Proceedings of the Eighth International Conference on Machine Learning and Cybernetics, Baoding, 12-15 July 2009

STATISTICAL PART-BASED MODELS FOR OBJECT CATEGORY RECOGNITION

XIAO-ZHEN XIA[1], SHU-WU ZHANG[1]

[1]Digital Content Technology Research Center, Institute of Automation, Chinese Academy of Sciences, Beijing, China
E-MAIL: xiaxiaozhen@hitic.ia.ac.cn, swzhang@hitic.ia.ac.cn

Abstract:

In this paper, we present a new method to learn statistical part-based structure models for object category recognition in a supervised manner. The method learns both a model of local part appearance and a model of the spatial relations between those parts. By using Histograms of Oriented Gradient (HOG) features to describe local part appearance within an image, we investigate whether richer appearance model is helpful in improving recognition performance. We learn the model parameters from training examples using maximum likelihood estimation. In detection, these models are used in a probabilistic way to classify and localize the objects in the images. The experimental results on a variety of categories demonstrate that our method provides both successful classification and localization of the object within the image.

Keywords:

Object categorization; Part-based recognition; Statistical models; HOG descriptor

1. Introduction

Techniques for representing, learning and recognizing object categories have attracted much attention in computer vision and artificial intelligence. There has been a progression from recognizing individual objects to classes of objects which are visually similar. But it still remains a challenging problem because of the large variance of objects in the same class. Such variance may be due to the view point variation, illumination and occlusion, change in the scale, background clutter, and deformable object shape, etc.

A great deal of algorithms have been proposed for object classification. One approach to classifying images is to treat them as a collection of regions. Similar models have been successfully used in the text community for analyzing documents and are known as "bag-of-words" models. Fei-Fei et al. [1] have applied such methods to the visual domain. However, "bag-of-words" models describe only part appearance and ignoring their spatial structure. Burl et al. [3] propose a statistical model in which "soft" part detectors is used and shape variability is modeled in a probabilistic setting. Weber et al. [4] introduce a particular type of model where objects are represented as flexible constellations of rigid parts. The variability within a class is represented by a joint probability density function on the shape of the constellation and the output of part detectors. Furgus et al. [5] have extended the work of Weber and added variability of appearance, relative scale to learn this model in a scale invariant manner. The drawback of the Constellation model is an exponential explosion in computational cost and being highly dependent on the consistent firing of the interest operator [6].

The pictorial structure models introduced by Fischler and Elschlager [7] represent an object by a collection of parts arranged in a deformable configuration, where the deformable configuration is represented by spring-like connections between pairs of parts. Crandall et al. propose k-fans model [9, 10] to study the extent to which additional spatial constraints among parts are actually helpful in detection and localization. It is proved that adding spatial constraints gives better performance but increases computational complexity.

In this paper, we consider the problem of detecting and localizing objects of various categories such as airplanes or faces within the images. Emphasizing on spatial relations between parts, we use tree-structured model similar to 1-fan model [9, 12] or star model [6]. The model is trained in a probabilistic way that requires labeled parts for the positive images. In particular, we describe local part appearance by using HOG [11] features which is different from [9] to investigate whether richer appearance model is helpful in improving recognition performance.

The rest of the paper is organized as follows. In section 2, the detailed procedure of model learning and object recognition are described. Section 3 presents the experimental results on three classes and the background from Caltech-101 database [2]. Section 4 gives the conclusions and future work.

978-1-4244-3703-0/09/$25.00 ©2009 IEEE

2. Approach

In this section, we first describe the local part appearance model and spatial relations model. Then HOG features are represented as local part appearance within an image and principle component analysis (PCA) is used to reduce the dimension of HOG feature vector. Finally, the procedure of learning the models, detection and localization are presented in detail.

2.1. Model structure

We use the model framework in [8, 9] where an object model $\theta = (A, S)$ consists of appearance A for each part and spatial relationships S between parts. Let I denote an image, and $L = \{l_1, ..., l_n\}$ denote the location for each part. Here the location of a part is given by a point in the image, $l_i = (x_i, y_i)$. Using Bayes' law the posterior distribution $p(L \mid I, \theta)$, which characterizes the probability that the object configuration is L given the model θ and the image I, can be written as,

$$p(L \mid I, \theta) \propto p(I \mid L, \theta) p(L \mid \theta) \tag{1}$$

where the distribution $p(I \mid L, \theta)$ measures the likelihood of seeing image I given a particular configuration for the object parts and the model θ. The distribution $p(L \mid \theta)$ measures the prior probability that the object configuration is L.

2.1.1. Appearance

Let $A = \{a_1, ..., a_n\}$ be appearance parameters for each part, and suppose the part positions are independent, then we model the appearance by the product of the individual likelihoods,

$$p(I \mid L, \theta) = p(I \mid L, A) \propto \prod_{i=1}^{n} p(I \mid l_i, a_i) \tag{2}$$

This is an approximation when parts overlap. To model the appearance of each individual part we use the HOG representation introduced in [11]. Let $F = \{f_1, ..., f_n\}$ be HOG features extracted from each part, and consider the distribution of HOG features of each part as a Gaussian. Under the Gaussian model, the appearance parameters for each part are $a_i = (\mu_i, \Sigma_i)$. We have,

$$p(I \mid l_i, a_i) = N(f_i, \mu_i, \Sigma_i) \tag{3}$$

2.1.2. Spatial relations

In tree-structured graph, let $G = (V, E)$ be a tree with a root node v_r and other independent nodes $v_i (i \neq r)$ conditioned on the value of v_r. Let $S = \{s_1, ..., s_n\}$ be parameters of spatial relationships, l_r be the location of the root part and l_i be the location of other part except for the root part, then we have,

$$p(L \mid \theta) = p(l_r \mid s_r) \prod_{v_i \neq v_r} p(l_i \mid l_r, s_i) \tag{4}$$

We model the conditional distribution of other part location given the root part location $p(l_i \mid l_r, s_i)$ as a Gaussian with mean $\mu_{i|r}$ and covariance $\Sigma_{i|r}$,

$$p(l_i \mid l_r, s_i) = N(l_i - l_r, u_{i|r}, \Sigma_{i|r}) \tag{5}$$

2.2. HOG representation

We use HOG to describe features for each part within an image. The image is first divided into 8*8 non-overlapping pixel cells. For each cell a 1D histogram of gradient orientations over pixels is accumulated, with nine orientation bins. For better invariance to illumination, shadowing and other small deformations, the histogram of each cell is normalized with respect to the gradient energy over 2*2 cells (called a block) around it. This leads to a vector of length 9*4 representing the local gradient information inside a block.

We extract fix-sized patches of the 3*3 non-overlapping blocks surrounding the labeled part location. Then the feature vector of each patch with 48*48 pixel size exists in a 36*9 dimensional space. We use PCA to reduce the dimensionality of the feature vector of each patch, in order to reduce the memory requirements and computational cost. In the learning stage, we collect the feature vector of each patch from all images and perform PCA on them. Let k be the number of dimensions in the dimensionally reduced subspace. We choose the smallest value of k so that the cumulative energy is above 90%. Then the feature vector of each patch is projected into the space spanned by the first k principal components.

2.3. Learning

The task of learning is to estimate the parameters $\theta = (A, S)$ of the model discussed above. The goal is to find the maximum likelihood (ML) estimate θ^* which best explain the data A and S from all the training images. We are given a set of images $\{I_1, ..., I_m\}$ and corresponding object configurations $\{L_1, ..., L_m\}$ for each image, the ML estimate of θ is,

$$\theta^* = \arg\max_{\theta} \prod_{k=1}^{m} p(I_k \mid L_k, \theta) \prod_{k=1}^{m} p(L_k \mid \theta) \quad (6)$$

From equation (2), (4) and (6) we get ,

$$A^* = \arg\max_{A} \prod_{k=1}^{m} p(I_k \mid L_k, A)$$

$$= \arg\max_{A} \prod_{k=1}^{m} \prod_{i=1}^{n} p(I_k \mid l_{k,i}, a_i) \quad (7)$$

$$S^* = \arg\max_{S} \prod_{k=1}^{m} p(L_k \mid S)$$

$$= \arg\max_{s} \prod_{k=1}^{m} p(l_{k,r} \mid s_r) \prod_{v_i \neq v_r} p(l_{k,i} \mid l_{k,r}, s_i) \quad (8)$$

2.4. Detection and Localization

The detection problem is to determine whether the image I contains an instance of the object (hypothesis w_1) or whether the image is background-only (hypothesis w_0). The decision is given by the likelihood ratio,

$$R = \frac{p(I \mid w_1)}{p(I \mid w_0)} = \frac{\sum_{L} \prod_{i=1}^{n} p(I \mid l_i, a_i) \cdot p(L \mid \theta)}{p(I \mid w_0)}$$

$$= \sum_{L} \frac{N(f_r; \mu_r, \Sigma_r)}{N(\mu_{bg}, \Sigma_{bg})} p(l_r \mid s_r) \prod_{v_i \neq v_r} \frac{N(f_i; \mu_i, \Sigma_i)}{N(\mu_{bg}, \Sigma_{bg})} p(l_i \mid l_r, s_i) \quad (9)$$

where the numerator can be expressed as a sum over all possible object configurations L, and (μ_{bg}, Σ_{bg}) are the background model parameters. By calculating the likelihood ratio R and comparing it to a threshold, the presence or absence of the object within the image can be determined.

For getting the location of the object within each image, we look for an object configuration L^* with maximum posterior probability. From equation (2) and (4) we get,

$$L^* = \max_{L} \frac{N(f_r; \mu_r, \Sigma_r)}{N(\mu_{bg}, \Sigma_{bg})} p(l_r \mid s_r) \cdot \prod_{v_i \neq v_r} \frac{N(f_i; \mu_i, \Sigma_i)}{N(\mu_{bg}, \Sigma_{bg})} p(l_i \mid l_r, s_i) \quad (10)$$

3. Experiments

We evaluated our system using the Caltech-101 database [2]. Five categories were selected from the databases, which were airplanes, motorbikes, faces, leopards and cars. Background images were also selected from them. The images were scaled so that they were uniform. Figure 1 shows the images from the five categories, each of which with six parts annotated, and background image.

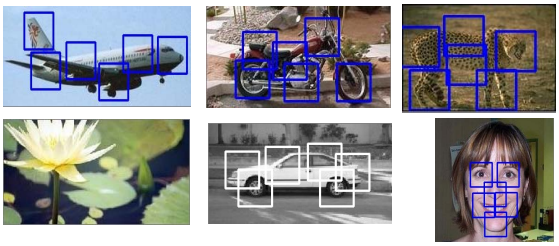

Figure 1. Six categories and background

Each dataset was split randomly into two separate sets. We used the first for training and the remaining for testing. Among the datasets, the car dataset was only 123 images originally, so another 77 were added by reflecting the original images, making 200 in total. The numbers of samples used for training and testing are shown in Table 1.

Table 1. Datasets in experiments

	Training data		Testing data	
	positive	negative	positive	negative
Planes	400	400	400	400
Bikes	400	400	400	400
Faces	200	200	200	200
Leopards	100	100	100	100
Cars	100	100	100	100

3.1. Learning

As in [9], we used supervised training method to learn the models. Six parts were labeled by hand in each training

images. To learn the appearance model for a given part, patches of 3*3 blocks surrounding the labeled part location were extracted from each training image. Each patch then formed a 36*9 dimensional feature vector. We performed PCA on feature vectors from all images (for motorbikes $k = 50$) , and then got mean and covariance parameters $a_i = (\mu_i, \Sigma_i)$ for each part. The background model parameters (μ_{bg}, Σ_{bg}) were estimated from the background images. For spatial relations model we compute the conditional distribution parameters $s_{i(i \neq r)} = (\mu_{i|r}, \Sigma_{i|r})$ for each part. In addition, we used the same training images and annotated file to learn 1-fan model, same testing images to evaluate the performance of 1-fan model in order to compare the results with ours.

3.2. Experimental Results

For detection, we used the procedure described in Section 2.4 to compute the likelihood ratio R and find an optimal configuration for the object in each test image. Figure 2 illustrates the ROC curves for airplanes and motorbikes generated from the experiments. We observe that our model performs slightly better, indicating that using HOG descriptors to model the appearance gives better performance for airplanes and motorbikes.

Table 2 shows the results of recognition accuracy at the equal ROC points by comparing the algorithm to 1-fan method. It can be seen that in the cases of motorbikes and airplanes dataset, the performance of the algorithm is superior to the 1-fan model, but is inferior to 1-fan model in the case of faces dataset.

Table 2. Equal ROC performance in experiments

	Ours	1-fan
Planes	93.25%	91.3%
Bikes	97.6%	97.0%
Faces	96.8%	98.2%
Leopards	92.7%	--
Cars	94.3%	--

As in [9], in order to test the ability of the models to differentiate between the three different object classes (airplanes, motorbikes and faces) and the background images, we conducted multi-class detection experiments. The results of our method and 1-fan model are illustrated in Table 3, where rows correspond to actual classes and columns correspond to predicted classes.

4. Conclusions

We have introduced a probabilistic method of learning part-based models for object category recognition. Combining HOG descriptors, we learn the models by estimating both part appearance and spatial relations between parts. Specifically, we investigate whether richer appearance model is helpful in improving recognition. Experimental results on a variety of categories demonstrate the power of our system in object detection. Currently, the framework does not use the context analysis, where the presence of a certain object class in an image probabilistically influences the presence of a second class. We are working on integrating scene context to improve detection.

Acknowledgements

The paper is supported by National Sciences & Technology Support Program of China (Grant No. 2008BAH26B02-3, Grant No. 2008BAH21B03-04 and Grant No. 2008BAH26B03)

References

[1] L. Fei-Fei, R. Fergus, and P. Perona, “Learning Generative Visual Models from Few Training Examples: An Incremental Bayesian Approach Tested on 101 Object Categories”, CVPR Workshop on Generative-Model Based Vision, 2004.

[2] L. Fei-Fei, R. Fergus, and P. Perona, “A Bayesian Hierarchical Model for Learning Natural Scene Category”, CVPR, 2005.

[3] M.C. Burl, M. Weber, and P. Perona, “A Probabilistic Approach to Object Recognition Using Local Photometry and Global Geometry”, ECCV, pp. 628-641, 1998.

[4] M. Weber, M. Welling, and P. Perona, “Unsupervised Learning of Models for Recognition”, ECCV, 2000.

[5] R. Fergus, P. Perona, and A. Zisserman, “Object Class Recognition by Unsupervised Scale-Invariant Learning”, CVPR, 2003.

[6] R. Fergus, P. Perona, and A. Zisserman, “A Sparse Object Category Model for Efficient Learning and Exhaustive Recognition”, CVPR, 2005.

[7] M.A. Fischler and R.A. Elschlager, “The representation and matching of pictorial structures”, IEEE Transactions on Computer, 1973.

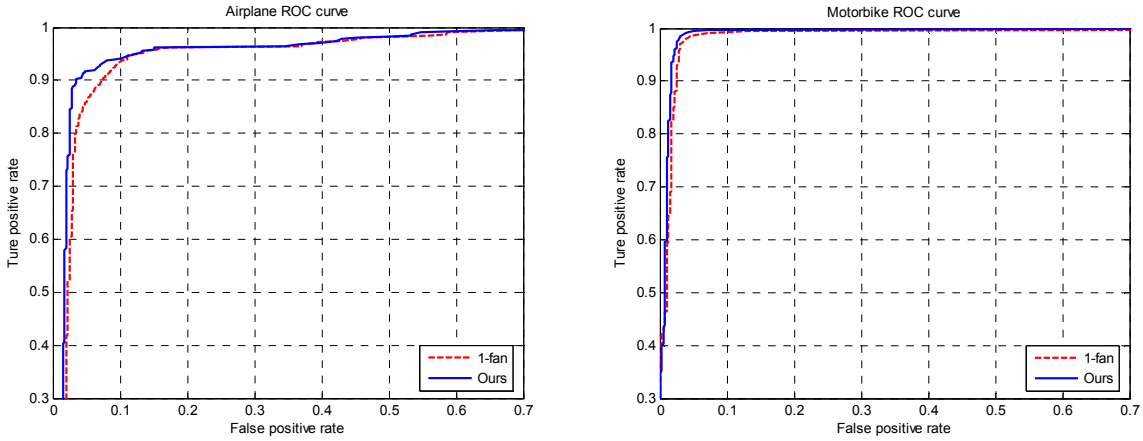

Figure 2. ROC curves for motorbikes and airplanes

Table 3. Multi-class detection experiments

	Ours				1-fan			
	Planes	Bikes	Faces	BG	Planes	Bikes	Faces	BG
Planes	371	7	0	22	362	5	0	33
Bikes	4	386	0	10	4	384	0	12
Faces	3	8	189	0	2	7	191	0
BG	27	7	0	366	36	11	0	353

[8] P.F. Felzenszwalb and D.P. Huttenlocher, "Pictorial Structures for Object Recognition", IJCV, 2005.

[9] D.J. Crandall, P.F. Felzenszwalb, and D.P. Huttenlocher, "Spatial Priors for Part-Based Recognition using Statistical Models", CVPR, pp. 10-17, 2005.

[10] D.J. Crandall and D.P. Huttenlocher, "Weakly Supervised Learning of Part-Based Spatial Models for Visual Object Recognition," ECCV, pp. 16-29, 2006.

[11] N. Dalal and B. Triggs, "Histograms of Oriented Gradients for Human Detection," CVPR, pp. 886-893, 2005.

[12] P. Felzenszwalb, D. McAllester, and D. Ramanan, "A Discriminatively Trained, Multiscale, Deformable Part Model," CVPR, 2008