

# Part Detection, Description and Selection based on Hidden Conditional Random Fields

Wenhao Lu, Shengjin Wang, Xiaoqing Ding

Dept. of Electronic Engineering  
Tsinghua University  
Beijing, P R China

e-mail: {luwenhao, wsj, dxq}@ocrserv.ee.tsinghua.edu.cn

**Abstract**—In this paper, the problem of part detection, description and selection is discussed. This problem is crucial in the learning algorithms of part-based models, but can't be solved well when some candidate parts are extracted from background. This paper studies this problem and introduces a new algorithm, HCRF-PS (Hidden Conditional Random Fields for Part Selection), for part detection, description, especially selection. Our algorithm is distinguished for its power to optimize multiple kinds of information at the same time, including texture, color, location and part label. Finally, we did some experiments with HCRF-PS algorithm which give good results on both virtual and real data.

**Keywords**—HCRF, part detection, part description, part selection, part-based model

## I. INTRODUCTION

The description of object classes is a fundamental and difficult problem in the computer vision field. Recently, many researchers have focused their research on the part-based model [1, 2, 3, 4, 5, 6, 7, 8, 11, 12] which can handle much greater in-class variation while keep enough discriminative information for classification or detection. Generally, a learning algorithm for part-based models can be split into two steps shown in Fig 1. In the first step, part detection and description is operated on every sample in the training set. In the second step, a geometry model is learned with some probabilistic graph, like Bayesian Network[3], or a discriminative model, like SVM, Adaboost[1].

In all these algorithms, part detection, description and selection play important roles though are not the main topic in most of the articles. Generally, part detection, description and selection need at least three steps: candidate part detection, candidate part description and part selection. It's hard to give a strict order to these steps, because they are not independent as shown in Fig 1. Current algorithms always cope with this problem by assuming that all the labels of candidate parts are foreground parts and clustering the parts according to their locations [3, 4]. This method simply reduced the information used for clustering parts and is easy to implement, but it sometimes fails because there might be many background patches in the candidate parts set or need more information for clustering.

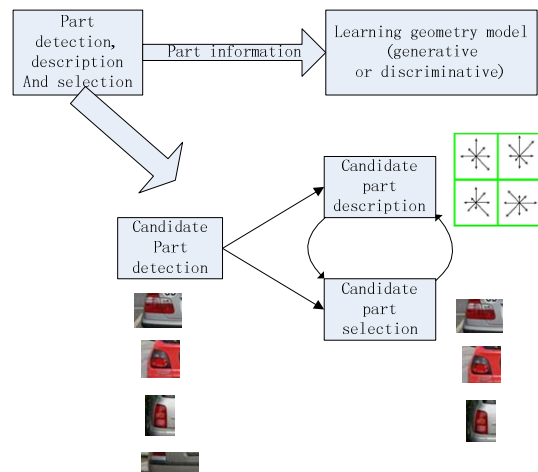


Figure 1. Learning algorithm for a part-based model

In this paper, an algorithm for solving the problems of part detection, description and selection is introduced. We develop this algorithm for a better solution when more candidate parts belong to background, which is a common situation in part selection problems (In this paper, 'part' indicate the kind of regions, while 'candidate parts' are specific regions in certain images which may be foreground or background patches). We begin in Section II by reviewing some related works and pointing out our main idea. Sections III and IV describe our algorithm. Section V gives some results of our algorithm and conclusion will be summarized in Section VI.

## II. RELATED WORK

Many papers discussed the part-base model [1, 2, 3, 4, 5, 6, 7, 8] which involves the problems of part detection, description and selection. Fergus et al [2] and Sivic et al [7] used the bag-of-words manner to form codebook. Each entry of the codebook can be more or less explained as a part. But the key point detector will also detect some key points in the background which make codebook a little chaotic and less semantic. Some region detectors, like MSER [9] and Saliency points [10], will enlarge the part's scale and bring some semantic information to the part, but can't reduce the influence of background. B.Ommer et al [3] used

agglomeration of key points to form candidate parts, fixed the location to learn the description of a part with NKDA, and selected the canonical parts using a validation set. Hao Su et al [4] described and selected parts in the same way. They trained a random forest classifier for each part while keeping the relative position information to the center of the object.

Intuitively, people cluster candidate parts following some rules: 1. a part must have less variance inside; 2. parts of the same kind always appear in the same location and have some appearance in common. For example, color and location information is more important for car bulb while shape is more important for the window shield. We emphasize our algorithm in two aspects. First, we introduce our candidate part detection and description method which is fit for the latter part selection algorithm. Second, we introduce the main part detection, description and selection algorithm which is based on HCRF (Hidden Conditional Random Fields). This algorithm optimized the location, color, texture and part-label information as a whole compared to the current algorithms which just fix some and optimize the others. Experiments show better performance than the method that simply clusters by one or two kind of information.

### III. CANDIDATE PART DETECTION AND DESCRIPTION

First, we introduce the way to generate candidate parts. We used the traditional MSER detector but ignore the step of fitting ellipse. Some incomplete regions can be extracted in the image, see Fig 2. Then, these regions are dilated to fill some gaps inside the regions. Region information is formed afterwards. In this paper, we extracted three kinds of information to describe a region which are texture, color and location. Given the area of the region, a rectangle can be fitted and SIFT [13] descriptor is extracted after normalizing the shape of the region. Color information is described by a color histogram which used only the pixels in MSER region instead of the rectangle area. Location is obtained from the weighted center of the MSER region.



Figure 2. MSER region detection results

### IV. HCRF-PS (HIDDEN CONDITIONAL RANDOM FIELDS FOR PART SELECTION)

Second, we give a detailed introduction of our HCRF-PS algorithm which covers the part detection, description and selection procedures.

As we assume that there are background patches in the candidate part set, a background label is included in the label set while the other labels indicate different part of the class object. Our purpose is to assign labels to all these candidate parts (or seen as a clustering problem) and optimizes each

part's description of location, texture and color. As described above, it's inefficient to fix the location and optimize other information when a proportion of candidate parts are obtained from background or just not discriminative enough. Therefore, we develop HCRF-PS algorithm to optimize all these information, including part label, texture, color and location together.

HCRF (Hidden Conditional Random Fields) is a popular probabilistic graph introduced by A.Quattoni et al in [5][6]. It is the hidden version of CRF (Conditional Random Fields). The traditional HCRF optimize the conditional probabilistic

$$P(y|x, \theta) = \sum_{\bar{h}} P(y, \bar{h}|x, \theta) = \frac{\sum_{\bar{h}} \exp\{\psi(y, \bar{h}, x; \theta)\}}{\sum_{y, \bar{h}} \exp\{\psi(y, \bar{h}, x; \theta)\}} \quad (1)$$

Where  $\bar{h}$  are the hidden label,  $\psi(y, \bar{h}, x; \theta)$  is described as a linear function of the parameters  $\theta$ .

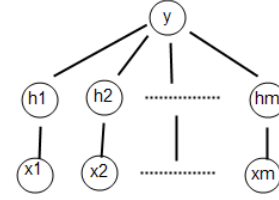


Figure 3. One kind of HCRF probabilistic graph

In our HCRF-PS algorithm, we use the graph in Fig 3 and assume that  $m$  candidate parts are extracted in each training image. The information of a candidate part, like color, texture and location, is represented by the observed variable  $x_j (j=1..m)$ ; part label for every candidate part is indicated by hidden variable  $h_j (j=1..m, h_j=1..H)$ ;  $y$  represents the class label of the training image.

The form of the optimization function in HCRF-PS is different from the traditional linear form.

The potential function:

$$P(x, \bar{h}, y; \theta) = \exp\{\psi(y, \bar{h}, x; \theta)\} \quad (2)$$

$$\psi(y, \bar{h}, x; \theta) = \sum_{j=1}^m \sum_{l=1}^L \psi_l(x_{jl}, h_j) + \sum_{j=1}^m \psi_0(h_j, y) \quad (3)$$

Where  $l$  is the indicator of different information;  $\psi_0$  is a function to represent the contribution of each kind of part for distinguishing a class object against backgrounds.

$$\begin{cases} \psi_l(x_{jl}, h_j) = \sum_{k=1}^{H-1} \delta(h_j - k) \phi(x_{jl}, \mu_{kl}, \sigma_{kl}) + \delta(h_j - H) \theta_l \\ \psi_0(h_j, y) = \sum_{c=1}^C \sum_{k=1}^H \delta(h_j - k, y - c) \theta_{kc} \end{cases} \quad (4)$$

Where

$$\phi(x_{jl}, \mu_{kl}, \sigma_{kl}) = -\frac{(x_{jl} - \mu_{kl})^T (x_{jl} - \mu_{kl})}{2\sigma_{kl}^2} - \frac{\log(2\pi\sigma_{kl}^2)}{2} \quad (5)$$

We give the  $\phi(x_{jl}, \mu_{kl}, \sigma_{kl})$  function a Gaussian distribution form, but it's a little different from the one introduced in

[14]. First, we treat it as a 1-order Gaussian distribution of the distance  $d = \sqrt{(x_{j_l} - \mu_{kl})^T (x_{j_l} - \mu_{kl})}$  instead of a multi-dimension Gaussian distribution of the vector  $(x_{j_l} - \mu_{kl})$ ; second, we use the variables to represent mean and variance instead of moments like  $x_1^2, \alpha x_2, \beta x_1 x_2$ . Obviously,  $\mu_{kl}$  is the mean of part-k's information l while  $\sigma_{kl}$  can be treated as the 'weight' ( $weight_{kl} \sim 1/\sigma_{kl}$ ) of different information when they are combined. It also needs to be mentioned that we introduce a background part label ( $h_j = H$ ). When a candidate part doesn't fit with any foreground part, the algorithm will assign the background part label to it automatically.

The optimization procedure usually uses Quasi-Newton with BFGS method or other gradient descent methods. The gradient can be calculated as follow:

The log-likelihood:

$$l(\theta) = \log(P(y|x)) = \log \sum_{\tilde{h}} \exp\{\psi(y, \tilde{h}, x; \theta)\} - \log \sum_{\tilde{h}'} \exp\{\psi(y', \tilde{h}', x; \theta)\} \quad (6)$$

The final derivative is given below:

$$\frac{\partial}{\partial \theta} l(\theta) = \sum_{j=1}^m P(h_j = k | x, y, \theta) \frac{\partial}{\partial \theta} \phi(\theta) - \sum_{y=1}^c \sum_{j=1}^m P(h_j = k, y' | x, \theta) \frac{\partial}{\partial \theta} \phi(\theta) \quad (7)$$

$$\frac{\partial}{\partial \theta_{kc}} l(\theta) = \sum_{j=1}^m P(h_j = k | x, y, \theta) \delta(y - c) - \sum_{y=1}^c \sum_{j=1}^m P(h_j = k, y' | x, \theta) \delta(y' - c) \quad (8)$$

Where  $\phi(\theta)$  is given in Equation (5), and  $\theta = \mu_{kl}, \sigma_{kl}$ ,  $k = 1 \dots H, l = 1 \dots L$ ,  $L$  is the number of information types used.

## V. EXPERIMENT RESULT

Two experiments will be shown in this section. The first experiment tests the HCRF-PS with data that generated manually, while the second experiment tests the HCRF-PS with candidate parts generated from real images. The experiment results show good clustering performance and reduction of the background candidate parts.

### A. Experiment I

In experiment I, the data is generated as follow: assume there are 4 foreground parts with 1 background part; every candidate part has two kinds of information-----appearance and location; candidate parts are generated following some manually set proportion. The initial parameters are generated by adding Gaussian noise  $N(0, \sigma_{init})$  on the original parameters which used to generate the data.

We test our algorithm on virtual data to evaluate its ability of coping with initial bias for large bias on initial parameter will lead to local extreme points in this optimization problem. Table I shows results changed with the variance  $\sigma_{init}$  of the Gaussian noise added on the initial parameters. As we normalized the parameters, the entries of the parameters are less than 0.4 on average. Therefore, we

add the noise from  $\sigma_{init} = 0.05$  to  $\sigma_{init} = 0.2$ . Keeping the initial parameters to be positive is important, for we find the algorithm can't converge when  $\sigma_{init} \geq 0.1$  if some negative entries exist in the initial parameters. That means the algorithm will probably converge to another extreme point if the initial parameters step over zero.

TABLE I. RESULT CHANGED WITH THE INITIAL PARAMETERS

| $\sigma_{init}$ | Converged or Not |
|-----------------|------------------|
| 0.05            | Y                |
| 0.1             | Y (set positive) |
| 0.2             | Y (set positive) |
| 0.25            | N (set positive) |

### B. Experiment II

In experiment II, the candidate parts are generated from the training images of PASCAL VOC06 car dataset [15] using MSER region detector. Every region is described with a 128-dimension SIFT-descriptor and a 72-dimension color histogram. We used HCRF-PS algorithm to cluster these candidate parts and find the optimal description of every kinds of parts. We compared our algorithm with a baseline using k-means. Fig 4 shows some positive and negative images used in our experiment. Table II compares the HCRF-PS algorithm and k-means algorithm. Fig 5 shows some part clusters of these two algorithms. As we can see, k-means might split one kind of parts into two clusters or keep some background candidate parts in a foreground cluster, but HCRF-PS algorithm performances much better and gains some synchronous results with human semantics. Table III gives the parameters  $\theta_{kc}$  which represent the discriminative power of every learned part. Purer part clusters and larger clusters get higher score:  $score = \theta_{k+} - \theta_{k-}$ , which means the parts are more important to cars. The parts with higher scores are car plate (part1), left bulb (part4), window shield (part5), car top (part6), right bulb (part8) and car bottom (part10).

TABLE II. COMPARISON OF HCRF-PS AND K-MEANS

| Algorithm   | Error classified parts |
|---|------------------------|
| K-means by location (K=14)                            | 34%                    |
| K-means by texture (K=14)                             | Chaotic                |
| HCRF-PS (13 foreground parts with 1 background parts) | 15.1%                  |

TABLE III. LEARNED  $\theta_{kc}$  WHICH REPRESENT THE DISCRIMINATIVE POWER OF EVERY LEARNED PART

| $\theta_{kc}$ | Part1 | Part2 | Part3  | Part4  | Part5  | Part6  | Part7   |
|---------------|-------|-------|--------|--------|--------|--------|---------|
| +             | 8.14  | 4.23  | 4.95   | 7.93   | 8.34   | 6.22   | 5.10    |
| -             | -7.01 | -3.37 | -3.94  | -6.81  | -7.15  | -5.26  | -4.13   |
|               | Part8 | Part9 | Part10 | Part11 | Part12 | Part13 | Bg part |
| +             | 6.28  | 4.54  | 6.29   | 5.28   | 5.55   | 4.71   | -3.38   |
| -             | -5.27 | -3.59 | -5.19  | -5.29  | -4.60  | -6.56  | 8.41    |



Figure 4. Some positive and negative images used in our experiment

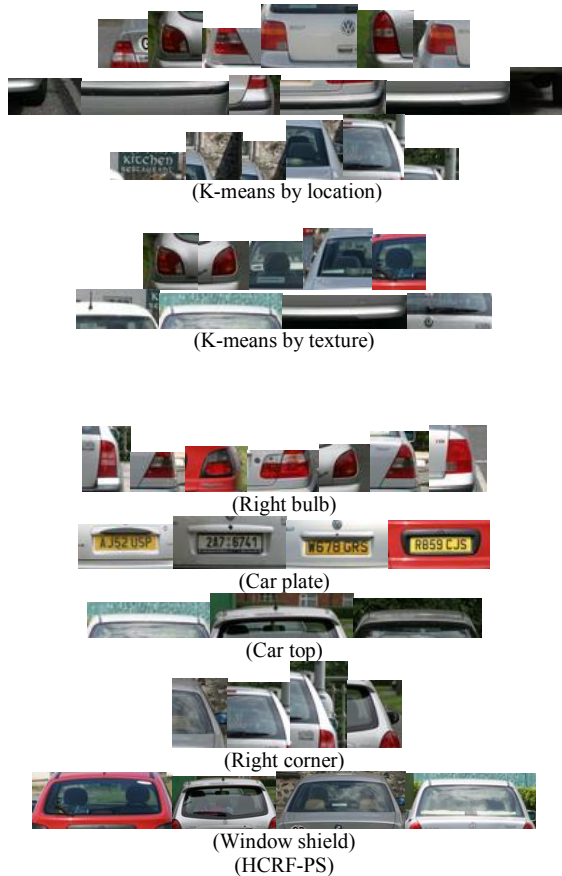


Figure 5. Some results of the k-means and HCRF-PS algorithm

## VI. CONCLUSION

In this paper, we introduce a new algorithm-----HCRF-PS for part detection, description and selection, experiments with this algorithm gives better results in part clustering problem. It can be used in the learning procedure of part-based models or can be just used for clustering. Many more information other than texture and color can be combined

which will make the algorithm more powerful but won't need to change its framework. For future work, we would like to incorporate the geometry relationship between parts into the algorithm to make the algorithm more stable.

## ACKNOWLEDGMENT

This work is supported by the National Basic Research Program of China (973 program) under Grant No.2007CB311004, the National High Technology Research and Development Program of China (863 program) under Grant No. 2009AA11Z214 and Doctoral Fund of Ministry of Education of China under Grant No. 20090002110077.

## REFERENCES

- [1] A. Opelt, A. Pinz, and A. Zisserman. A boundary-fragment model for object detection. In Proc. ECCV, volume 2, pages 575–588, May 2006.
- [2] R. Fergus, L. Fei-Fei, P. Perona, and A. Zisserman, Learning object categories from google's image search. in IEEE ICCV 2005., pp. 1816–1823
- [3] B. Ommer, M. Sauter, and J. M. Buhmann, Learning top-down grouping of compositional hierarchies for recognition, in Proc IEEE CVPR, 2006.
- [4] Hao Su, Min Sun, Li Fei-Fei, Silvio savarese, Learning a dense multi-view representation for detection, viewpoint classification and synthesis of object categories, IEEE ICCV 2009
- [5] A. Quattoni, M. Collins, and T. Darrell, Conditional Random Fields for Object Recognition, Proc. IEEE Conf. Neural Information Processing Systems, 2004.
- [6] Ariadna Quattoni, Sybor Wang, Louis-Philippe Morency, Michael Collins and Trevor Darrell, Hidden Conditional Random Fields, IEEE Trans Pattern Analysis and Machine Intelligence, VOL. 29, NO. 10, OCTOBER 2007
- [7] J. Sivic, B. C. Russell, A. A. Efros, A. Zisserman, and W. T. Freeman, Discovering objects and their localization in images. in IEEE ICCV 2005, pp. 370–377
- [8] R. Fergus, P. Perona, and A. Zisserman, Object class recognition by unsupervised scale-invariant learning, in Proc IEEE CVPR 2003, pp. 264–271.
- [9] J. Matas, O. Chum, M. Urban, and T. Pajdla. Robust wide baseline stereo from maximally stable extremal regions. In Proc. BMVC, pages 384–393, 2002.
- [10] T. Kadir and M. Brady. Scale, saliency and image description. International Journal of Computer Vision, s45(2):83–105, 2001.
- [11] D. Hoeim, C. Rother, and J. Winn. 3d layoutcrf for multi-view object class recognition and segmentation. CVPR, 2007.
- [12] A. Kushal, C. Schmid, and J. Ponce. Flexible object models for category-level 3d object recognition. CVPR, 2007.
- [13] D. Lowe. Object recognition from local scale-invariant features. ICCV, 1999.
- [14] A. Gunawardana, M. Mahajan, A. Acero, and J. Platt, Hidden conditional random fields for phone classification, in Proc. Eurospeech, 2005, pp. 1117–1120
- [15] M. Everingham, A. Zisserman, C. K. I. Williams, and L. Van Gool. The PASCAL Visual Object Classes Challenge 2006 Results. Technical Report, PASCAL Network, 2006.