Conditional Random Field for Natural Scene Categorization

Yong Wang and Shaogang Gong
Department of Computer Science
Queen Mary, University of London
{ywang,sgg}@dcs.qmul.ac.uk

Abstract

Conditional random field (CRF) has been widely used for sequence labeling and segmentation. However, CRF does not offer a straightforward approach to classify whole sequences. On the other hand, hidden conditional random field (HCRF) has been proposed for whole sequences classification by viewing the segment labels as hidden variables. But the objective function of HCRF is non-convex because of its hidden variable structure. In this paper, we propose a classification oriented CRF (COCRF) adapted from HCRF for natural scene categorization by taking an image as an ordered set of local patches. Our approach firstly assigns a topic label to each segment on the training data by the probabilistic latent semantic analysis (PLSA) and train a COCRF model given these topic labels. PLSA provides a higher level of semantic grouping of image patches by considering their co-occurrence relationships while COCRF provides a probabilistic model for the spatial layout structure of image patches. The combination of PLSA and COCRF can not only classify but also interpret scene categories. We tested our approach on two well-known datasets and demonstrated its advantage over existing approaches.

1 Introduction

This paper addresses the problem of natural scene categorization. Scene understanding underlies many other problems in visual perception such as object recognition and environment navigation. Although scene categorization can be achieved at a glance by a human, it poses great challenges to a computer vision system. Different instances of the same category can vary a lot in their color distribution, texture patterns and more importantly, a scene category does not have a well-defined shape as an object category does.

Recent work in scene image classification focus on image classification based on an intermediate level of features. They can be further divided into two categories. The first relies on self-defining the intermediate features. Oliva and Torralba [7] proposed a set of perceptual dimensions (naturalness, openness, roughness, expansion and ruggedness) that represent the dominant spatial structure of a scene. Each of these dimensions can be automatically extracted and scene images can then be classified in this low-dimensional representation. Vogel and Schiele [8] used the occurring frequency of different concepts (water, rock, *etc*) in an image as the intermediate features for scene image classification,

and they need manual labeling of each image patch in the training data. While manual labeling can improve the semantic interpretation of images, it is still a luxury for a large dataset and it can also be inconsistent in defining a common set of concepts [8]. The second kind of approach is aimed to alleviate this burden of manual labeling and learn the intermediate features automatically. This is achieved by making an analogy between a document and an image and taking advantage of the existing document analysis approaches. For example, Fei-Fei and Perona [2] proposed a Bayesian hierarchical model extended from latent dirichlet allocation (LDA) to learn natural scene categories. Bosch et al. [1] achieved good performance in scene classification by combining probabilistic latent semantic analysis (PLSA) [3] and a KNN classifier. A common point of these approaches is that they represent an image as a bag of orderless visual words. An exception is the work done by Lazebnik et al. [6] where they proposed spatial pyramid matching for scene image classification by partitioning an image into increasingly fine sub-regions and taking each sub-region as a bag of visual words.

As a simple but discriminative enough representation, the bag of visual words has shown its advantage in the above approaches. However, its assumption of an orderless bag makes it inevitably sacrifice certain amount of discriminative capability. The order statistics are actually quite helpful in our understanding of scenes. At least two cues can be applied. The first is the spatial layout of the patches. For example, *sky* always appear in the upper part of an image and *ground* almost always appear in the bottom part. Lazebnik et al. [6] have demonstrated the advantage of this cues, but they did not do it in a probabilistic model. The second cue is the spatial pairwise interaction between two neighboring patches. For example, it is more likely to find a *water* patch as the neighbor as a *sand* patch in a *beach* scene, while in a *coast* scene water patches are usually adjacent to *stone* patches. None of the existing approaches have modeled both of these two relations explicitly in a probabilistic model.

A good candidate for modeling a set of ordered local patches is the conditional random field (CRF) [5]. For example, Kumar and Hebert [4] attempted to use a discriminant random field to model contextual interaction between image patches. But their work was for image region classification, instead of whole image classification. Generally speaking, CRF is aimed for segment labeling and segmentation. It does not offer a straightforward approach to classify whole sequences and requires the labeling of the segments in the training data. Hidden conditional random field (HCRF) [9] was proposed for whole sequences classification by viewing the segment labels as hidden variables, but the hidden variable structure makes the objective function of HCRF non-convex and only local optimum can be achieved in training. In this paper, we proposed a combinational approach of PLSA and a classification oriented CRF (COCRF) adapted from HCRF for natural scene categorization by taking an image as an ordered set of image patches. COCRF takes the advantage of automatic labels generated by PLSA and is capable of reaching a global optimum in the training stage. The motivations of PLSA here are not only that it can provide labeling of the image patches, but also that it is complimentary to COCRF, i.e., PLSA can discover the *co-occurrence* relationship between image patches, while COCRF can only model *spatial* relation between patches. Thus our PLSA+COCRF model can take into account both of these two factors. An obvious advantage of our approach is to provide a probabilistic way to model both the spatial layout of image patches and their neighboring interaction. We tested our approach on two scene image image datasets and show that it outperforms existing approaches.

The rest of this paper is organized as follows. Section 2 describes the topic labeling of image patch by PLSA. Section 3 introduces COCRF and focus on the features we have deployed. Section 4 discusses the learning and inference of COCRF for classification. We show some experimental results in section 5 and conclude in section 6.

2 Automatic Topic Labeling of Image Patches via PLSA

In our approach, an image is represented as a number of image patches. Each patch is assigned a topic label automatically through PLSA [3]. PLSA can be summarized as follows. Suppose we have a collection of text documents $\mathscr{D}=\{d\}$, a vocabulary $\mathscr{W}=\{w\}$ and a number of topics $\mathscr{S}=\{s\}$. Each document d is represented as a bag of words, i.e, we keep only the counts $n(d,w)$ which indicates the number of occurrence of word w in document d. PLSA assumes that each word in a document is generated by a specific topic. Given the topic distribution of a document, its word distribution is independent from the document. More precisely, the probability of a word w in a document d is a marginalization over topics, i.e.,

$$P(w|d) = \sum_{s\in\mathscr{S}} P(w|s)P(s|d) \tag{1}$$

Given $\mathscr{D}$ and $P(w|d)$, the parameters $P(s|d)$ and $P(w|s)$ can be estimated by an EM algorithm [3]. To adapt PLSA to image data, we transform images into the bag of visual words representation by the following procedures: (i) Partition each image into a number of small patches. (ii) Learn a visual vocabulary on the descriptors of a subset of local patches by k-means clustering. (iii) Assign a visual word to each local patch. After a PLSA model is learned from the training images, we can obtain the topic labeling s of a visual word w in a specific document d by the following equation

$$P(s|w,d) = \frac{P(w|s)P(s|d)}{P(w|d)} \tag{2}$$

The ending results of PLSA is that each image patch has a topic label.

3 Classification Oriented Conditional Random Field (COCRF)

Our final objective is to assign a scene category label to a given image. The training data is $\{(y^{(k)},\mathbf{x}^{(k)},\mathbf{s}^{(k)})\}$, where $y^{(k)}$ is the category label, $\mathbf{x}^{(k)}=\{x_1^k,x_2^k,x_{n_k}^k\}$ are the visual features of each image patch, $\mathbf{s}^{(k)}=\{s_1^k,s_2^k,s_{n_k}^k\}$ are the corresponding topic labels of the image patches obtained by PLSA. k is the index of the training image. The graphical structures of CRF, HCRF and COCRF are illustrated in Fig. 1. In these graphic models, we have taken an image with four local patches (which we also refer to as segments) as an example. The scene category label is denoted by variable y and $\mathbf{s}=\{s_1,s_2,s_3,s_4\}$ are the topic labels of the image patches. The image observation is denoted by variables $\mathbf{x}=\{x_1,x_2,x_3,x_4\}$. The edges between nodes represent their inter-dependence. The shaded nodes in HCRF indicate these nodes are hidden variables. In our model, we consider the graphic structure of nodes $\mathbf{s}$ as a lattice with pairwise potentials. In a CRF model, we have only the topic labels and the image observation. In HCRF we have an additional node y but $\mathbf{s}$ is not observed. In COCRF we have the node y and all the nodes $\mathbf{s}$ are observed.

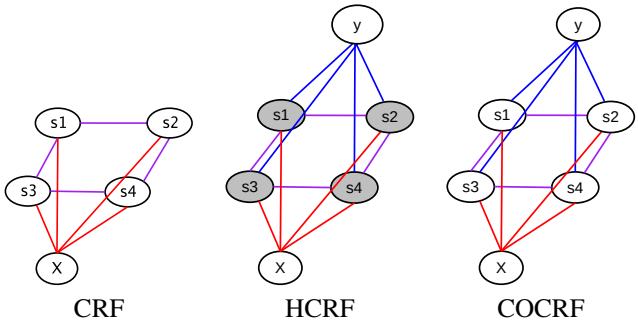


Figure 1: Graphical models of conditional random field (CRF), hidden conditional random field (HCRF) and classification oriented conditional random field (COCRF).

Following the definition of a CRF model, the conditional probability for the topic labels $\mathbf{s}$ and the category label y given the observation $\mathbf{x}$ can be expressed as

$$P(y,\mathbf{s}|\mathbf{x};\theta) = \frac{e^{\psi(y,\mathbf{s},\mathbf{x};\theta)}}{\sum_{y',\mathbf{s}'} e^{\psi(y',\mathbf{s}',\mathbf{x};\theta)}} \tag{3}$$

where θ represents the parameters of the model. $e^{\psi(y,\mathbf{s},\mathbf{x};\theta)}$ is the potential function. In COCRF, we consider three types of potential and we write the log potential function $\psi(y,\mathbf{s},\mathbf{x};\theta)$ as the summation of three terms. Each term can be viewed as a different type of features deployed for classification.

$$\psi(y,\mathbf{s},\mathbf{x};\theta) = \underbrace{\psi^a(y,\mathbf{s},\mathbf{x};\theta)}_{\text{node appearance potential}} + \underbrace{\psi^e(y,\mathbf{s},\mathbf{x};\theta)}_{\text{edge potential}} + \underbrace{\psi^s(y,\mathbf{s};\theta)}_{\text{node spatial potential}} \tag{4}$$

3.1 Appearance Potential

The appearance potential measures the compatibility between a topic label and its appearance. This potential is a kind of low-level features and it is shared among different scene categories.

$$\psi^a(y,\mathbf{s},\mathbf{x};\theta) = \sum_{j=1}^{m} \phi(\mathbf{x},j)\cdot\theta^a(s_j) \tag{5}$$

where j is the index of a segment (patch) and m is the total number of segments. $\phi(\mathbf{x},j) \in \mathbb{R}^d$ is a feature extraction function which maps the observation at site j to a d-dimensional feature vector. $\theta^a(s_j)$ is the appearance parameter vector corresponding to the segment label $s_j \in \mathscr{S}$.

Considering the diversity in appearance of each topic, we map the local observation to a feature vector by a Gaussian Mixture Model (GMM). Suppose we have a set of Gaussian components $\{g_1,g_2,\ldots,g_d\}$, each of which has its own parameters of the mean and variance. The feature extraction function is represented as,

$$\phi(\mathbf{x},j) = \left[g_1(x_j), g_2(x_j), \cdots, g_d(x_j)\right]^t \tag{6}$$

where x_j is the appearance descriptor of segment j. To obtain the set of Gaussian components $\{g_1,g_2,\ldots,g_d\}$, we firstly collect a subset of local patches of each topic and fit a GMM to each topic. The final set of Gaussian components are the combination of all the Gaussian components for each topic.

3.2 Edge Potential

The edge potential models the interaction between neighboring patches. It is similar to that in CRF but it is category dependent. This provides COCRF more discriminative capability between different categories, as follows

$$\psi^e(y,\mathbf{s},\mathbf{x};\theta) = \sum_{(j,k)\in E} \theta^e(s_j,s_k,y) \tag{7}$$

where θ^e is symmetric with respect to s_j and s_k. E is the set of all the edge links between the segment nodes depending on the 2-D lattice structure.

3.3 Spatial Layout Potential

Here we take an explicit approach by dividing the image area into 3× 3=9 sub-regions. We examine the the spatial layout distribution of each topic on this 3×3 grid.

$$\psi^s(y,\mathbf{s};\theta) = \sum_{j}^{m} \theta^s(y,s_j,\eta(j)) \tag{8}$$

where $\eta(j) \in \{1,2,\ldots,9\}$ denotes the deterministic mapping function of a site j into the sub-region it sits in. It is worth noting that if θ^s does not depend on the spatial location of node j, this potential will degrade to the one as same as that in HCRF [9].

4 Learning

In the training process we learn the model parameter $\hat{\theta}$ by maximizing its log likelihood on the training data. Assume the training data is *i.i.d.*, $\hat{\theta}$ is obtained by,

$$\hat{\theta} = \arg\max_{\theta} \mathscr{L}(\theta) = \arg\max_{\theta} \sum_{k=1}^{n} \mathscr{L}^k(\theta) \tag{9}$$

where $\mathscr{L}^k(\theta)$ is the log likelihood of the k-th sample and n is the total number of training samples. Since $\mathbf{s}^{(k)}$ is observed, we have

$$\mathscr{L}^k(\theta) = \log P(y^{(k)},\mathbf{s}^{(k)}|\mathbf{x}^{(k)};\theta) = \log\left(\frac{e^{\psi(y^{(k)},\mathbf{s}^{(k)},\mathbf{x}^{(k)};\theta)}}{\sum_{y',\mathbf{s}'} e^{\psi(y',\mathbf{s}',\mathbf{x}^{(k)};\theta)}}\right) = \psi(y^{(k)},\mathbf{s}^{(k)},\mathbf{x}^{(k)};\theta) - \log\sum_{y',\mathbf{s}'} e^{\psi(y',\mathbf{s}',\mathbf{x}^{(k)};\theta)} \tag{10}$$

This equation is different from that in HCRF [9], where the topic labels $\mathbf{s}^{(k)}$ have to be marginalized out because they are not observed. Unlike HCRF, $\mathscr{L}^k(\theta)$ is concave because the first term is a linear function of θ and the second term is a log-sum-exp which is convex. The optimization is based on the quasi-newton algorithm, so we need the first-order derivatives of the log likelihood with respect to the model parameters θ. For convenience, we reformulate $\psi(y,\mathbf{s},\mathbf{x};\theta)$ as a linear function of the model parameters [5, 9], i.e.,

$$\psi(y,\mathbf{s},\mathbf{x};\theta) = \sum_{j}\sum_{l\in L^1} \theta_l^1 f_l^1(j,y,s_j,\mathbf{x}) + \sum_{(j,k)\in E}\sum_{l\in L^2} \theta_l^2 f_l^2(j,k,y,s_j,s_k,\mathbf{x}) \tag{11}$$

where θ_l^1 is the clamped parameters of θ^a and θ^s. θ_l^2 is the clamped parameters [1] of θ^e. f_l^1 and f_l^2 are the corresponding binary feature functions. The dependency of f^1 and f^2 on site index j and k is for the general formulation. In our problem, we have only one feature function for nodes and edges respectively, i.e., $|L^1| = |L^2| = 1$. We consider the derivative with respect to the node potential parameters θ_l^1 based on this formulation. For simplicity, we omit the upper index k for a specific training sample so that $(y, \mathbf{s}, \mathbf{x})$ actually refers to $(y^{(k)}, \mathbf{s}^{(k)}, \mathbf{x}^{(k)})$. It can be derived that,

$$\frac{\partial \mathscr{L}^k(\theta)}{\partial \theta_l^1} = \sum_j f_l^1(j, y, s_j, \mathbf{x}) - \sum_{y', j, a} P(y', s_j{=}a|\mathbf{x}; \theta) f_l^1(j, y', a, \mathbf{x}) \tag{12}$$

Similarly, the derivative with respect to the edge potential parameters θ_l^2 can be written as

$$\frac{\partial \mathscr{L}^k(\theta)}{\partial \theta_l^2} = \sum_{(j,k)\in E} f_l^2(j, k, y, s_j, s_k, \mathbf{x}) - \sum_{y', j, k, a, b} P(y', s_j{=}a, s_k{=}b|\mathbf{x}; \theta) f_l^2(j, k, y', a, b, \mathbf{x}) \tag{13}$$

where,

$$P(s_j{=}a, y|\mathbf{x}; \theta){=}P(s_j{=}a|y, \mathbf{x}; \theta) P(y|\mathbf{x}; \theta) \tag{14}$$

$$P(s_j{=}a, s_k{=}b, y|\mathbf{x}; \theta){=}P(s_j{=}a, s_k{=}b|y, \mathbf{x}; \theta) P(y|\mathbf{x}; \theta) \tag{15}$$

By belief-propagation (BP) [10], we can calculate the two marginals in Eq. (14) and Eq. (15). As a by-product, BP can also calculate the partition function,

$$Z(y, \mathbf{x}; \theta) = \sum_{\mathbf{s}} e^{\Psi(y, \mathbf{s}, \mathbf{x}; \theta)} \tag{16}$$

so that we can calculate the marginal $P(y|\mathbf{x}; \theta)$ as

$$P(y|\mathbf{x}; \theta) = \frac{\sum_{\mathbf{s}} e^{\Psi(y, \mathbf{s}, \mathbf{x}; \theta)}}{\sum_{y', \mathbf{s}'} e^{\Psi(y', \mathbf{s}', \mathbf{x}; \theta)}} = \frac{Z(y, \mathbf{x}; \theta)}{\sum_{y'} Z(y', \mathbf{x}; \theta)} \tag{17}$$

Given the observation $\mathbf{x}$ of a new image and the learned parameter vector $\hat{\theta}$, we infer its category label $\hat{y}$ by maximizing the posterior probability. Since predicting the class label $\hat{y}$ is our ultimate goal, we marginalize out the topic labels $\mathbf{s}$, giving out

$$\hat{y} = \arg\max_y \sum_{\mathbf{s}} P(y, \mathbf{s}|\mathbf{x}; \hat{\theta}) = \arg\max_y P(y|\mathbf{x}; \hat{\theta}) \tag{18}$$

As noted in the above section, this can be efficiently calculated by BP.

5 Experiments

5.1 Datasets

We used two well known scene image datasets for our experiments: the Oliva and Torralba [7] dataset which we referred to as the OT dataset, and the Vogel and Schiele [8] dataset,

[1] The whole set of parameter is represented by a vector and the vector again is divided into blocks. The parameters in the same block can be updated together. *Clamped* means several parameters are put in the same block, this is for the convenience of implementation.

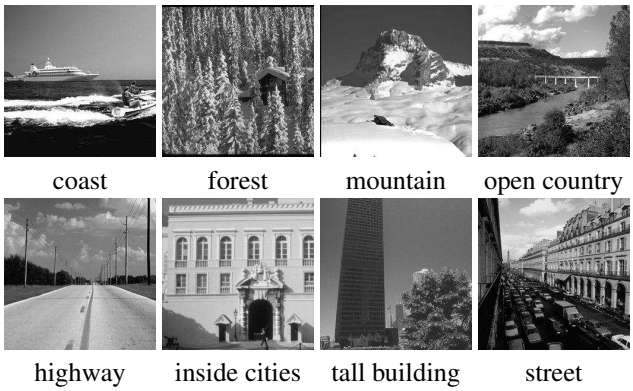


Figure 2: Sample images from the OT datasets.

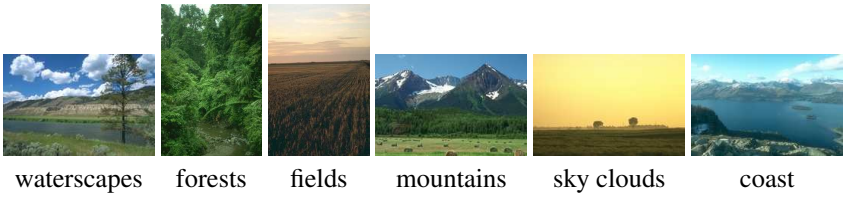


Figure 3: Sample images from the VS datasets.

referred to as the VS dataset. The OT dataset contains grayscale images of 8 scene categories. The category labels and the number of images of each category (in brackets) are: coasts (360), forest (328), mountain (374), open country (410), highway (260), inside of cities (308), tall buildings (356) and streets (292). All the images are in the same size as 250×250 pixels. The VS dataset contains 700 color images of 6 categories. The category labels and the number of images (in brackets) are: coast (142), waterscape (111), forest (103), field (131), mountain (179) and sky clouds (34). All the images in the VS dataset have been resized to 250 pixel in the maximum dimension. In Fig. 2 and Fig. 3 we show some sample images from these two datasets. Grayscale images are from the OT dataset and color images are from the VS dataset. We are aware that there are other datasets with more categories. The most complete set to our best knowledge is the 15 scene categories proposed by Lazebnik et al. [6], of which the OT dataset is only a subset. We have not chosen this one mainly because at this stage we have paid no effort on the speed of our algorithm. Working on the OT subset, we can have a more comprehensive evaluation. It is worth noting that although COCRF is computational more expensive compared to other approaches, it provides a probabilistic model to interpret the scene categories which other approaches cannot. The Bayesian approach by Fei-Fei and Perona [2] has this capability but they can not interpret the spatial layout structures of scenes.

5.2 Implementation

In our implementation, we partition each image into patches of 18×18 pixels and overlapping by 9 pixels. The number of patches of each image varies from 700 to 961. For

Table 1: Classification results in percentage on the OT and VS datasets.

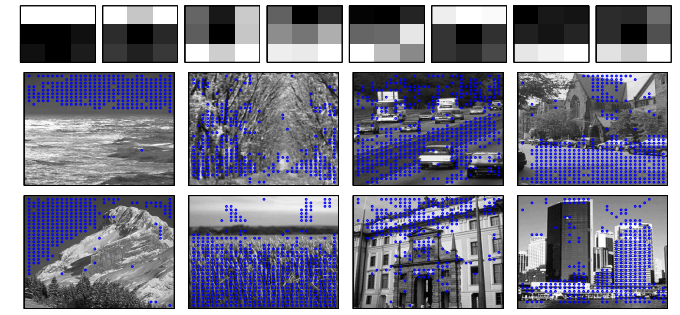
Performance on OT dataset					
Method	[1]	[6]	Task 1	Task 2	Task 3
Accuracy	86.65	86.85	82.3	87.13	90.2
Performance on VS dataset					
Method	[1]	[8]	Task 1	Task 2	Task 3
Accuracy	85.7	74.1	84.2	87.1	88.0

the grayscale images from OT dataset, we use SIFT descriptor as the feature vector for each patch. For the color images from VS dataset, we concatenated SIFT descriptor with another 6 dimensional color descriptor. The color descriptor represents the mean and variance of R,G and B. The visual vocabulary is generated by clustering a subset of 50000 image patches into 500 visual words on these two dataset respectively. PLSA is applied to group these visual words into 8 topics for both OT and OS. In generating the Gaussian components, the appearance of each topic is modeled by a mixture of 2 Gaussian components. Thus the final local appearance feature vector is a $2\times8=16$ dimensional vector. On the OT dataset, we take 100 images from each category for the training and the rest images for test (the same setup as [2] and [6]). On the VS dataset, we take half of the images from each category as training and the rest as testing (the setup as [1]). We have done several experiments including: (1) In task 1, we train COCRF with node potential but ignore the spatial location of each patch and edge potential. (2) In task 2, we train COCRF with spatial layout potential but without edge potential. (3) In task 3, we train COCRF with spatial layout potential and edge potential.

5.3 Results

Table 1 shows the classification results on the two datasets. The classification accuracy is calculated as the average of the classification accuracy of each category. In the following discussion we focus on the OT dataset. Task 1 is equivalent to take the number of occurrence of each topic in an image as the features and train a logistic classifier for image classification. Compared to the result (86.65%) in [1], our result (82.3%) in task 1 is a little worse. This is because their approach takes more training samples and trains a KNN as a non-linear classifier although the features are similar while ours is equivalent to a linear classifier. In task 2 we consider the number of occurrence of each topic and also the spatial layout of topics. This incorporation of spatial information of patches raise the recognition rate to 87.13%. It is better than that of [1] and [6] (86.65%). In [6], they also takes into account the spatial layout of each patches. Nevertheless, the result of their approach listed in Table 1 is conservative because we have taken out the classification accuracy of 8 categories from their 15 scene categories classification results. With less categories, the classification performance is expected to be slightly better. The best performance of of 90.2% is obtained in task 3. With 5 runs of task 3, each having a differnt partition of training and testing set, the deviation is 0.4%. This shows that the combination of spatial layout of individual patch and the pairwise interaction between patches is helpful for classification. The experimental results on the VS dataset shows the similar behavior.

As mentioned before, a benefit of COCRF is that it can discover the spatial layout



coast	forest	highway	street
mountain	open country	inside cities	tall building
topic=6	topic=3	topic=8	topic=5

Figure 4: Spatial distribution of topics per category. Each column illustrates two scene categories and the spatial distribution of a specific topic. The blue dots superimposed on the images illustrated the location of those image patches labeled as the corresponding topics. See text for explanation. (This figure is best viewed in color).

distribution of local patches and their pairwise interaction for a category. The ability of probabilistic modeling can not be achieved by those approaches such as those of Bosch et al. [1] and Lazebnik et al. [6]. In Fig. 4 we illustrate the learned 3×3 spatial layout distribution of different topics in some categories. In this figure, we compare the spatial layout distributions of a specific topic of two categories in each column. The first row shows the two distribution probability maps of a certain topic for the two categories. For example, in the first row and the first column, we show the spatial layout distribution of topic 6 for a coast scene in the left and that for a mountain scene in the right. The second and third rows in each column show an instantiation for each category respectively. The blue dots superimposed on the images illustrated the location of those image patches labeled as the corresponding topics. The fourth row is the text description explaining which categories and which topic are compared. It is interesting to discover that topic 6 in the moutain scene has a special distribution (mass in left top and right top part of an image) while the same topic in a coast scene is more evenly distributed in the top part of an image. In Fig. 5, we show the pairwise interaction potential map between different topics for four categories. The intensity of the cell in row i and column j represents the probability of that topic i and topic j appear as neighbors to each other. Since in scene images, it is very common that the same topic appears as neighbors, we have depressed the pairwise interaction between two same topics (diagonal cells). This is to highlight the pairwise interaction potential between different topics. From this figure we can find that different categories can have very different pattern of pairwise interaction potential between patches.

6 Conclusion

We have presented a classification oriented conditional random field (COCRF) for natural scene categorization. COCRF is adapted from HCRF and is a fully observed model for classifying a whole sequence instead of labeling each segment of a sequence. Our

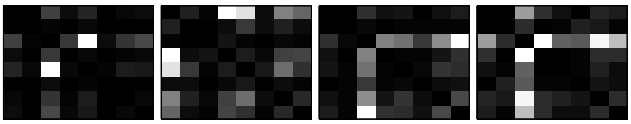


Figure 5: Illustration of the pairwise interaction potential between topics for four categories. The intensity of the cell in row i and column j represents the probability of that topic i and topic j appear as neighbors to each other.

approach is based on representing each image as an ordered set of local image patches. The training of COCRF needs both the topic labels and category labels of the training data. However, we do not need manual labeling of each segment. This is achieved by an automatic segment labeling process based on PLSA. PLSA can provide a higher level of semantic grouping of local patches by taking into account the co-occurrence relationship between different patches. COCRF provides a discriminative probabilistic model of the spatial layout of patches and their spatial pairwise interaction. Unlike HCRF, the objective function of training a COCRF model is convex, so we can avoid the concerns about local optimum and careful initialization. We have done experiments on two well-known scene image datasets. Our results demonstrate that COCRF outperforms the existing approaches for scene categorization.

References

[1] A. Bosch, A. Zisserman, and X. Munoz. Scene Classification via pLSA. In *Proceedings of the European Conference on Computer Vision*, 2006.

[2] L. Fei-Fei and P. Perona. A bayesian hierarchical model for learning natural scene categories. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2005.

[3] T. Hofmann. Unsupervised learning by probabilistic latent semantic analysis. In *Proc. of Uncertainty in Artificial Intelligence*, 1999.

[4] S. Kumar and M. Hebert. A discriminative framework for contextual interaction in classification. In *Proceedings of International Conference on Computer Vision*, 2003.

[5] J. Lafferty, A. McCallum, and F. Pereira. Conditional random fields: probabilistic models for segmenting and labelling sequence data. In *Proceedings of International Conference on Machine Learning*, 2001.

[6] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2006.

[7] A. Oliva and A. Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *International Journal of Computer Vision*, 42(3):145–175, May 2001.

[8] J. Vogel and B. Schiele. Semantic modeling of natural scenes for content-based image retrieval. *International Journal of Computer Vision*, 72(2):133–157, 2007.

[9] S. B. Wang, A. Quattoni, L.-P. Morency, and D. Demirdjian. Hidden conditional random fields for gesture recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2006.

[10] Y. Weiss and W. Freeman. On the optimality of solutions of themax-product belief propagation algorithm in arbitrary graphs. *IEEE Transactions on Information Theory*, 47(2):723–735, 2001.