2D CONDITIONAL RANDOM FIELDS FOR IMAGE CLASSIFICATION

Ming Wen, Hui Han, Lu Wang and Wenyuan Wang
Department of Automation, Tsinghua University
Beijing, 100084, P.R.China
{wenm03, hanh01, l-wang02}@mails.tsinghua.edu.cn, wwy-dau@tsinghua.edu.cn

Abstract For grid-based image classification, an image is divided into blocks, and a feature vector is formed for each block. Conventional grid-based classification algorithms suffer from inability to take into account the two-dimensional neighborhood interactions of blocks. We present a classification method based on two-dimensional Conditional Random Fields which can avoid the limitation. As a discriminative approach, the proposed method offers several advantages over generative approaches, including the ability to relax the assumption of conditional independence of the observations.

Keywords: multimedia data mining, image classification, 2D conditional random fields, loopy belief propagation

1. Introduction

Image classification is one of the most actively researched areas in multimedia data mining. Given a training set (x^k, y^k) for $k = 1, ..., K$, where x^k is the k'th image and y^k is the corresponding label, i.e. the category of the image, we would like to learn a model that maps images to labels. In general, current image classification algorithms can be divided into two groups according to the features used in classification: global approaches and component-based approaches [1]. The global approaches use global features which are usually computed with little cost. For example, Chapelle et al. [2] trained Support Vector Machines (SVMs) on color histograms to classify the images into the predefined categories. Vailaya et al. [11] extracted edge direction histograms and used Bayesian classifiers to discriminate between city and landscape images. However, global features are often unable to depict the internal structure and important details of an image. Therefore, a lot of component-based approaches have been proposed to exploit local and spatial information of the images. Fergus et al. [3] proposed a generative model to recognize object classes from unsegmented cluttered scenes. This classification system models

the appearance, shape, occlusion and relative scale of local parts extracted by an interest point detector. Smith and Li [9] proposed a method for classifying and querying images based on the spatial orderings of regions or objects using composite region templates. In the method introduced by Szummer and Picard [10], an image is partitioned into non-overlapping blocks; color and texture features are extracted for each block. The image is then classified as indoor or outdoor scenes by combining the classification results of these blocks.

Our work was intended to proceed along the same philosophical lines as Szummer and Picard's method [10] which is referred to as the grid-based method. For grid-based image classification, an image x is divided into M-by-N blocks $x = \{x_{0,0}, x_{0,1}, ..., x_{M-1,N-1}\}$, and a feature vector $\Phi(x_{i,j})$ is formed for each block $x_{i,j}$. Traditional grid-based methods don't take into account the two-dimensional neighborhood interactions of image blocks. Generative models such as Bayesian networks or Markov random fields can be used to address this problem. However, generative models have fundamental limitations. One limitation is that they require specification of the data generation process, i.e., how data can be sampled from the model [7]. In many cases, this process is unknown and not of interest for the classification task. A second limitation is that to make the model support tractable inference, one has to assume conditional independence of the observed data given the labels. Conditional random fields [5] (CRFs) are a probabilistic framework for labeling and segmenting sequential data. The conditional nature of CRFs means that no effort is wasted on specification of the data generation process and one don't need to make unwarranted independence assumptions about the observations.

In this paper, we present an image classification method based on two-dimensional conditional random fields (2D CRFs). We introduce a sequence of image block labels $h = \{h_{0,0}, h_{0,1}, ..., h_{M-1,N-1}\}$ and assume (x, h) is a CRF. Since the image blocks are two-dimensionally laid out, we specify the graphical structure of this CRF as a 2D grid, where the relative location of vertex $h_{i,j}$ is determined by the relative location of patch $x_{i,j}$ in an image x. Borrowing ideas from [8], we define a conditional probabilistic model $p(y, h|x)$ to combine the 2D CRF (x, h) and image labels y into a unified framework for image classification. Hence $p(y|x) = \sum_h p(y, h|x)$. In this model, inference and parameter estimation can be carried out using loopy belief propagation [6].

The rest of this paper is organized as follows. We introduce 2D CRFs in the next section. Section 3 describes the details of our model. In section 4, we present our experimental setup and results. Section 5 brings this paper to a conclusion. Finally, we give our acknowledgements.

2. 2D Conditional Random Fields

2.1 Standard CRFs

A conditional random field is an undirected graphical model that defines a single exponential distribution over label sequences given a particular observation sequence. Let X be a random variable over the observations to be labeled, and H be a random variable over corresponding labels. All components H_i of H are assumed to range over a finite label alphabet $\mathcal{H}$. In a discriminative framework, CRFs construct a conditional model $p(H|X)$ from paired observations and labels. Formally, we have the following definition of CRFs [5]:

DEFINITION 1 *Let $G = (V, E)$ be an undirected graph such that $H = \{H_v\}_{v \in V}$. Then (X, H) is a conditional random field if, when conditioned on X, the random variables H_v obey the Markov property with respect to the graph: $p(H_v|X, H_{V-\{v\}}) = p(H_v|X, H_{N_v})$, where $V - \{v\}$ is the set of nodes in the graph except the node v and N_v is the set of neighbors of the node v in graph G.*

Thus, a CRF is a random field globally conditioned on the observations X. In theory the structure of graph G can be arbitrary, provided it represents the conditional independencies in the models.

If the graph G is a tree (of which a chain is the simplest case), its cliques are the edges and vertices. According to the Hammersley-Clifford Theorem [4], the conditional distribution of the label sequences H given the observations X has the form:

$$p(h|x) = \frac{1}{Z(x)} \exp\{\psi(h, x; \theta)\}$$

$$Z(x) = \sum_{h} \exp\{\psi(h, x; \theta)\}$$

$$\psi(h, x; \theta) = \sum_{v \in V, l} \theta_l^1 f_l^1(v, h|_v, x) + \sum_{e \in E, l} \theta_l^2 f_l^2(e, h|_e, x)$$

where $Z(x)$ is a normalization factor known as the partition function; $h|_v$ and $h|_e$ are the components of h associated with vertex v and edge e respectively; f_l^1 and f_l^2 are feature functions and θ (including θ_l^1 and θ_l^2) are parameters to be estimated from the training data.

2.2 2D CRFs

2D CRFs are a particular case of CRFs. The graphical structure of 2D CRFs is a 2D grid (see Figure 1). Here X denotes the random variable over observations, and H denotes the random variable over the corresponding label sequences. $H_{i,j}$ is the component of H at the vertex (i, j). Apparently, the

cliques of this graph are its edges and vertices, so the conditional distribution of 2D CRFs has the same form as tree-structured CRFs. 2D CRFs can also be viewed as a finite-state model [12]. Each variable $H_{i,j}$ has a finite set of states. Out labels are associated with the states. It is possible for several states to have the same label, but in this paper we assume a one-to-one correspondence.

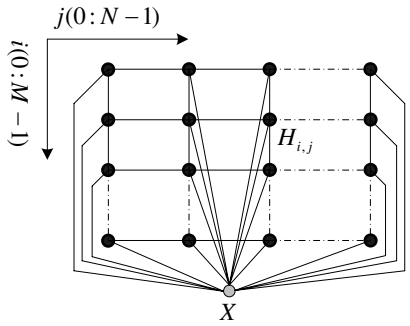

Figure 1. The graphical structure of 2D CRFs

3. Application to Image Classification

3.1 2D CRFs with Hidden Variables

Our task is to learn a model that maps images x to labels y. These labels belong to a finite image label alphabet $\mathcal{Y}$, e.g. $\mathcal{Y} = \{City, Landscape\}$.

We divide an image into M-by-N non-overlapping equal-sized blocks, and assume these image blocks can be classified into several categories, although these categories might not carry exact semantic meanings. Thus, we introduce a set of hidden variables $h_{0,0}, h_{0,1}, ..., h_{M-1,N-1}$, that correspond to block labels of an image x and form a label sequence h. Intuitively, (x, h) can be modeled with a 2D CRF. However, what we are concerned about is not block labels but image labels. Motivated by [8], we define a conditional probabilistic model:

$$p(y, h|x, \theta) = \frac{\exp\{\psi(y, h, x; \theta)\}}{\sum_{y', h} \exp\{\psi(y', h, x; \theta)\}}$$

$$\psi(y, h, x; \theta) = \sum_{v \in V, l} \theta_l^1 f_l^1(v, y, h_{i,j}, x) + \sum_{e \in E, l} \theta_l^2 f_l^2(e, y, h_{m,n}, h_{i,j}, x) \quad (1)$$

where (x, h) is a 2D CRF; $G = (V, E)$ is the graph of the 2D CRF; $h_{i,j}$ is the component of h associated with vertex v; $(h_{m,n}, h_{i,j})$ are the components of h associated with edge e; f_l^1 and f_l^2 are feature functions and θ (including θ_l^1

and θ_l^2) are the parameters of the model. It follows that

$$p(y|x,\theta) = \sum_h p(y,h|x,\theta) = \frac{\sum_h \exp\{\psi(y,h,x;\theta)\}}{\sum_{y',h} \exp\{\psi(y',h,x;\theta)\}}$$

Given the parameters θ^* estimated from the training data, a test image x will be labeled with

$$y = \arg\max_{y \in \mathcal{Y}} p(y|x,\theta^*)$$

We define ψ to take the following form as described in [8] :

$$\psi(y,h,x;\theta) = \sum_{v \in V} \phi(x_{i,j}) \cdot \theta(h_{i,j}) + \sum_{v \in V} \theta(h_{i,j},y) + \sum_{e \in E} \theta(h_{i,j},h_{m,n},y) \quad (2)$$

Here $\theta(p) \in \mathbb{R}^d$ for $p \in \mathcal{H}$ is a parameter vector corresponding the p'th block state (block label). The inner-product $\phi(x_{i,j}) \cdot \theta(p)$ can be viewed as a measure of the compatibility between block $x_{i,j}$ and state p. $\theta(p,y) \in \mathbb{R}$ for $p \in \mathcal{H}, y \in \mathcal{Y}$ can be interpreted as a measure of the compatibility between state p and image label y. $\theta(p,q,y) \in \mathbb{R}$ for $p,q \in \mathcal{H}$ and $y \in \mathcal{Y}$ measures the compatibility between the label y and an edge with states p and q. Apparently, Eq. 2 can be written in the same form as Eq. 1.

3.2 Parameter Estimation

Given the training set (x^k, y^k) for $k = 1, ..., K$, we use the following objective function in training the parameters:

$$L(\theta) = \sum_k \log p(y^k|x^k,\theta) - \frac{1}{2\sigma^2} \|\theta\|^2 \quad (3)$$

where the first item is the log-likelihood of the training data, and the second item is the logarithm of a Gaussian prior with variance σ^2, i.e. $p(\theta) \sim \exp(\frac{1}{2\sigma^2} \|\theta\|^2)$. The parameter estimation problem is to find the parameters θ^* that maximize $L(\theta)$. It is worth noting that due to the use of hidden variables, $L(\theta)$ has multiple local extrema, i.e., this method is not guaranteed to reach the global optimal point [8]. During the course of optimization, it's very important to compute the gradient of $L(\theta)$. In the rest of this section, we discuss how the gradient can be calculated efficiently. Consider the likelihood term that is contributed by the k'th training data:

$$L^k(\theta) = \log p(y^k|x^k,\theta)$$

Taking the partial derivatives of $L^k(\theta)$ with respect to the parameters θ, we have the following equations:

$$\begin{aligned}\frac{\partial L^k(\theta)}{\partial \theta_l^1} &= \sum_{v\in V,a} p(h_{i,j}=a|y^k,x^k,\theta) f_l^1(v,y^k,a,x^k) \\ &\quad - \sum_{y',v\in V,a} p(h_{i,j}=a,y'|x^k,\theta) f_l^1(v,y',a,x^k)\end{aligned}$$

$$\begin{aligned}\frac{\partial L^k(\theta)}{\partial \theta_l^2} &= \sum_{e\in E,a,b} p(h_{i,j}=a,h_{m,n}=b|y^k,x^k,\theta) f_l^2(e,y^k,a,b,x^k) \\ &\quad - \sum_{y',e\in E,a,b} p(h_{i,j}=a,h_{m,n}=b,y'|x^k,\theta) f_l^2(e,y',a,b,x^k)\end{aligned}$$

It is obvious that $\partial L^k(\theta)/\partial\theta_l^1$ can be expressed in terms of components $p(h_{i,j}=a|y,x^k,\theta)$ and $p(y|x^k,\theta)$, which can be approximately calculated using loopy belief propagation, for 2D grid contains cycles. Similarly, $\partial L^k(\theta)/\partial\theta_l^2$ can also be expressed in terms of expressions which can be approximately calculated using loopy belief propagation.

4. Experiments

4.1 Experimental Setup

We carried out three sets of experiments to distinguish car from background, city from landscape, and indoor scene from outdoor scene.

The image data set consists of 600 Corel images. All the images are in JPEG format of size 384×256 or 256×384. As a result of the tradeoff between cost and accuracy, every image is partitioned into 8-by-8 blocks, and a feature vector is formed for each block. The feature vector is made up of color histogram (CH), edge direction histogram (EDH), texture statistics based on Gabor filters and Discrete Cosine Transform (DCT) coefficients. CH is obtained by quantizing each component of the RGB color space into 16 bins. For the shape feature, EDH is selected. Sobel edge detector is applied to obtain the edge images. The computed EDH from the edge image of each block is quantized into 36 bins. To calculate the texture feature, we first apply a set of 2D Gabor filters to the blocks, and then calculate the means and standard deviations of the transformation coefficients. The filter bank is created with 5 orientations ($0°, 30°, 60°, 90°, 135°$) and 6 frequencies (0, 2, 4, 8, 16, 32). The DCT transform is performed on each block, and the 16 coefficients from the uppermost left 4-by-4 matrix are taken as features, representing the energy in the lower frequencies. Hence, a feature vector of 160 dimensions is formed

for each block. Finally, Principal Components Analysis (PCA) is applied to reduce the feature vector of each block to 2 dimensions.

In our experiments, images within each category were randomly partitioned in half to form a training set and a test set, and five-state models were trained. We repeated each experiment for 5 random splits, and reported the average of the results obtained over 5 different test sets. The parameter σ^2 of the Gaussian prior in 2D CRFs was selected according to a two-fold cross-validation on the training set.

4.2 Experimental Results

To provide a more objective evaluation, we compared our method with a SVM based method. Different from the method in [2], the SVM based method we used in our experiments is a grid-based method. It packs the block feature vectors of an image into a single feature vector.

The average classification accuracies are presented in Table 1. From the results shown in Table 1, we can see that the proposed 2D CRFs based method performs better than the SVM based method. Though the SVM based method extracts local features and partially considers the spatial information of image blocks, it loses sight of the fact that the blocks are two-dimensionally laid out.

Table 1. Comparison of the classification results for different methods

Experiment	2D CRFs	SVM
Car vs. Background	88.8%	87.0%
City vs. Landscape	89.0%	86.2%
Indoor scene vs. Outdoor scene	90.6%	86.4%

5. Conclusions and Future Work

In this paper, we have presented a novel probabilistic model for grid-based image classification. Based on two-dimensional conditional random fields, the model not only takes into account the spatial information of image blocks, but also incorporates the two-dimensional neighborhood interactions of blocks. Experimental results show our method outperforms the SVM based classification method. In the future, we'll try to deduce the forward-backward vectors of this model for efficient inference.

Acknowledgments

We thank Chong Wang for many helpful discussions. We also thank Kevin Murphy very much for publishing the Conditional Random Field Toolbox for Matlab in the Web so that we can develop our experimental codes efficiently.

References

[1] Bi, J. and Chen, Y. (2005). A sparse support vector machine approach to region-based image categorization. In *Proc. of CVPR*.

[2] Chapelle, O., Haffner, P., and Vapnik, V. (1999). Support vector machines for histogram-based image classification. *IEEE Trans. on Neural Networks*, 10(5):1055–1064.

[3] Fergus, R., Perona, P., and Zisserman, A. (2003). Object class recognition by unsupervised scale-invariant learning. In *Proc. of CVPR*, volume 2, pages 264–271.

[4] Hammersley, J. and Clifford, P. (1971). Markov fields on finite graphs and lattices. Unpublished Manuscript.

[5] Lafferty, J., McCallum, A., and Pereira, F. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proc. of ICML*.

[6] Murphy, K. P., Weiss, Y., and Jordan, M. I. (1999). Loopy belief propagation for approximate inference: An empirical study. In *Proc. of UAI*.

[7] Qi, Y., Szummer, M., and Minka, T. P. (2005). Bayesian conditional random fields. In *Proc. of the Tenth International Workshop on Artificial Intelligence and Statistics*.

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

[9] Smith, J. R. and Li, C.-S. (1999). Image classification and querying using composite region templates. *Journal of Computer Vision and Image Understanding*, 75(1/2):165–174.

[10] Szummer, M. and Picard, R. W. (1998). Indoor-outdoor image classification. In *Proc. of IEEE International Workshop on Content-Based Access of Image and Video Databases*, pages 42–51.

[11] Vailaya, A., Figueiredo, M. A. T., Jain, A. K., and Zhang, H.-J. (2001). Image classification for content-based indexing. *IEEE Trans. on Image Processing*, 10(1):117–130.

[12] Zhu, J., Nie, Z., Wen, J.-R., Zhang, B., and Ma, W.-Y. (2005). 2d conditional random fields for web information extraction. In *Proc. of ICML*.