliably evaluate the proposed approaches, we create a moderate scale image set with region-level ground truth. The experimental results show that (i) spatial context constraints indeed help for accurate region annotation, (ii) the approaches combining the merits of discriminative learning and context constraints perform best, (iii) image retrieval can benefit from accurate region-level annotation.

Categories and Subject Descriptors

H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing—*Abstracting methods,Indexing methods*; I.5.1 [**Pattern Recognition**]: Models—*Statistical, Structural*

General Terms

Algorithms, Experimentation, Performance

Keywords

Spatial Context, Image Region Annotation, Graphical Model

MM'07, September 23–28, 2007, Augsburg, Bavaria, Germany.

1. INTRODUCTION

In this paper we consider a special task of image annotation, namely region annotation, which aims to automatically assign semantic labels to individual segmented regions rather than entire images.

1.1 Background and Motivation

The primary goal of image annotation is to facilitate image retrieval through the use of text [6]. Many existing image annotation approaches adopt part-based visual features, either grids [10, 13, 19] or segmented regions [1, 11, 12], but they do not aim to establish the correspondences between individual parts and semantic labels. Instead, they focus on assigning labels to entire images. Image-level annotation seems sufficient for the purpose of image retrieval. It is natural to ask whether region-level annotation (or object recognition [1, 2, 8, 29]), which seems more challenging, can bring some benefit to image retrieval. Recently, Fan *et al.* [9] and Yang *et al.* [34] have proposed different approaches for image-level annotation by firstly establishing correspondences between salient (or representative) regions and semantic labels. Their experiments reveal that, if region annotation is accurate enough, it can effectively boost the performance of image-level annotation. We continue this idea and focus on improving the accuracy of region annotation.

The first difficulty of region annotation is the lack of training set with region-level ground truth [2, 27, 32, 34]. In most available image set, descriptive keywords are associated with entire images rather than individual regions. Various approaches for learning from such weakly labeled data can be employed, e.g., expectation maximization (EM) algorithm [19, 8, 2] and multiple-instance learning (MIL)[4, 34]. However, image set with more detailed annotation is urgently needed to quantitatively measure performances of different approaches and focus research attention [27]. Some existing data set provides region-level annotation, but the scale is too small, usually at most several hundred images [8, 29, 2]. In this paper, we create a moderate size image set with mannually assigned region-level annotations and will study region annotation in supervised learning setting.

The second significant challenge of region annotation is how to exploit various context constraints to reduce ambiguities. It is frequent that regions with similar appearances correspond to distinct semantic concepts. For example, a smooth region in blue may be a part of sky or a part of ocean. It is difficult for computers even for human observers

to classify such regions without context. Previously, context constraints such as *scene context* [24, 9, 33] and *spatial context* [29, 2, 20] have been discussed. The former describes the co-occurrence relationships of different concepts while the latter characterizes the spatial layout constraints of concepts. In this paper we focus on investigating the spatial context.

Several methods exist to exploit spatial context constraints for tasks similar to region annotation. Typical examples include 2D Hidden Markov Model (2D HMM) [19, 13], Markov Random Fields (MRFs) [2] and Conditional Random Fields (CRFs) [20]. The above work makes beneficial exploration to exploit spatial context information, yet still several reasons motivate the work in this paper. First, most of the above work annotates fixed grids in images rather than segmented regions. Second, there is no comparative evaluation on context methods versus non-context ones previously. The problem to what extent spatial context can help remains unclear. Third, previous methods usually employ generative graphical models for the convenience of characterizing statistical dependencies among semantic labels. Nevertheless, some practical evaluations show that discriminative learning is preferable to generative approach in terms of classification performance [25, 15]. We wonder whether it is possible to combine the merits of both ideas, that is, integrating the classification performance of discriminative learning and the representation capability of graphical models. This idea has been explored in several other applications [15, 18]. Our goal is to investigate it in the scenario of image region annotation.

1.2 The Major Work

1. We present a straightforward yet effective method to regularize the segmented regions into 2D lattice layout, so that simple grid-structure graphical models can be employed.

2. We present a relatively complete study on how to exploit spatial context constraints for image region annotation. We show how to represent the spatial context constraints in graphical models and present related learning and inference algorithms. It is the first time for CRFs model with plugged-in SVMs to be used for region annotation.

3. We create a moderate scale image set with region-level annotation. We carry out extensive experiments to evaluate both the context versus non-context methods and the generative versus discriminative approaches. To our best knowledge, our experiment is so far the largest scale evaluation for region annotation in supervised learning setting. We also show the potential applications of region annotation in image retrieval.

1.3 Organization

The structure of this paper is as follows. In Section 2 we present the method of adjusting segmented regions into regular lattice layout. In Section 3 we formulate our task as a probabilistic inference problem. In Section 4 and Section 5, we present the representation, learning and inference algorithms for the models without context and with context constraints respectively. In Section 6, we describe the detailed procedure of creating image set with region-level annotation. In Section 7, we carry out the comparative experiments. Finally, we conclude the paper in Section 8.

2. REGULARIZING THE SPATIAL LAYOUT OF REGIONS INTO 2D LATTICE

We will only consider four types of neighboring dependencies (i.e., left, top, right and bottom). Such simple layout relationships are robust to familiar transformations. Moreover, they are convenient for representation,learning and inference (details in Section 5). However, it is not easy to obtain such neighborhood dependencies among segmented regions. On the one hand, the shapes of regions are so complex that usually no accurate top-bottom and left-right spatial relationships exist. On the other hand, the arbitrary sizes of regions often lead to the fact that several small regions may share the same large neighboring region in the same direction. These difficulties may explain why region-based representation has seldom been adopted in existing work exploiting spatial context information. For example, only symmetrical adjacency relation rather than unsymmetrical location relation of regions has been considered in [2, 23]. Singhal et al. [29] employ a modified *weighted walktrhough* algorithm to quantify the spatial relationship of two regions, and then define a set of rules to obtain spatial arrangement of regions. However, it is computation expensive and the resultant layout of regions is too complex to describe.

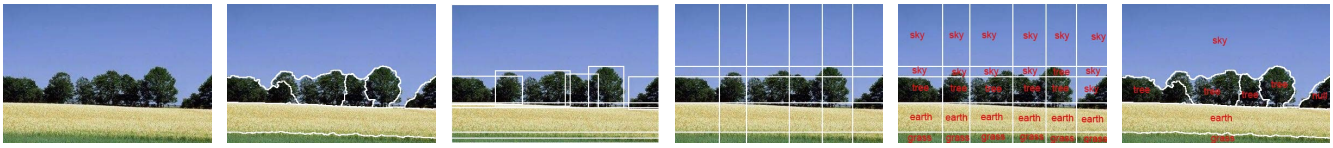
Algorithm 1: Convert Regions to 2D lattice

Input : Region set $\{R_i\}_{i=1}^{k}$ of image I
Output: Grid partition of image I

$\{t_i, b_i, l_i, r_i\}_{i=1}^{k}$ =`BoundingBox`($\{R_i\}_{i=1}^{k}$)
$\{h_j\}_{j=1}^{2k}$ =`AscendingSort`($\{t_i, b_i\}_{i=1}^{k}$)
$\{v_j\}_{j=1}^{2k}$ =`AscendingSort`($\{l_i, r_i\}_{i=1}^{k}$)
repeat
  $p = \arg\min_{j>2}(h_j - h_{j-1})$, $q = \arg\min_{j>2}(v_j - v_{j-1})$
  $\alpha_i = h_i - h_{i-1}$, $i \in \{p-1, p+1\}$
  $\beta_i = v_i - v_{i-1}$, $i \in \{q-1, q+1\}$
  if $\alpha_{p-1} < \alpha_{p+1}$ **then** *remove* h_{p-1} *from* $\{h_j\}$
  else *remove* h_p *from* $\{h_j\}$
  if $\beta_{q-1} < \beta_{q+1}$ **then** *remove* v_{q-1} *from* $\{v_j\}$
  else *remove* v_q *from* $\{v_j\}$
until *stop criterion is true*
draw horizontal line at $\{h_j\}$ *to* I
draw vertical line at $\{v_j\}$ *to* I

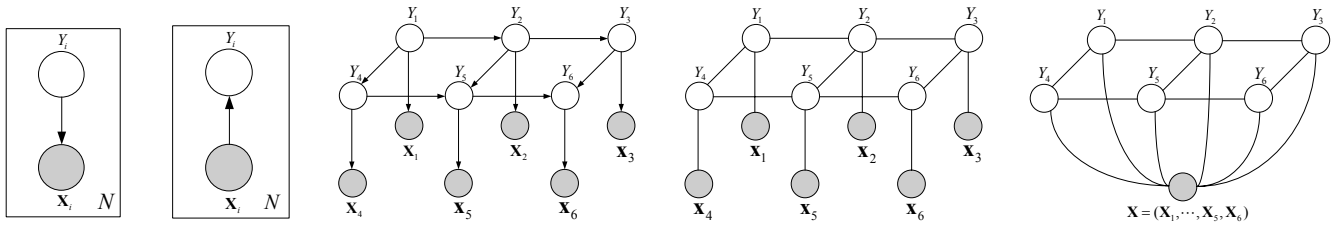
We propose a straightforward approach as illustrated in Algorithm 1 to adjust irregular regions into regular 2D lattice. The basic idea is: (a) using quadrate grids to approximate the regions so that exact left-right and top-bottom relationships can be obtained, (b) if one-to-many neighboring relationship occurs, partitioning large region into several grids to get one-to-one neighboring correspondence. In the algorithm, $\{t_i, b_i, l_i, r_i\}$ indicates the top, bottom, left and right coordinates of the i-th region's bounding box. The routine *BoundingBox* gets the coordinates of the given region's bounding box, and the routine *AscendingSort* sorts the coordinates into ascending order. To avoid over grid partition, an iteration is designed to merge too thin slices. We can stop the iteration once the width of the thinnest slice achieves a threshold or the number of remaining slices is below a predetermined value. Figure 1 illustrates the above procedure on an example image.

While preparing the training data, we do not manually



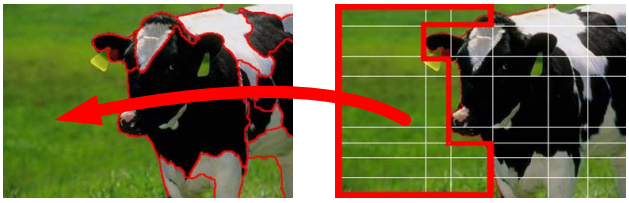
(a) original image (b) segmentation (c) bounding box (d) grid partition (e) annotated grid (f) annotated region

Figure 1: Procedure of regularizing the spatial layout of regions. After image segmentation, the bounding boxes of regions are obtained. With the help of the coordinates of bounding boxes, region adaptive grid partition can be obtained. Automated grid annotation is firstly carried out, and then grid level labels are propagated to the corresponding regions.



(a) (b) (c) (d) (e)

Figure 2: (a) i.i.d generative approach, (b) i.i.d discriminative approach, (c) 2D Hidden Markov Model (2D HMM), (d) Pairwise Markov Random Fields (MRFs), (e) Conditional Random Fields (CRFs).



Region segmentation by JSEG Region-adaptive grid partition

Figure 3: All the grids enclosed by the red polygon in the right image will inherit the label and feature from the region pointed by the arrow in the left image.

assign semantic labels to grids or extract grid-based visual features. Instead, we will firstly manually assign labels to regions and extract features from each region. And then we use Algorithm 1 to obtain region-adaptive grid partition and each grid automatically inherits semantic label and region-based visual features from the region to which most of its pixels belong, as shown in Figure 3. Hence, the grids belonging to the same region will automatically have the same labels and visual features. With above mappings between grids and regions, the learning and inference can be firstly carried out in regular grid-structure. Finally, by propagating the semantic labels from grids to regions, we can obtain region-based annotation. It is worth noting that the above strategy differs from the previous fixed grid partition methods [2, 13, 19, 20] in several aspects. Firstly, it is difficult to determine suitable grid size for fixed grid partition [13]. The grid partition yielded by Algorithm 1 is region adaptive and most of partitioning positions lie close to the region boundaries. Secondly, the labels and features inherited from regions avoid two obvious drawbacks of that of fixed grid partition, in which grids simultaneously covering several different objects lead to ambiguities for manually assigning semantic labels to them, and furthermore, the features extracted from grids may mix the visual appearances of different concepts. Finally, by propagating semantic labels from grids to corresponding regions, we can get object-level annotation rather than hard block-level annotation (e.g., example in Figure 1).

3. FORMULATING REGION ANNOTATION AS PROBABILISTIC INFERENCE

We denote random variables in upper-case letters while denote the the realizations of random variables in lower-case letters. Suppose we have N regions, let $(\mathbf{x}_i, y_i)$ denote the pair of feature and semantic label for the i-th region, where $\mathbf{x}_i$ is a d-dimensional vector and y_i is a discrete value in $\{1, \cdots, K\}$. Given the training data $\{(\mathbf{x}_i, y_i), i = 1, \cdots, N\}$, our task is to learn the mappings from region-based low level features to high level semantic labels. Since there is no determinate one-to-one correspondences between features and semantic labels, the desired mapping can be characterized by a posterior distribution of semantic labels conditioned on features, from which the label can be determined by *maximum a posteriori* (MAP) criterion.

If we assume the training data are independent, identically distributed (i.i.d), we can learn posterior distribution $p(Y|\mathbf{X})$ for individual regions as shown in Figure 2(a) and 2(b). However, the above i.i.d assumption is over-simplified. There actually exist strong spatial dependencies among the features and semantic labels for the regions in the same image. Let $\{(\mathbf{x}_i, y_i), i = 1, \cdots, R\}$ denote features and semantic labels of R regions in the same image, learning the posterior distribution of label configuration $p(Y_{1:R}|\mathbf{X}_{1:R})$ is more reasonable, with examples shown in Figure 2(c),2(d) and 2(e). There are basically two approaches to obtaining posterior distribution, i.e., generative approach and discriminative approach. Take the classifier under i.i.d assumption

as an example, generative and discriminative learning can be interpreted graphically in terms of the edge direction between $\mathbf{X}_i$ and Y_i in Figure 2. Generative approach firstly learns the class-conditional density $p(\mathbf{X}|y)$ for each discrete value y and the prior distribution $p(Y)$, and then employs the Bayes rule to calculate the posterior distribution $p(Y|\mathbf{X})$. On the contrary, the discriminative approach directly models the posterior distribution $p(Y|\mathbf{X})$. When classifying a region, it simply plugs the corresponding feature vector $\mathbf{x}$ into the conditional distribution and calculates $p(Y|\mathbf{x})$ directly.

4. MODELS WITHOUT SPATIAL CONTEXT CONSTRAINTS

Both Gaussian Mixture Models (GMM) and Support Vector Machines (SVMs) are models without considering spatial context constraints. Since they are very familiar techniques in many areas, in this section we will only present a brief description on them. The two approaches will also act as the baseline methods for the models exploiting the spatial dependencies in Section 5. GMMs is a generative modeling approach. We assume the prior probability $p(Y)$ obeys a multinomial distribution and the the class conditional density of each concept $p(\mathbf{x}|k), k \in 1, \ldots, K$ follows GMM. We can learn the parameters via maximum likelihood estimates (MLE) and expectation maximization (EM) algorithm. Finally, the label y of a region with feature vector $\mathbf{x}$ can be determined by MAP criterion. The basic model of SVMs is for binary-classification, while region annotation is a multi-classification task. Cusano *et al.* [5] employ one-against-all strategy to extend binary class to multi-class SVMs for image region annotation. Here we adopt one-against-one approach [3].

5. MODELS WITH SPATIAL CONTEXT CONSTRAINTS

2D HMM [19, 13], MRFs [23, 2] and CRFs [20] have been previously proposed for tasks similar to region annotation. In this section, we employ these models to region annotation. We present the details on how to represent the spatial dependencies among concepts as well as the related learning and inference algorithms. We identify the major differences between this work and the previous work.

5.1 2D Hidden Markov Model

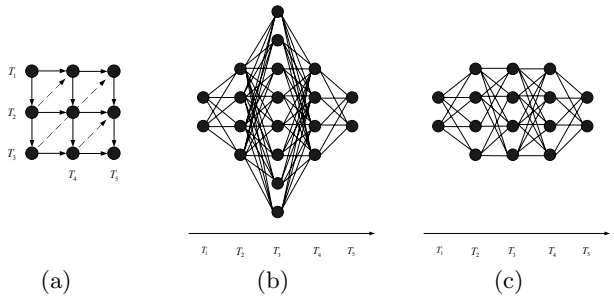
Several versions of 2D HMM exist [19, 13], of which we will adopt the one presented in [19]. Assuming homogeneous state transitions and using the fact that 2D HMM is a directed graphical model, we can factorize the log-likelihood of training data $\{(\mathbf{x}_i, y_i), i = 1, \cdots, N\}$ into

$$\mathcal{L}(\theta, \lambda) = \sum_{i=1}^{N} \log p(y_i|y_{\pi_i}) + \sum_{i=1}^{N} \log p(\mathbf{x}_i|y_i),$$

where π_i is the set of parents of node i, e.g., in Figure 2(c), $\pi_5 = \{2, 4\}$. Suppose $p(Y_i|y_{\pi_i})$ follows a multinomial distribution, the MLE can be obtained by counting the frequencies of state transitions. The class conditional density of each state can be further assumed as a GMM which can be learned by EM algorithm. It is worth noting that it is a supervised learning here rather than the unsupervised learning in [19]. Different from the MAP labeling for individual region in Section 4, 2D HMM pursues an MAP label configuration for the regions in the same image

$$y^*_{1:R} = \arg\max_{y_{1:R}} p(y_{1:R}|\mathbf{x}_{1:R}).$$

Li *et al.* show that path-constrained variable-state Viterbi algorithm can be used to approximately infer the MAP configuration [19]. We will present a brief introduction to the idea, since we will also employ it to obtain the sub-optimal MAP solutions for MRFs and CRFs.



(a) (b) (c)

Figure 4: (a) Label-field of 2D HMM, (b) Full state spaces of the corresponding chain-HMM, (c) State spaces of the path-constrained chain-HMM.

As shown in Figure 4(a) [1], the nodes on the same diagonal are firstly isolated to form a novel super-node, acting as the element of another chain-HMM. In this way, the MAP labeling of 2D HMM becomes the decoding problem of a chain-HMM. However, as shown in Figure 4(b), the scale of state spaces of the constructed chain-HMM increases exponentially with the width of the diagonal. To reduce the required computation, Li *et al.* propose to constrain the path of Viterbi algorithm to get a sub-optimal solution as shown in Figure 4(c). Concretely, only a sub-set of most likely state sequences are chosen into the state space at each phase. Following the notation in [36], let T_d denote the state sequence on diagonal d, the key problem is how to calculate the $|T_{d-1}| \times |T_d|$ state transition matrix $\mathbf{M}_d$ for the constructed chain-HMM. Let $M_d(T_{d-1}, T_d)$ denote the transition probability from state sequence T_{d-1} to T_d, we can calculate it by

$$M_d(T_{d-1}, T_d) = \prod_{y \in T_d} p(y|y_\pi \in T_{d-1}).$$

With the constrained state transition matrix $\mathbf{M}_d$, standard dynamic programming algorithm (i.e., Viterbi) can be used to obtain the approximate MAP solution.

5.2 Markov Random Fields

Different from 2D HMM, MRFs is a type of undirected graphical models as shown in Figure 2(d). Carbonetto *et al.* employ MRFs to perform generic object recognition [2]. Here we use it for region annotation with different representation, learning and inference algorithms. We adopt pairwise MRFs model, with which the joint probability of regions in the same image can be written as

$$p(\mathbf{x}_{1:R}, y_{1:R}) = \frac{1}{Z} \prod_{(i,j)} \Psi(y_i, y_j) \prod_k \Phi(y_k, \mathbf{x}_k),$$

[1]The figure originally appears in [19]

where (i, j) indicates the indices of neighboring nodes y_i and y_j, Z is known as the partition function, Ψ and Φ are pairwise potential functions characterizing the state-state interactions and observation-state associations respectively. The potential functions can be thought of as compatibility functions. A good definition of potential function assigns high values to the clique settings that are most compatible with each other under the given distribution.

5.2.1 Definition of Potential Functions

We firstly define some binary feature functions to capture the information on state transitions

$$\begin{aligned} h_{u',u}(Y_i, Y_j) &= \delta(Y_i = u')\delta(Y_j = u)\delta(i \dashv j) \\ v_{u',u}(Y_i, Y_j) &= \delta(Y_i = u')\delta(Y_j = u)\delta(i \perp j), \end{aligned} \quad (1)$$

where $h_{u',u}$ and $v_{u',u}$ indicate the horizontal and vertical transition functions respectively, u' and u are discrete values in $\{1, \cdots, K\}$, $i \dashv j$ indicates the i-th node is to the left of the j-th node and $i \perp j$ indicates the i-th node is above the j-th node, δ is an indicator function such that $\delta(\cdot)$ is 1 only if the contained assertion is true, and 0 otherwise. In the rest of this paper, we do not distinguish horizontal and vertical transitions, but denote both $h_{u',u}$ and $v_{u',u}$ by a single notation $f_{u',u}$. With the help of binary feature functions, the state-state interaction potential $\Psi(\cdot)$ can be defined as

$$\Psi(Y_i, Y_j) = \exp\left(\sum_{u',u} \lambda_{u',u} f_{u',u}(Y_i, Y_j)\right),$$

where $\lambda_{u',u}$ is the weight indicating the importance of transition from u' to u. These weights can be learned from the annotated training set. It is easy to see that $\Psi(\cdot)$ may be unsymmetrical, different from that of [2]. For observation-state potential $\Phi(\cdot)$, we directly define it as the probability of generating $\mathbf{X}_k$ given label Y_k

$$\Phi(Y_k, \mathbf{X}_k) = \exp\left(\log p(\mathbf{X}_k | Y_k)\right), \quad (2)$$

where $p(\mathbf{X}_k | Y_k)$ can be assumed to be the output of GMM.

5.2.2 Parameter Estimation

Given M i.i.d labeled images, the log-likelihood of the training data can be written as

$$\mathcal{L}(\lambda, \theta) = \underbrace{\sum_{m=1}^{M}\left(\sum_{(i,j)} \log \Psi(y_i, y_j) - \log Z_m\right)}_{prior} + \underbrace{\sum_{i=1}^{N} \log p(\mathbf{x}_i | y_i)}_{c.c.d},$$

where the parameters of class conditional density (c.c.d) and prior distribution of label-fields can be separately estimated. Again the GMM $p(\mathbf{X}|y)$ for each value of y can be learned by EM algorithm firstly. The MLE of λ can be obtained by maximizing the prior term of the log-likelihood. We use gradient ascent method to solve the optimization problem. The derivative of log-likelihood with respect to $\lambda_{u',u}$ can be given by

$$\frac{\partial \mathcal{L}}{\partial \lambda_{u',u}} = E_{\tilde{p}(Y)}[f_{u',u}] - E_{p(Y)}[f_{u',u}]$$

where $E_{\tilde{p}(Y)}[\cdot]$ is the expectation with respect to the empirical distribution and can be computed by counting the occurring of the corresponding event, $E_{p(Y)}[\cdot]$ is the expectation with respect to the model distribution. For feature functions $f_{u',u}$, we have

$$E_{p(Y)}[f_{u',u}] = \sum_{m=1}^{M} \sum_{(i,j)} p(y_i, y_j) f_{u',u}(y_i, y_j).$$

The major computation involved is to calculate the marginal probability $p(y_i, y_j)$, which is needed at each iteration of gradient ascent. However, due to the combinatorial property of partition function Z, exact computation of marginal probability is problematic even for problems of moderate size. Various approximate inference algorithms, e.g., Markov Chain Monte Carlo (MCMC), loopy belief propagation (LBP) and mean field (MF), can be used to obtain the approximate marginal probabilities. We adopt MF to calculate the approximate gradients and choose L-BFGS algorithm to maximize the log-likelihood because of their recent empirical success in training CRFs [28, 36].

5.2.3 MAP Labeling

Given the learned MRFs model and the feature vectors $\mathbf{x}_{1:R}$ of an image, the MAP label configuration can be approximately obtained by the approach similar to path-constrained Viterbi algorithm in Section 5.1. Using the fact that the partition function Z and the marginal probability $p(\mathbf{x}_{1:R})$ are constant for the given image, it is straightforward to have

$$y^*_{1:R} = \arg\max_{y_{1:R}} \prod_{(i,j)} \Psi(y_i, y_j) \prod_k \Phi(y_k, \mathbf{x}_k).$$

To use the path-constrained Viterbi algorithm, we only need to change the calculation of the transition probability matrix $\mathbf{M}_d$ according to

$$M_d(T_{d-1}, T_d) = \prod_{(i,j)} \Psi(y_i \in T_{d-1}, y_j \in T_d).$$

5.3 Discriminative Random Fields

Both 2D HMM and MRFs are generative frameworks that model the joint probability of the observed data and the corresponding labels. For computation tractability, they make strong assumptions on the data generating mechanism, that is, $p(\mathbf{x}_{1:R}|y_{1:R})$ is usually assumed to have a factorized form $p(\mathbf{x}_{1:R}|y_{1:R}) = \prod_k p(\mathbf{x}_k|\mathbf{y}_k)$, which may be oversimplified. Lafferty *et al.* propose a better alternative, i.e., conditional random fields (CRFs), which directly models the posterior distribution $p(\mathbf{x}_{1:R}|y_{1:R})$ as a Gibbs fields [17]

$$p(y_{1:R}|\mathbf{x}_{1:R}) = \frac{1}{Z(\mathbf{x}_{1:R})} \prod_{(i,j)} \Psi(y_i, y_j, \mathbf{x}_{1:R}) \prod_k \Phi(y_k, \mathbf{x}_{1:R}),$$

where the notation $\Psi(y_i, y_j, \mathbf{x}_{1:R})$ and $\Phi(y_k, \mathbf{x}_{1:R})$ make it explicit the fact that the potentials can depend on the features of the entire image. It has shown superior performance over the generative models in a variety of applications [17, 15]. Kumar *et al.* further clarify that the potential $\Phi(y_k, \mathbf{x}_{1:R})$ can be the output of any discriminative classifier and propose the discriminative random fields (DRFs)[15]

$$\begin{aligned} &p(y_{1:R}|\mathbf{x}_{1:R}) \\ &= \frac{1}{Z(\mathbf{x}_{1:R})} \prod_{(i,j)} \exp\left(I(y_i, y_j, \mathbf{x}_{1:R})\right) \prod_k \exp\left(A(y_k, \mathbf{x}_{1:R})\right), \end{aligned}$$

where A is the *association potential* which decides the association of a given site to a certain class ignoring its neighbors, I is the *interaction potential* which serves as a data dependent smoothing function. To sum up, there are three main differences between DRFs and MRFs [15]. First, the association potential in DRFs is a kind of discriminative classifier while in MRFs it is a generative classifier. Second, in DRFs, the association potential at any site, can be a function of all the observations, i.e., $\mathbf{X}_{1:R}$ while in MRFs it is the function of the data only at that site, i.e., $\mathbf{X}_k$. Third, the interaction potential in MRFs is a function of only labels, while in DRFs it can be a function of labels as well as the observations.

5.3.1 Definition of Potential Functions

Local Association Potential Both the output of logistic function [15] and the probability output of SVMs [18] have been previously adopted as the association potential in DRFs. Nevertheless, the previous work [15, 18] limits in binary classification applications. We extend it to multi classification case

$$A_l(Y_k, \mathbf{X}_{1:R}) = \log p(Y_k|\mathbf{X}_k), \tag{3}$$

where $p(Y_k|\mathbf{X}_k)$ is the probability output of multi-class SVMs [3]. Note that although A_l is allowed to be dependent on the features of entire image $\mathbf{X}_{1:R}$, we still limit it to $\mathbf{X}_k$ to emphasize it is a local association potential. It is different from the potential function in Equation 2 which is defined as the log-likelihood output of GMM.

Global Association Potential DRFs relaxes the strong assumption on the conditional independence of features, so that arbitrary complex features are allowed while defining the association potential. We design a simple method to explore this merit. We use k-means to group all the images into B clusters based on the global image features. Let t denote this assignment, we have $t(\mathbf{x}_{1:R}) = b$ where b is the index of the cluster to which the image with global features $\mathbf{x}_{1:R}$ belongs. We define a binary feature function to characterize such association between global features and semantic label

$$g_{u,b}(Y_k, \mathbf{X}_{1:R}) = \delta(Y_k = u)\delta(t(\mathbf{X}) = b).$$

With the help of binary feature, the global association potential can be defined as

$$A_g(Y_k, \mathbf{X}_{1:R}) = \sum_{u,b} \omega_{u,b} g_{u,b}(Y_k, \mathbf{X}_{1:R}). \tag{4}$$

This global potential can be added to the potential in Equation 3 to form a combined potential reflecting the evidence from both the local and global features.

Observation Independent Interaction Potential For simplicity, we can directly adopt the binary feature functions defined in Equation 1 to derive interaction potential

$$I(Y_i, Y_j, \mathbf{X}_{1:R}) = \sum_{u',u} \lambda_{u',u} f_{u',u}(Y_i, Y_j). \tag{5}$$

Note that this type of I is observation independent.

Observation Dependent Interaction Potential DRFs provides the choices of defining observation dependent interaction potentials. However, defining suitable such potentials for multi-classification is difficult. Meanwhile, whether observation dependent interaction potential will outperform the above observation independent interaction potential remains unclear [15, 16, 18]. Here, we make an initial attempt to define and evaluate observation dependent interaction potential in the task of region annotation. Following the way in defining global association potential, we encode feature vectors $(\mathbf{X}_i, \mathbf{X}_j)$ of neighboring regions into discrete values by k-means clustering. Assuming the assignment function as $s(\mathbf{X}_i, \mathbf{X}_j) = b$, we can define binary feature function

$$f_{u',u,b}(Y_i, Y_j, \mathbf{X}_i, \mathbf{X}_j) = \delta(s(\mathbf{X}_i, \mathbf{X}_j) = b)\delta(Y_i = u')\delta(Y_j = u).$$

With this feature function, the observation dependent interaction potential can be written as

$$I(Y_i, Y_j, \mathbf{X}_{1:R}) = \sum_{u',u,b} \lambda_{u',u,b} f_{u',u,b}(Y_i, Y_j, \mathbf{X}_i, \mathbf{X}_j), \tag{6}$$

which differs from the potential in Equation 5 that the pairwise feature vectors are encoded to force the label transitions consistent with the observation in the pair of parts.

5.3.2 Parameter Estimation

We can learn the parameters of DRFs similar to that of MRFs. We will only present the parameter estimation procedure for DRFs with association potential in Equation 3 and interaction potential in Equation 6. Other cases can be derived in similar way. DRFs can be trained by maximizing the log-likelihood of the M given images with respect to the conditional distribution

$$\mathcal{L}(\lambda, \mathbf{w}) = \sum_{m=1}^{M} \left(\sum_{(i,j)} I(y_i, y_j, \mathbf{x}_{1:R}) + \sum_k \log p(y_k|\mathbf{x}_k) - \log Z(\mathbf{x}) \right).$$

The parameters in local association potential within SVMs can be firstly learned by standard quadratic optimization. The parameters in interaction potential can be obtained by gradient ascent. The derivative of log-likelihood with respect to $\lambda_{u',u,b}$ can be given by

$$\frac{\partial \mathcal{L}}{\partial \lambda_{u',u,b}} = E_{\tilde{p}(\mathbf{X},Y)}[f_{u',u,b}] - \sum_{m=1}^{M} E_{p(Y|\mathbf{x_m})}[f_{u',u,b}],$$

where $E_{\tilde{p}(\mathbf{X},Y)}[\cdot]$ is the expectation with respect to the empirical distribution and $E_{p(Y|\mathbf{X})}[\cdot]$ is the expectation with respect to the conditional model distribution. It is straightforward that the techniques discussed in Section 5.2.2 can also be applied here to obtain approximate the MLE of parameters.

5.3.3 MAP Labeling

Given the trained DRFs model and the feature vectors $\mathbf{x}_{1:R}$ of an image, the MAP label configuration can be approximately obtained by path-constrained Viterbi algorithm similar to Section 5.1 [36]. Using the fact that the partition function $Z(\mathbf{x}_{1:R})$ is constant for the given image, it is straightforward to get

$$y^*_{1:R} = \arg\max_{y_{1:R}} \prod_{(i,j)} \Psi(y_i, y_j, \mathbf{x}_{1:R}) \prod_k \Phi(y_k, \mathbf{x}_{1:R}).$$

To use the path-constrained Viterbi algorithm, we only need to change the calculation of the transition probability matrix $\mathbf{M}_d$ according to

$$M_d(T_{d-1}, T_d) = \prod_{(i,j)} \Psi(y_i \in T_{d-1}, y_j \in T_d, \mathbf{x}_{1:R}).$$

Table 1: The number of images containing each concept and the number of regions for each concept.

Concept Name	Sky	Water	Mountain	Grass	Tree	Flower	Rock	Earth	Ground	Building	Animal	All
Image No.	3382	1690	1215	1660	2234	251	580	953	553	1852	477	4002
Region No.	13540	9257	9809	12820	19454	1701	6573	7598	1753	19422	2699	104626

Table 2: The Definition of Concept Lexicon.

Concept	Description
Sky	atmosphere, cloud, smoke, etc.
Water	river, sea, lake, fountain, etc.
Mountain	specifically for a distant sight of mountain
Grass	any vegetation except trees and flowers
Tree	trunks and leaves of trees
Flower	colorful plants
Rock	close observation of stone material
Earth	bare and natural land surface
Ground	manmade land surface such as road, square
Building	manmade structures such houses, bridges, etc.
Animal	skin of animals such as tigers, horses, etc.

6. DATA PREPARATION

To our best knowledge, no existing data corpora nicely fit the needs of automatic region annotation by supervised learning. To accurately study the context modeling approaches, here we create a novel moderate scale image set with manually assigned region-level semantic labels.

6.1 Data Corpora

To narrow the scope of detected concepts, we confine the selected images to those of outdoor scenes, including urban and natural pictures. Totally 4002 images out of 60,000 are chosen from Corel Stock Photo CDs. The Corel collection is the most broadly adopted data set in the community of image retrieval. Several drawbacks of using Corel set have been pointed out, of which an obvious one is that, in Corel CDs, every 100 images sharing the related semantics are stored in the same directory. There is sometimes a lack of diversity among the images in the same group. Therefore, we have paid special attention to address this drawback by selecting images in diverse appearances and prohibiting taking too many images from the same directory.

6.2 Lexicon Definition

The detailed description of the defined concepts is shown in Table 2. The lexicon is complete to cover all the concepts occurring in the image set. Therefore, we do not need to define an *outlier* class. The lexicon is also exclusive, that is, different concepts are not intersectant. With completeness and exclusiveness, each region has one and only one suitable semantic label. To achieve the above goals, we define concepts in the lexicon as the names of materials rather than the names of objects [29]. For example, we do not distinguish rivers, seas and lakes, but uniformly define the corresponding regions as *water*. Similarly, we do not distinguish tigers, lions, cows and horses, but uniformly define the corresponding regions as *animal* skins. When ambiguities exist, we categorize regions as concepts of the surface material. Take a region with forest on mountain as an example, we consider it as *tree* rather than *mountain* since trees cover on the mountain.

6.3 Automatic Image Segmentation

We adopt a state-of-the-art color texture image segmentation methods, i.e., JSEG [7]. Because of the integrated seed growing mechanisms, each region yielded by JSEG is spatially connected, which makes it convenient to describe the spatial relationships among regions. We choose a fixed set of parameters so that JSEG produces preferable over-segmentation results. With such fixed set of parameter values for whole image set, JSEG works well on most images. Finally, we get totally 104,626 regions for all the 4002 images, that is, average 26 regions for each image.

6.4 Manual Region Annotation

Manually assigning semantic labels to more than one hundred thousand segmented regions is both time-consuming and problematic. Nevertheless, it deserves the special laborious treatment for the accurate study and reliable evaluation of the proposed context modeling methods. We develop a human-computer interaction tool to facilitate the manual annotation of image regions. With this tool, users can browse the image segmentation results and assign predefined descriptive keywords to regions simply by mouse clicks. To avoid the inevitable subjective judgement of different annotators, all the images are annotated by the same person. For under-segmented regions, we assign the concept occupying the largest area to them. For too ambiguous regions, we infer and determine the concepts by taking the surrounding context into account. The region number of each concept is shown in Table 1.

6.5 Region Feature Extraction

Region-based low level visual features can be extracted to characterize visual appearance of corresponding regions. Since adopting what kind of visual features is not the crucial part of the proposed methods, here we simply extract two kinds of features, i.e., 9-dimensional color moment in HSV color space and 20-dimensional Pyramid-structured wavelet texture, which are then combined into a 29-dimensional feature vector.

7. EXPERIMENT RESULTS

In the experiments, 4002 images are randomly grouped into two sets in equal size as training and testing data respectively. The common adopted recall, precision, F-score and average precision (AP) are used to measure the performance of different approaches.

7.1 Statistical Spatial Dependencies

We count the frequencies of four neighboring relationships, i.e., *above*, *below*, *left*, *right*, among all the 11 concepts. The statistical results show that strong spatial dependencies exist among concepts. For example, Figure 5 shows the probability of one concept being above *sky*, *flower*, *building* respectively. Specifically, the first set of bars show that any region above *sky* is most likely to be *sky*, followed by *tree*, and no

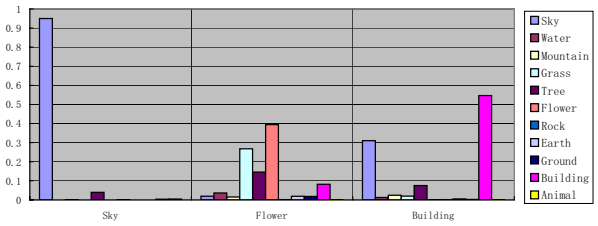


Figure 5: The probability of all the concepts being the above neighbor of *sky*, *flower* and *building*.

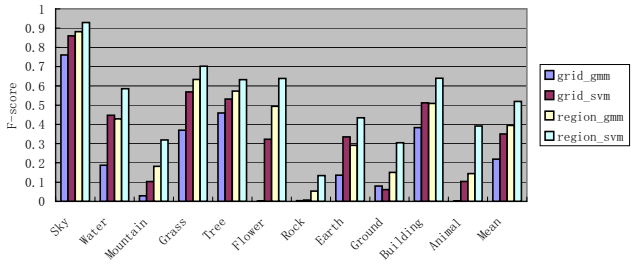


Figure 6: Fixed grid partition versus region adaptive grid partition.

other options. The second set of bars show that the region above *flower* is most probably *flower*, *grass*, and *tree*.

7.2 Fixed Grids vs. Region Adaptive Grids

We evaluate fixed grid partition and region adaptive grid partition using both GMM and SVMs. In fixed grid partition, we partition each image into 9×9 grids in equal size, so that the total number of grids can be almost equal to that of region adaptive grid partition [2]. For fixed grid partition, the labels are inherited from corresponding regions while the visual features are extracted for each grid. Their performances are evaluated by grid-level F-score. As shown in Figure 6, region adaptive methods significantly outperform the ones based on fixed grid partition.

7.3 Context Free vs. Spatial Context

We implement seven approaches, including *gmm* (GMM), *svm* (SVMs), *hmm* (2D HMM), *mrf* (MRFs), *drf_1* (DRFs with potentials defined in Equation 3 and Equation 5), *drf_2* (DRFs with potentials defined in Equation 3 and Equation 6) and *drf_3* (DRFs with potentials defined in Equation 3, Equation 4 and Equation 5). For *gmm*, 30 components are determined and a diagonal covariance matrix is assumed for each component. For *svm*, Gaussian kernel is adopted and parameters are chosen by a 5-fold cross validation procedure. The option of probability output is turned on and one-against-one strategy is used to perform multi-classification. For *hmm* and *mrf*, the class conditional density for each concept directly adopts the models obtained in *gmm* approach. For *drfs*, the local association potentials directly adopt the outputs of the trained SVMs models in *svm* approach. The cluster number B in Equation 4 and Equation 6 is fixed as 30. For the path-constrained Viterbi algorithm in *hmm*,*mrf* and *drfs*, the maximum size of the state spaces is restricted to 1000. While training MRFs and DRFs, we adopt several special treatments. First, the likelihood objective is penalized with a spherical Gaussian weight prior to avoid overfitting [28]. Second, mean field (MF) approach is adopted for approximate learning, since we find MF usually yields superior performance though it converges slightly slower than loopy belief propagation (LBP). Third, a simple feature selection procedure is devised to filter out too rare feature functions. L-BFGS algorithm is used to maximize the log-likelihood for both MRFs and CRFs [21, 36].

Table 3 shows the evaluation results. Though no single approach yields the best performance for all the concepts simultaneously. We have two prominent observations by comparing the overall performances. First, spatial context information indeed help. For example, *hmm* and *mrf* have respectively gained 12.8% and 8% improvement of F-score compared to *gmm*; *drfs* have gained 8% improvement against *svm*. Second, discriminative approaches uniformly outperform generative ones. For example, *svm* itself outperforms all the generative approaches, i.e., *gmm*, *hmm*, *mrf*. The best performance is achieved by *drfs*, which combines the merits of both discriminative learning and context constraints. *drf_3* gains 41% improvement against *gmm*. With a closer analysis to the results of *drfs*, we can also find, (a) *drf_3* outperforms *drf_1*, showing that the global association potential helps a little, (b) *drf_1* outperforms *drf_2*, showing that observation dependent interaction potential works worse than observation independent interaction potential though the latter leads to improvement on specific concepts, e.g., *flower* and *rock*. The possible reason is that the former imports too many sparse feature functions, which leads to a more challenging training problem.

[2]The average number of region adaptive grids for each image is different from that of regions for each image

7.4 Performance in Image Retrieval

We evaluate whether image retrieval can benefit from region annotation. The results are shown in Figure 7. g_svm, adopting global image feature and one-against-all strategy, trains SVMs models for the 11 concepts. r_svm and r_drf indicate the region annotation methods using SVMs and DRFs respectively, in which the labels with maximum confidence are propagated to image-level so that they can be compared with the results of g_svm. $g\_svm\_r\_drf$ fuses the results of g_svm and r_drf with equal weight. All the ranking lists returned by querying the 11 concepts are measured by average precision (AP). Note that *sky* is deliberately excluded since most of the images contain this concept(see statistics in Table 1). We have several observations according to the results shown in Figure 7. First, region annotation methods outperform g_svm in most concepts, especially for *flower* and *ground*, while for *animal*, g_svm outperforms region annotation methods. Second, region annotation methods outperform g_svm in overall performance, with nearly 16% improvement on mean-AP compared to g_svm. Third, the fusion of image-level annotation (g_svm) and region-level annotation ($g\_svm\_r\_drf$) performs best, with 21.2% improvement of mean-AP with respect to g_svm. A close analysis to the ranking lists reveals why region annotation significantly outperforms g_svm on *flower* and *ground* while performs poor on *animal*. Both *flower* and *ground* are rare concepts (see the statistics in Table 1), and the region areas containing these concepts are not dominant in the global images. For such concepts, if the region annotation is ac-

Table 3: Evaluation results of context free versus spatial context. Note that we abbreviate the concept names to save space. The measures in the last row are the average metric of the corresponding columns.

	Precision							Recall							F-score						
	gmm	svm	hmm	mrf	drf_1	drf_2	drf_3	gmm	svm	hmm	mrf	drf_1	drf_2	drf_3	gmm	svm	hmm	mrf	drf_1	drf_2	drf_3
sky	.937	**.961**	.933	.914	.955	.908	.960	.832	.899	.889	.871	.909	**.942**	.921	.882	.929	.911	.892	.931	.925	**.940**
wat.	.410	.583	.531	.390	.598	**.613**	.612	.449	.588	.489	.496	**.639**	.524	.607	.429	.585	.509	.437	.618	.610	**.630**
mnt.	.134	.269	.215	.299	.395	**.442**	.374	.282	.392	.315	.313	**.435**	.393	.367	.182	.319	.255	.306	.401	.402	**.408**
grs.	.616	.655	.651	.591	.661	**.689**	.671	.652	.757	.679	.696	**.780**	.670	.708	.633	.702	.665	.639	.715	.699	**.719**
tre.	.709	**.765**	.611	.615	.755	.583	.745	.481	.538	.556	.532	.570	.570	**.594**	.573	.632	.582	.571	**.650**	.588	**.650**
flr.	.475	.591	.584	.561	.615	.619	**.622**	.513	.694	.447	.411	.695	**.808**	.701	.494	.639	.507	.474	.653	.658	**.672**
rck.	.033	.088	.198	.242	.220	**.413**	.300	.132	.281	.216	.189	**.341**	.235	.234	.050	.134	.207	.212	.268	.298	**.309**
ert.	.230	.386	.337	.184	.294	**.412**	.351	.397	.497	.328	.372	**.539**	.417	.487	.291	.434	.332	.246	**.445**	.366	.423
grd.	.099	.208	.433	**.461**	.379	.230	.372	.316	**.569**	.220	.187	.509	.513	.465	.151	.305	.292	.265	**.424**	.418	.413
bld.	.610	.730	.484	.481	.625	.658	**.725**	.437	.569	.582	.550	.645	.646	**.677**	.509	.640	.529	.513	**.687**	.650	**.687**
anl.	.096	.297	.295	.312	.480	.484	**.495**	.294	.573	.285	.204	.540	**.751**	.497	.144	.392	.290	.247	.508	.491	**.520**
avg.	.395	.503	.479	.459	.560	.550	**.566**	.435	.578	.455	.438	**.600**	.586	.571	.414	.538	.467	.448	.579	.561	**.583**

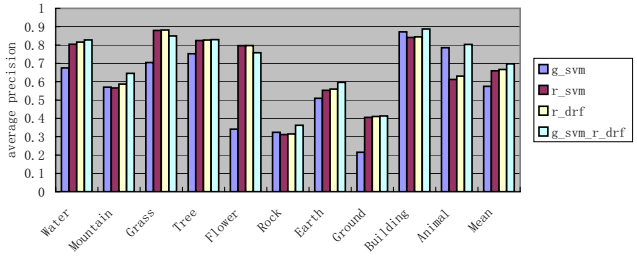


Figure 7: Performance in concept retrieval. *Mean* indicates the mean-AP over all the concepts.

curate, it will rank relevant images at the top positions in the lists, that is why it beats the method using global image features (i.e., *g_svm*). *animal* possesses similar characteristics, e.g., rare and not dominant. However, on the one hand, due to both the simple 29-dimensional region-based features and the limited local context (only first-order Markov property considered in DRFs), region annotation is not accurate enough for *animal*. On the other hand, *animal* has strong correlations with the global scene. In the result, *g_svm* using global features beat region annotation methods.

We also compare DRFs with the state-of-art Cross Media Relevance Model (CMRM) [11], which employs region-based visual features while performs image-level annotation. CMRM does not support ranking, we can not evaluate them by AP. Instead, we compare them in the cases of assigning fixed number of keywords to images, which can be evaluated by F-score. Again the keywords with maximum confidence are propagated to image-level in DRFs. In CMRM, we adopt the same parameter settings to [11]. Table 4 shows that DRFs uniformly outperforms CMRM. The above results show that region-level annotation not only provides the possibility of locating objects in images, but also brings improvements to image-level CBIR.

Table 4: CMRM versus DRFs for fixed length image-level annotation.

	1 word	2 words	3 words	4 words	5 words
cmrm	0.240	0.380	0.459	0.499	0.526
drf	**0.281**	**0.469**	**0.584**	**0.635**	**0.642**

8. CONCLUSIONS AND DISCUSSIONS

In this paper we present a relatively complete study on how to exploit spatial context constraints for automated image region annotation. We design a simple yet effective approach to regularize the segmented regions into 2D lattice layout, so that simple grid-structure graphical models can be employed. We create a moderate size image set with region-level annotation and carry out extensive experiments to evaluate several classical methods. To our best knowledge, our experiment is so far the largest scale evaluation for region annotation in supervised learning setting. The experimental results show that (i) spatial context constraints indeed help for accurate region annotation, (ii) the approaches combining the merits of discriminative learning and context constraints perform best. These experiments provide useful guide for building real-world systems.

There are several limitations on current work. First, our approach to combining SVMs and CRFs is not seamless. The parameters in SVMs and CRFs are estimated sequentially but not simultaneously, which may not be optimal. There exist models unifiedly integrating the large margin mechanisms into CRFs such as Max-margin Markov network [30, 31], but the required computation is too expensive to be used to large scale applications. Second, we have only considered the local spatial constraints (i.e., first-order Markov). Overall scene context [24, 9, 33] may be incorporated for further improvement. Finally, so far we have only discussed models in supervised learning settings. However, manually annotating each region is rather tedious and extremely costly. There are two possible solutions for this problem. The first choice is to transform the tedious manual annotation to an enjoyable game similar to *Peekaboom*[32]. Another choice seems more appealing, that is, using advanced machine learning techniques to learn the correspondences between regions and labels from weakly labeled data (i.e., image-level ground truth). Such machine learning approaches include multiple instance learning (MIL) [4, 34, 35] and expectation maximization (EM) [2, 8, 19, 26].

Therefore, a critical problem is whether the supervised training in our work is significantly better than those approaches without requiring region-level annotated training

data (i.e., MIL and EM). Unfortunately, we have not conducted the comparative study on this topic. We conjecture that the supervised training will indeed significantly outperform both MIL and EM, based on the evidences from the evaluation in related fields. Specifically, Jun Yang's [35] empirical evaluation on image retrieval shows that MIL algorithms running on local region features only achieve comparable performance to that of SVMs running on global image features, while in our experiments region annotation approach apparently outperforms SVMs on global image features. As for EM algorithms, though no comparative study exists for image region annotation, similar evaluation was previously conducted on the task of part-of-speech tagging (POS) [14, 22]. The evaluation results show that supervised training HMM significantly outperforms the unsupervised training HMM (i.e.,EM) on POS task (more than 10% improvement) [14, 22].

9. ACKNOWLEDGMENTS

We would like to thank the anonymous reviewers for their insightful suggestions. We also thank Dr. Chih-Jen Lin for the code of *libSVM*, Dr. Kevin Murphy for the code of *2DCRFs*. Finally, special thanks go to Jun Zhu for helpful discussions on graphical models. This work was supported by National Natural Science Foundation of China (60621062, 60605003) and Chinese National Key Foundation Research & Development Plan(2003CB317007, 2004CB318108).

10. REFERENCES

[1] K. Barnard, P. Duygulu, D. Forsyth, N. de Freitas, D. M. Blei, and M. I. Jordan. Matching words and pictures. *J. Mach. Learn. Res.*, 3:1107–1135, 2003.

[2] P. Carbonetto, N. de Freitas, and K. Barnard. A statistical model for general contextual object recognition. In *Proc. of ECCV 2004*, pages 350–362.

[3] C.-C. Chang and C.-J. Lin. LIBSVM: A Libary for Support Vector Machines. 2005. available at http://www.csie.ntu.edu.tw/~cjlin/libsvm/.

[4] Y. Chen and J. Z. Wang. Image categorization by learning and reasoning with regions. *J. Mach. Learn. Res.*, 5:913–939, 2004.

[5] C. Cusano, G. Ciocca, and R. Schettini. Image annotation using SVM. In *Proc. of SPIE IS&T Electronic Imaging 2004*, pages 330–338.

[6] R. Datta, J. Li, and J. Z. Wang. Content-based image retrieval - approaches and trends in the new age. In *Proc. of MIR Workshop 2005*, pages 253–262.

[7] Y. Deng and B.S.Manjunath. Unsupervised segmentation of color-texture regions in images and video. *IEEE Trans. Pattern Anal. Mach. Intell.*, 23(8):800–810, 2001.

[8] P. Duygulu, K. Barnard, N. de Freitas, and D. Forsyth. Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary. In *Proc. of ECCV 2002*, pages 97–112.

[9] J. Fan, Y. Gao, and H. Luo. Multi-level annotation of natural scenes using dominant image components and semantic concepts. In *Proc. of ACM Multimedia 2004*, pages 540–547.

[10] A. Ghoshal, P. Ircing, and S. Khudanpur. Hidden Markov models for automatic annotation and content-based retrieval of images and video. In *Proc. of SIGIR 2003*, pages 544–551.

[11] J. Jeon, V. Lavrenko, and R. Manmatha. Automatic image annotation and retrieval using cross-media relevance models. In *Proc. of SIGIR 2003*, pages 119–126.

[12] R. Jin, J. Y. Chai, and L. Si. Effective automatic image annotation via a coherent language model and active learning. In *Proc. of ACM Multimedia 2004*, pages 892–899.

[13] J. Jiten, B. Mérialdo, and B. Huet. Semantic feature extraction with multidimensional hidden Markov model. In *Proc. of SPIE CMCAMR 2006*, volume 6073, pages 211–221.

[14] M. Johnson. Why doesn't EM find good HMM POS-taggers? In *Proc. of EMNLP 2007*, pages 296–305.

[15] S. Kumar and M. Hebert. Discriminative random fields: A discriminative framework for contextual interaction in classification. In *Proc. of ICCV 2003*, pages 1150–1159.

[16] S. Kumar and M. Hebert. Multiclass discriminative fields for parts-based object detection. In *Proc. of Snowbird Learning Workshop*, 2004.

[17] J. Lafferty, A. McCallum, and F. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proc. of ICML 2001*, pages 282–289.

[18] C. H. Lee, R. Greiner, and M. Schmidt. Support vector random fields for spatial classification. In *Proc. of PKDD 2005*, pages 121–132.

[19] J. Li, A. Najmi, and R. M. Gray. Image classification by a two-dimensional hidden Markov model. *IEEE Trans. Signal Processing*, 48(2):517–533, 2000.

[20] W. Li and M. Sun. Semi-supervised learning for image annotation based on conditional random fields. In *Proc. of CIVR 2006*, pages 463–472.

[21] D. C. Liu and J. Nocedal. On the limited memory BFGS method for large scale optimization. *Mathematical Programming*, 45:503–528, 1989.

[22] B. Merialdo. Tagging english text with a probabilistic model. *Computational Linguistics*, 20(2):155–171, 1994.

[23] J. W. Modestino and J. Zhang. A Markov random field model-based approach to image interpretation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 14(6):606–615, 1992.

[24] M. R. Naphade and J. R. Smith. A hybrid framework for detecting the semantics of concepts and context. In *Proc. of CIVR 2003*, pages 196–205.

[25] M. R. Naphade and J. R. Smith. On the detection of semantic concepts at TRECVID. In *Proc. of ACM Multimedia 2004*, pages 660–667.

[26] A. Quattoni, M. Collins, and T. Darrell. Conditional random fields for object recognition. In *Proc. of NIPS 2004*, 2005.

[27] B. C. Russell, A. Torralba, K. P. Murphy, and W. T. Freeman. Labelme: a database and web-based tool for image annotation. In *MIT AI Lab Memo AIM-2005-025*, 2005.

[28] F. Sha and F. Pereira. Shallow parsing with conditional random fields. In *Proc. of NAACL 2003*.

[29] A. Singhal, J. Luo, and W. Zhu. Probabilistic spatial context models for scene content understanding. In *Proc. of CVPR 2003*.

[30] B. Taskar, C. Guestrin, and D. Koller. Max-margin markov networks. In *Proc. of NIPS 2003*, 2004.

[31] I. Tsochantaridis, T. Joachims, T. Hofmann, and Y. Altun. Large margin methods for structured and interdependent output variables. *J. Mach. Learn. Res.*, 6:1453–1484, 2005.

[32] L. von Ahn, R. Liu, and M. Blum. Peekaboom: a game for locating objects in images. In *Proc. of SIGCHI 2006*, pages 55–64.

[33] R. Yan, M.-Y. Chen, and A. Hauptmann. Mining relationship between video concepts using probabilistic graphical models. In *Proc. of ICME 2006*, pages 301–304.

[34] C. Yang, M. Dong, and F. Fotouhi. Region based image annotation through multiple-instance learning. In *Proc. of ACM Multimedia 2005*, pages 435–438.

[35] J. Yang. MILL: A multiple instance learning library. 2006. available at http://www.cs.cmu.edu/~juny/MILL.

[36] J. Zhu, Z. Nie, J.-R. Wen, B. Zhang, and W.-Y. Ma. 2D conditional random fields for web information extraction. In *Proc. of ICML 2005*, pages 1044–1051.