

Weakly Supervised Learning of Component-Based Hierarchical Model for Object Detection

Xiaozhen Xia¹, Wuyi Yang², Wei Liang¹, Shuwu Zhang¹

¹Hi-tech Innovation Center, Institute of Automation, Chinese Academy of Sciences, Beijing, China

²Key Laboratory of Underwater Acoustic Communication and Marine Information Technology, Xiamen University, China

Abstract—In this paper, we present a hierarchical framework for detecting and localizing object by components. The system is structured with a root detector and several component detectors that are trained to separately find the object and different parts of the object on the first level. On the second level the spatial relations model performs detection by combining the root detector and the component detectors. We learn the component models in a weakly supervised manner, where object labels are provided but component labels are not. The root model and each component model are learned by using boosting. The weak classifiers are vector-valued HOG features which are projected from d-dimensional to 1-dimensional subspace by Fischer Linear Discriminant (FLD). The experimental results demonstrate that our method is comparable with the previous ones.

Index Terms— object detection, boosting, component-based hierarchical models

I. INTRODUCTION

The detection of objects in images has been one of the most challenging problem in computer vision. The main difficulty in developing a reliable object detection approach arises from view point variation, illumination and occlusion, change in the scale, background clutter, and deformable object shape. Recently, there has been a tremendous improvement in object classification performance. However, detection has not reached such level of performance for the greater difficulty of the task, though there have been notable improvements for several classes.

Recent works have demonstrated that boosted classifiers can be effective to detect object. Viola and Jones [2] construct a boosted cascade of simple feature classifiers for rapid face detection. Laptev [10] combines histogram-based image descriptors with a boosting classifier to provide a state of the art object detector. Opelt et al. [14] use boosting to select boundary fragments as weak detectors and combine them to form a strong “Boundary-Fragment-Model” detector for object detection. Dollar et al. [13] automatically learn individual component classifiers and combine them into an overall classifier. The boosted classifiers have proven to be powerful in detection, however, the selection of the most discriminative features in the training process is still a difficult problem.

The way that objects can be represented by a collection of parts arranged in a deformable configuration achieves good

performance in object detection [3, 5, 6, 7, 8]. While these models offer an intellectually satisfying way of representing the objects, it remains complex and computationally intensive to establish their value in practice. There are many challenges for part-based models. The first challenge is how to learn a set of discriminatory object parts. Instead of manually selecting the parts, it is desirable to learn the parts automatically from a set of examples based on their discriminative power and their robustness against pose and illumination changes. The second challenge is to combine each part by learning their geometrical configuration.

In this paper, we develop a new component-based hierarchical model for object detection. Our work is motivated by [10] and [12]. Different from [10], we add component classifiers and spatial relations between the root and the components. The difference between [12] and our work lies in the way to learn the deformable part model. We learn each component classifier in a weakly supervised manner. GentleBoost [4] is applied to select weak classifiers and combine them to form a strong component classifier. The weak classifiers are vector-valued HOG [1] features which are projected from d-dimensional to 1-dimensional subspace by FLD. In detection, the root classifier and component classifiers independently detect the parts of the object. Then the spatial relations model combines the root and each component classifiers for detecting the object.

The rest of this paper is organized as follows. In section 2, we describe the model framework. In section 3, we provide a detailed description of learning method for the proposed model. Section 4 gives the procedures of detection and localization. The experimental results are presented in section 5. Finally, we conclude the paper in section 6.

II. MODEL

An overview of our two-level component-based model is shown in Figure 1. The model consists of a root model and several component models on the first level, as well as spatial relations model on the second level. The root model and component models are trained using boosting. The weak classifiers are vector-valued HOG features which are projected from multi-dimensional to 1-dimensional subspace by FLD. For each part, HOG features are extracted from all possible blocks that vary in sizes, locations and aspect ratio. The spatial relations between the root and each component are equivalent

The research was supported by National Sciences & Technology Support Program of China (Grant No. 2008BAH21B03, Grant No. 2008BAH26B02, and Grant No. 2008BAH26B03).

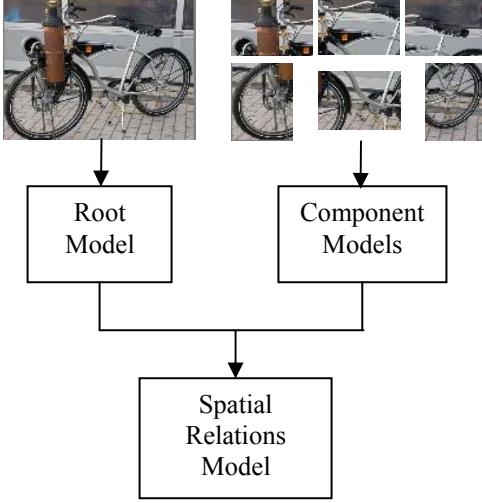


Figure 1. Two-level component-based models

to the star graph [5] with a root part and other parts conditioned on the root part.

A. Feature

We use HOG to describe features for each $w \times h$ component within an image, and integral histogram [9] to efficiently compute HOG features over arbitrary rectangular image regions. The gradient at each pixel is discretized into one of 9 orientation bins, and an integral image for each bin of the HOG is computed and stored. Then the HOG features for any rectangular regions can be computed quickly from those integral images. To encode more information for each component, we consider blocks whose size ranges from 8×8 to $w \times h$ in the component. The ratio between block width and block height can be any of the following ratios (1 : 1), (1 : 2) and (2 : 1). Each block contains a 36-dimensional histogram vector of concatenating the 9 orientation bins in 2×2 sub-blocks. In our system each feature corresponds to the 36-dimensional vector used to describe a block.

B. Weak Classifiers

We choose to base our algorithm on GentleBoost [4] which has been shown experimentally [15] to outperform other boosting algorithms. The weak learners are vector-valued HOG features which are projected from multi-dimensional to 1-dimensional subspace. Like [9], we use FLD as a weak learner to minimize a weighted classification error as required by GentleBoost. FLD is designed to find an optimal direction of projection to separate the positive and negative samples. Given the image features f , the projection function is defined as,

$$g = w^T f, \quad (1)$$

where $w = (S_1 + S_2)^{-1}(u_1 - u_2)$, u_1 , u_2 are the means of the two classes, and S_1 , S_2 are the covariance matrices.

C. Spatial Relations

We use the star-graph model proposed in [5] to model the spatial relations between the parts. In our system, the root part

corresponds to the object and the non-root part corresponds to each component. Let $G = (V, E)$ be a star graph with root part v_r and other independent parts $v_i (i \neq r)$ conditioned on the root part v_r . Let $S = \{s_1, \dots, s_n\}$ be the parameters of spatial model. The location of the object within an image is defined by a configuration of its parts $L = \{l_1, \dots, l_n\}$, where l_r is the location of the root part and $l_{i(i \neq r)}$ is the location of non-root part. The spatial relations can be written in terms of conditional distributions as,

$$p(L | S) = p(l_r | s_r) \prod_{v_i \neq v_r} p(l_i | l_r, s_i). \quad (2)$$

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

$$p(l_i | l_r, s_i) = N(l_i - l_r, \mu_{i|r}, \Sigma_{i|r}). \quad (3)$$

III. LEARNING

In learning we are given a set of images annotated with bounding boxes around each instance of an object. The root classifier is trained first, and then a set of discriminative object components are selected automatically by the feature selection process. Once the components have been determined, each component classifier is trained by the corresponding selected component. Finally, the spatial relations classifier is trained by the location of the object and each component.

A. Training Root Classifier

For each category, we automatically select the dimensions of the root part by looking at the statistics of the bounding boxes in the training data. We train an initial root classifier using cascade of boosted classifiers. The positive examples are constructed from the unoccluded training examples. We use random subwindows from negative images to generate negative examples. After we train the first stage for the boosted cascade, we obtain new negative training samples for subsequent training stage by collecting false positive detections from training images. The first row of Fig. 2 shows the top six discriminative blocks that are chosen with minimum error rates.

B. Training Component Classifiers

We select the most discriminative blocks as our initial components from the root classifier trained above. We train the initial classifiers for each component respectively, then for each positive bounding box in the training data, we apply each component detector at all positions within it and select the highest scoring placement as new positive component. Negative samples are selected by finding high scoring detections in images not containing the target component. A new model for each component is trained by corresponding updated component on the positive and negative examples. We update the model 4 times using the way described above. The second row of Fig. 2 shows the components of the bicycle class and some of the selected discriminative blocks for each component.

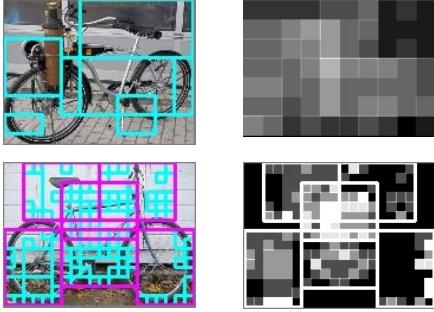


Figure 2. Selected blocks for the bicycle class as well as root and component models. top-left: regions of the top six discriminative blocks that are chosen with minimum error rates; top-right: selected features superimposed using transparent color; bottom-left: selected discriminative blocks for each component; bottom-right: our component models.

C. Training Spatial Relations

We use the trained component classifiers to locate each component within the labeled object and get the position of each component within all the training images. For a set of images $\{I_1, \dots, I_m\}$ with bounding box on the object, we get the object configuration $\{L_1, \dots, L_m\}$, $L_k = \{l_{k,1}, \dots, l_{k,n}\}$. We learn the spatial relations model from geometrical configurations using a maximum likelihood (ML) criterion. The goal is to find the ML estimate S^* which best explain the data from all the training images. The ML estimate of the model parameters S is,

$$\begin{aligned} S^* &= \arg \max_S \prod_{k=1}^m p(L_k | S) \\ &= \arg \max_S \prod_{k=1}^m p(l_{k,r} | s_r) \prod_{v_i \neq v_r} p(l_{k,i} | l_{k,r}, s_i). \end{aligned} \quad (4)$$

Given the set of detected object component configurations, the estimated parameters involve the means and covariances in (3). These can be obtained from the sample mean and covariance of the detected component configurations.

IV. DETECTION AND LOCALIZATION

In detection, the root classifier and component classifiers independently detect the object and the components of it first. The spatial relations model then combines the root and component classifiers for detecting and localizing the object. We use the standard window scanning technique and apply the classifier to the large number of image sub-windows with densely sampled positions and sizes. The root and component classifiers scan across the image at multiple scales and compute the boosting score within all sliding windows, and then localize the object within the image by integrating the spatial relations model. Let $L = \{l_1, \dots, l_n\}$ denote the location of each part, $c_i(l_i)$ denote the score of the presence of the i^{th} part in the candidate windows given the location l_i . $c_i(l_i)$ is derived

from the boosted cascade detector. For localization, we look for an object configuration L^* with maximum probability,

$$L^* = \max_L p(l_r | s_r) c_r(l_r) \prod_{v_i \neq v_r} p(l_i | l_r, s_i) c_i(l_i).$$

There is a large number of placements for the parts of the object. We use distance transforms technique [6] to compute the best location for the parts of the object as a function of the root part location. We score root part locations according to the best possible placement of the parts and threshold this score. In this case the running time of the localization algorithm is reduced to $O(nk)$, where n is the number of parts and k is the number of detection windows (number of locations) for each part within the image.

V. EXPERIMENTS

We evaluated the proposed method on PASCAL VOC 2007 challenge [11] on the task of detecting object classes. The VOC 2007 dataset contains 5011 training and validation images, manually annotated with bounding boxes for 20 image classes, of which we select motorbike, bicycle, people, bird, horse and cow classes. The training and the test sets contain substantial variation of objects in terms of scale, pose occlusion and within-class variability. A detection is considered correct when it overlaps more than 50% with a grounding-truth bounding box. One scores a system by the average precision (AP) of its precision-recall curve across a test set.

Our root model was learned by the labeled object in the images. We trained a cascade of boosted classifier for the root model. We required the minimum detection rate to be 0.995 and the maximum false positive to be 0.5 for each stage. The number of weak classifiers in the first four layers of the detector was 3, 5, 8, 10 respectively. We located the components of the object automatically by the most informative blocks selected in the training process of root model. Then each component model was learned and updated by selecting the highest scoring placement according to previous trained component model. Finally, the spatial relations model was trained by the locations of the object and the components. The training process took 5 to 6 hours using a PC with 2.66GHz CPU and 1GB memory.

We investigated the effect of combining the root model and each component model on the six datasets. Table 1 shows the detection results of our model for the six categories, which are comparable with the ones that entered the competition [11]. Fig. 3 illustrates some detection and localization results of our algorithm for the six categories, showing precise localization of the parts despite substantial variability in their appearances and locations. We also compared our method to the way of using root model without component information. The results in table 1 show that our model provided a substantial improvement in accuracy for all but the bird class, which implying that combining root model and component models is effective in object detection. We think the lack of improvement for bird is due to the difficulty of learning different components of the bird class.

TABLE I. AVERAGE PRECISION SCORES ON THE 6 CATEGORIES

Method	Bicycle	Motorbike	Person	Bird	Horse	Cow
INRIA Normal	0.246	0.153	0.121	0.012	0.225	0.039
INRIA Plus	0.287	0.249	0.092	0.041	0.335	0.127
MPI Center	0.110	0.170	0.091	0.028	0.198	0.044
MPI ESSOL	0.157	0.208	0.117	0.098	0.034	0.061
TKK	0.078	0.135	0.061	0.043	0.186	0.100
Ours(root)	0.139	0.188	0.096	0.130	0.124	0.076
Ours(root+ components)	0.144	0.191	0.098	0.125	0.138	0.082

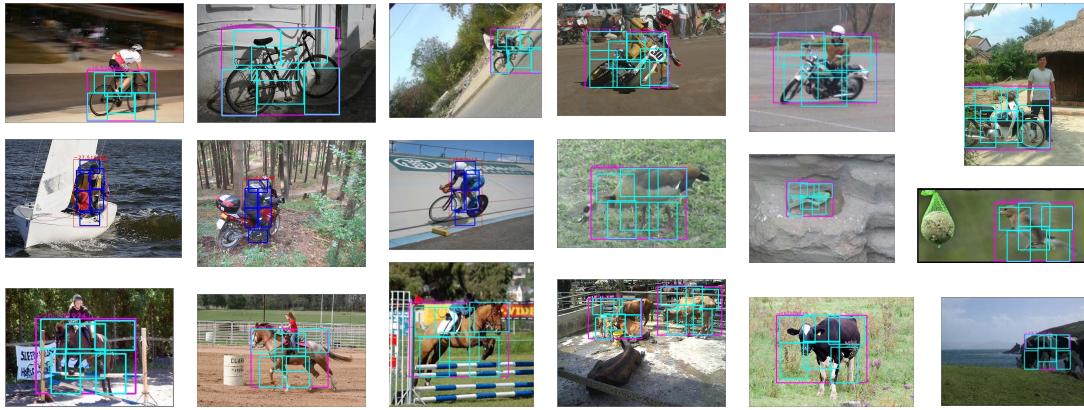


Figure 3. Some results from the six categories.

VI. CONCLUSION

We have proposed a component-based hierarchical framework for object detection and localization in images. The framework learns the root and each component classifier respectively, then combines spatial model to realize object detection. Specifically, we train the component classifiers in a weakly supervised manner, where object labels are provided but component labels are not. We evaluate the method on recent benchmark for object recognition [11] and demonstrate the competitive performance of our system in object detection. Currently, we train detectors independently for each class, which brings the computational complexity. We will consider the problem of multiclass detection by finding common features that can be shared across the classes.

REFERENCES

- [1] N. Dalal and B. Triggs, “Histograms of oriented gradients for human detection,” CVPR, 2005
- [2] P. Viola and M. Jones, “Rapid object detection using a boosted cascade of Simple Features,” CVPR, 2001
- [3] S. Agarwal, A. Awan, and D. Roth, “Learning to detect objects in images via a sparse, part-based representation,” PAMI, 2004
- [4] J. Friedman, T. Hastie, and R. Tibshirani, “Additive logistic regression: a statistical view of boosting,” Annals of Statistics, 2000
- [5] R. Fergus, P. Perona, and A. Zisserman, “A sparse object category model for efficient learning and exhaustive recognition,” CVPR, 2005
- [6] F. Felzenszwalb and D.P. Huttenlocher, “Pictorial structures for object recognition,” IJCV, 2005
- [7] D.J. Crandall, P.F. Felzenszwalb, and D.P. Huttenlocher, “Weakly supervised learning of part-based spatial models for visual object recognition,” ECCV, 2006
- [8] Y. Amit and A. Trouve, POP: Patchwork of parts models for object recognition. IJCV, 2007
- [9] F. Porikli, “Integral Histogram: A fast way to extract histograms in cartesian spaces,” CVPR, 2005
- [10] I. Laptev, “Improvements of object detection using boosted histograms,” BMVC, 2006.
- [11] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, The PASCAL Visual Object Classes Challenge 2007 (VOC2007) Results, <http://www.pascal-network.org/challenges/VOC/voc2007/workshop/index.html>
- [12] P. Felzenszwalb, D. McAllester, and D. Ramanan, “A discriminatively trained, multiscale, deformable part model,” CVPR, 2008
- [13] P. Dollar, B. Babenko, S. Belongie, P. Perona, and Z. Tu, “Multiple component learning for object detection,” ECCV, 2008
- [14] A. Opelt, A. Pinz, and A. Zisserman, “A boundary-fragment-model for object detection,” ECCV, 2006
- [15] R. Lienhart, A. Kuranov, and V. Pisarevsky, “Empirical analysis of detection cascades of boosted classifiers for rapid object detection,” Proc. DAGM 25th Pattern Recognition Symp, 2003