# SEMI-SUPERVISED OBJECT RECOGNITION USING STRUCTURE KERNEL

*Botao Wang, Hongkai Xiong* — *Xiaoqian Jiang* [*] — *Fan Ling* [†]

Department of Electronic Engineering
Shanghai Jiao Tong University
Shanghai, 200240, China

Division of Biomedical Informatics
University of California, San Diego
La Jolla 92093, CA, USA

Wireless Comm. Tech. Limited
QUALCOMM
Shanghai 200040, China

## ABSTRACT

Object recognition is a fundamental problem in computer vision. Part-based models offer a sparse, flexible representation of objects, but suffer from difficulties in training and often use standard kernels. In this paper, we propose a positive definite kernel called "structure kernel", which measures the similarity of two part-based represented objects. The structure kernel has three terms: 1) the global term that measures the global visual similarity of two objects; 2) the part term that measures the visual similarity of corresponding parts; 3) the spatial term that measures the spatial similarity of geometric configuration of parts. The contribution of this paper is to generalize the discriminant capability of local kernels to complex part-based object models. Experimental results show that the proposed kernel exhibit higher accuracy than state-of-art approaches using standard kernels.

***Index Terms***— Object recognition, kernel, image features, machine learning, data mining

## 1. INTRODUCTION

Object recognition is a fundamental problem in computer vision and artificial intelligence. Partial occlusion, background clutters and non-rigid deformation are among the most challenging problems in object recognition. In recent years, significant improvements have been made in object recognition with machine learning techniques, such as support vector machine [1] and kernel methods [2].

Kernels are symmetric bivariate functions that capture resemblance between input data. We assume $\Phi(\boldsymbol{x}) : \mathcal{X} \rightarrow \mathcal{H}$ to be a function that maps $\boldsymbol{x}$ from original data space to a high dimensional Hilbert space. The kernel function calculates the inner product of the mapped data in $\mathcal{H} : K(\boldsymbol{x}, \boldsymbol{x}') = \Phi(\boldsymbol{x}) \cdot \Phi(\boldsymbol{x}')$ without explicitly computing the mapped data. To ensure the existence of such mapping, the kernel function must satisfy positive definiteness condition, which is also called the Mercer condition. In general, kernel methods can be utilized in many machine learning techniques as long as they handle only inner products of the input data.

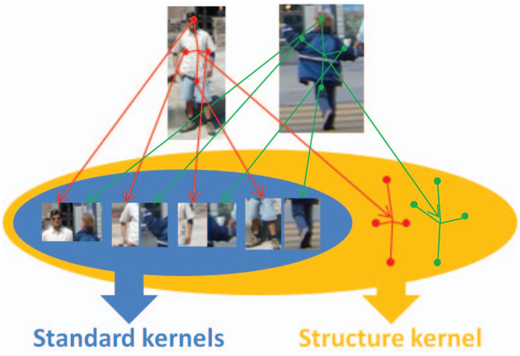


**Fig. 1**. Examples of standard kernels and the structure kernel. Standard kernels do not take spatial dependency into consideration while the structure kernel uses spatial dependency to improve the quality of object detection.

In this paper, we propose a positive definite kernel called "structure kernel" to measure the similarity of two part-based represented objects. The main contribution of this paper is to incorporate the discriminant capability of local kernels into complex object models. In our approach, we use a part-based representation of objects, which models an object as a collection of visual parts arranged in a deformable configuration. Visual appearance of each part is encoded with a region-based image descriptor, which is robust to background clutters and semantically significant. Parts are arranged in a deformable configuration, which offers more flexibility and invariance to partial occlusion. We propose a structure kernel to measure the similarity of complex models, which is more discriminant than standard kernels because it takes both visual appearance and the spatial configuration of corresponding parts into account, as indicated in Figure 1. Experimental results show that the proposed method has accurate performance and outperform many state-of-the-art approaches.

## 2. RELATED WORK

Summation kernel [3] calculates the sum of all of the cross-similarities between all of the possible combination of feature vectors. It is a Mercer kernel but its discriminative ability is compromised because good matchings can be easily swamped by bad matchings. The "max" kernel [4] improves

[*] X. Jiang was partially funded by NIH, U54 HL108640.
[†] This work was partially supported by QUALCOMM.

the summation kernel by summing only the similarities of the best matched feature vectors. But the "max" operation makes the kernel non positive definite and this is risky because the implicit mapping may not exist. In [3], a circular-shift invariant kernel was proposed to measure the neighborhood similarity of two keypoints, but the geometric configuration is only reflected by the orientation of the keypoint. In [5], a context-dependent kernel was proposed, which considers the context as a part of the alignment process in designing kernels.

## 3. STRUCTURE KERNEL

### 3.1. Object Representation

Our model represents an object as a collection of parts arranged in certain geometric configuration. We use both the global feature and local features of parts to represent an object. Global feature is typically coarse and sensitive to background clutters and partial occlusion, and through incorporating local features that are more distinctive among the object category, we expect the detection result can be more robust.

An example is shown in Figure 2(b), where the object is represented as a $n+1$ tuple:

$$\boldsymbol{x} = (F_0, P_1, \ldots, P_n) \tag{1}$$

where $F_0$ is the feature vector of the whole object. $P_i = (F_i, g_i), i = 1, \ldots, n$ are part models, where $F_i$ is the feature vector of part $i$ and $g_i$ is the 2D coordinate of the part relative to the whole object.

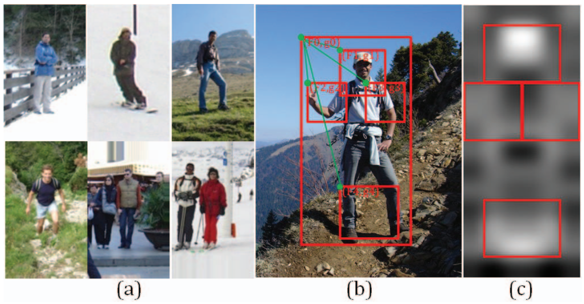


**Fig. 2**. (a) Positive examples from INRIA dataset; (b) Object representation; (c) Object model and part selection.

To encode the visual appearances of object and parts $F_i, i = 1, \ldots n$, we utilize a variation of histogram of oriented gradients used in [6]. Image patches are divided into $8 \times 8$ non-overlapping cells, each of which is represented by a 31 dimensional feature vector consisting of 4 sums over 9 contrast insensitive orientations, 9 sums over different normalizations for contrast insensitive orientations, and 18 sums over different normalizations for contrast sensitive orientations.

### 3.2. Definition of Structure Kernel

Let $\mathcal{F}_i, i = 0, 1, ..., n$ be the feature space of the feature vectors of the whole object ($i = 0$) and parts ($i = 1, ..., n$) and $\mathcal{G} = \mathbb{N} \times \mathbb{N}$ is the space of 2D coordinate. The feature space of part $i$ can be represented as $\mathcal{P}_i = \mathcal{F}_i \times \mathcal{G}$. Together, the feature space of an object is $\mathcal{X} = \mathcal{F}_0 \times \mathcal{P}_1 \times ... \times \mathcal{P}_n$.

Given two structured represented objects $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$, a structure kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is defined as:

**Definition 1** (Structure Kernel)**.** *Let $\mathcal{X}$ be the input space, $\boldsymbol{x}, \boldsymbol{x}' \in \mathcal{X}$ are two structured data. We define the structure kernel $\mathcal{K} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ between $\boldsymbol{x}$ and $\boldsymbol{x}'$ as*

$$\begin{aligned}\mathcal{K}(\boldsymbol{x}, \boldsymbol{x}') = \mathcal{K}_{00}\left(\boldsymbol{F}_0, \boldsymbol{F}_0'\right) + \sum_{i=1}^{n} \mathcal{K}_{0i}(\boldsymbol{F}_i, \boldsymbol{F}_i') \\ + \lambda \sum_{i=1}^{n} \exp\left\{-\frac{(\boldsymbol{g}_i - \boldsymbol{g}_i')^2}{2\sigma^2}\right\}\end{aligned} \tag{2}$$

*where $\mathcal{K}_{0i}(\cdot) : \mathcal{F}_i \times \mathcal{F}_i \to \mathbb{R}$ is a standard positive definite kernel, $\lambda : \lambda > 0$ is a kernel parameter that balances the relative weights between the appearance similarity and the geometric similarity, and $\sigma$ is the parameter of the measurement of the geometric similarity.*

The structure kernel consists of three terms: the first term in Eq. (2) is a "global term", which measures the resemblance of the feature vectors of the whole objects. The second term in Eq. (2) is a "part term", which measures resemblance of the feature vectors of of all the corresponding part pairs. Both the global term and the part term compare visual appearances of objects and parts, which are calculated with a pre-defined, positive definite, standard kernel function $\mathcal{K}_0$. The third term in Eq. (2) is a "spatial term", which reflects the similarity of the spatial configurations of these two objects.

### 3.3. Mercer Condition

To guarantee the existence of the high dimensional reproducing kernel Hilbert space, the structure kernel $\mathcal{K}$ must satisfy the Mercer condition: for any selection of examples $\boldsymbol{x}_1, \ldots \boldsymbol{x}_m \in \mathcal{X}$, the Gram matrix $\boldsymbol{K}$ of the structure kernel $\mathcal{K} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ , which is defined as $\boldsymbol{K}(i,j) = \mathcal{K}(\boldsymbol{x}_i, \boldsymbol{x}_j)$, is positive definite.

**Proposition 1.** *Structure kernel is a Mercer kernel.*

*Proof.* Recall that a matrix $\boldsymbol{K}$ is positive definite if and only if $\boldsymbol{\alpha}^T \boldsymbol{K} \boldsymbol{\alpha} > 0$ for all non-zero vector $\boldsymbol{\alpha}$. We denote the Gram matrix for kernel $\mathcal{K}_{0i}$ as $\boldsymbol{K}_{0i}$ and the Gram matrix for $d_i(\boldsymbol{g}_i, \boldsymbol{g}_i') = \exp\left\{-\frac{(\boldsymbol{g}_i - \boldsymbol{g}_i')^2}{2\sigma^2}\right\}$ as $\boldsymbol{D}_i$, $i = 0, \ldots, n$. As $\mathcal{K}_{0i}$ are Mercer kernels and $\exp\left\{-\frac{(\boldsymbol{g}_i - \boldsymbol{g}_i')^2}{2\sigma^2}\right\}$ is in the form of Gaussian function. So $\boldsymbol{K}_{0i}$ and $\boldsymbol{D}_i$ are positive definite. For any $m$-dimensional non-zero vector $\boldsymbol{\alpha}$,

$$\begin{aligned}\boldsymbol{\alpha}^T \boldsymbol{K} \boldsymbol{\alpha} &= \boldsymbol{\alpha}^T \left( \sum_{i=0}^{n} \boldsymbol{K}_{oi} + \lambda \sum_{i=1}^{n} \boldsymbol{D}_i \right) \boldsymbol{\alpha} \\ &= \sum_{i=0}^{n} \boldsymbol{\alpha}^T \boldsymbol{K}_{oi} \boldsymbol{\alpha} + \lambda \sum_{i=1}^{n} \boldsymbol{\alpha}^T \boldsymbol{D}_i \boldsymbol{\alpha} \geq 0\end{aligned}$$

Therefore, the structure kernel satisfies the Mercer condition. □

### 3.4. Training

Our object classifier can be trained in a semi-supervised way where only positive examples are labeled with bounding boxes (i.e., each covers an instance of the object of interest). The classifier is capable of automatically identifying effective parts that have similar visual appearance and geometric configuration in the object category in the training process.

#### *3.4.1. Initializing Global Detector and Part Detectors*

Candidate objects and parts are extracted from images by global detector and part detectors, using linear Support Vector Machine (SVM) classifiers. The detector can be represented as $D = \{\boldsymbol{\beta}, b\}$, where $\boldsymbol{\beta}$ is a linear filter and b is the bias term. An image patch is scored by $s = \boldsymbol{\beta}^T \boldsymbol{F} + b$, where $\boldsymbol{F}$ corresponds to the feature vector of the image patch. The training of global detector is similar to [7].

Parts are defined to be sub-regions of objects which have similar visual appearance in the object category. These regions typically have high values in the filter of the global detector. We greedily place a pool of rectangular average filters to search for high value regions. Parts are symmetric and do not overlap with each other, and total area of all parts covers at least 60% of the object. An example of detectors for "person" category is shown in Figure 2(c).

#### *3.4.2. Train Classifier*

An object hypothesis $\boldsymbol{x} \in \mathcal{X}$ is scored by the discriminant function:

$$f_{\mathcal{K}}(\boldsymbol{x}) = \sum_{i=1}^{N} \alpha_i y_i \mathcal{K}(\boldsymbol{x}_i, \boldsymbol{x}) + b \tag{3}$$

where $\mathcal{K}$ is the structure kernel, $\boldsymbol{x}_i$ are training examples, $y_i \in \{-1, +1\}$ are labels of training examples, $N$ is the number of training examples, and $b$ is the bias term.

The training procedure is shown in Algorithm 1. In training, we solve $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_n)$ by optimizing the dual formulation of the primal SVM formulation:

$$\min_{\alpha_i} \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j \mathcal{K}(\boldsymbol{x}_i, \boldsymbol{x}_j) - \sum_{i=1}^{N} \alpha_i$$

$$\text{s.t.} \begin{cases} 0 \leq \alpha_i \leq C \\ \sum_{i=1}^{N} \alpha_i y_i = 0 \end{cases} \tag{4}$$

Problem (4) is solved by the quadratic programming technique, which is denoted as *Train($P_a$, $N_a$)* in Algorithm 1.

We treat the position of parts as latent variables, and use a latent SVM [6] to train the classifier. In each iteration, we relabel the positive examples and negative examples with the current kernel so that their geometric configurations give the highest scores. With the fixed geometric configuration for each example, the classifier can be trained with standard kernelized SVM techniques.

As there is a large number of training examples, it is inefficient to make use of all of them in the training procedure due to the memory limitation. Therefore, we use a fixed cache of positives and negatives, denoted as $P_a$ and $N_a$ for training, iteratively removing easy examples from the cache, and adding hard examples to the cache. The algorithm below guarantees to converge to the exact solution of the training problem using all training data.

**Algorithm 1** Training classifier

```
Input: positive examples $P$, negative examples $N$
Output: classifier $\mathcal{C}$

for r = 1 to relabel do
    $[P_a, N_a]$ = RandomSelect($P$, $N$);
    $\mathcal{C}$ = Train($P_a$, $N_a$);
    for m = 1 to datamine do
        $[P_a, N_a]$ = RemoveEasyExamples($P_a$, $N_a$, $\mathcal{C}$);
        $[P_a, N_a]$ =AddHardExamples($P - P_a$, $N - N_a$, $\mathcal{C}$);
        $\mathcal{C}$ = Train($P_a$, $N_a$);
    end for
    $[P, N]$ = relabel($P$, $N$);
end for
```

## 4. EXPERIMENTAL RESULTS

We use the INRIA person dataset to evaluate the performance of the proposed structure kernel. Some images in the INRIA person dataset are shown in Figure 2(a). Positive training set consists of 1,208 images and their left-right reflections, i.e. 2,416 images in all. Negative training set is more than 10,000 image patches randomly sampled from 1,218 images that do not contain any instance of person. We set $\lambda = 1$, $C = 0.1$ and $\sigma = 1$.

We evaluate the structure kernel $\mathcal{K}$ using three settings and apply a simple configuration of kernel parameters.

1. Linear kernel: $\mathcal{K}_0(F, F') = F \cdot F'$;

2. Polynomial kernel: $\mathcal{K}_0(F, F') = (\frac{1}{d} F \cdot F' + r)^m$, where d is the dimension of the feature vector, $r = 1, m = 2$;

3. Radial basis kernel: $\mathcal{K}_0(F, F') = \exp(-\gamma \left\| F - F' \right\|^2)$ with $\gamma = 0.001$.

The test set consists of 1,216 positive images and 12,160 negative images. The precision-recall curves are displayed in Figure 3, and average precision values are shown in Table 1. Furthermore, we also compare our result with Felzenszwalb detector [6] and Dalal&Triggs detector [7].

In Figure 3, linear structure kernel achieves the best performance at low-recall-high-precision region and radial basis

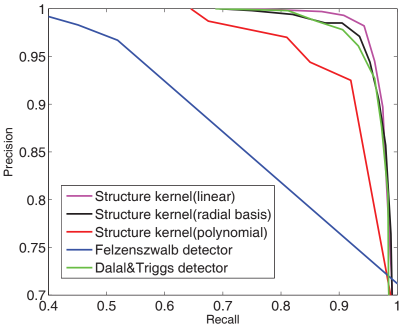


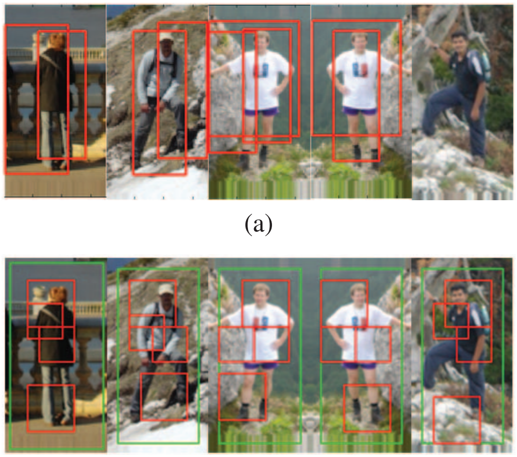
**Fig. 3**. Precision-recall curves of structure kernel with different $\mathcal{K}_0$ and other approaches for comparison.

**Table 1**. Average precision of different methods

| SK(linear) | SK(RBF) | SK(poly) | Dalal | P. F. F |
|---|---|---|---|---|
| 0.988 | 0.986 | 0.958 | 0.983 | 0.857 |

structure kernel achieves best performance at high-recall-low-precision region. The linear kernel outperforms the RBF kernel in the experiment because we limit the search space of $\gamma \in (0.0001, 0.01)$ of the RBF kernel for tractable computation, but in theory, RBF should always reduce to linear kernel.

In Figure 4, we illustrate some test images from the INRIA dataset, which are incorrectly detected by Felzenszwalb's approach [6] but are correctly detected by our proposed structure kernel.

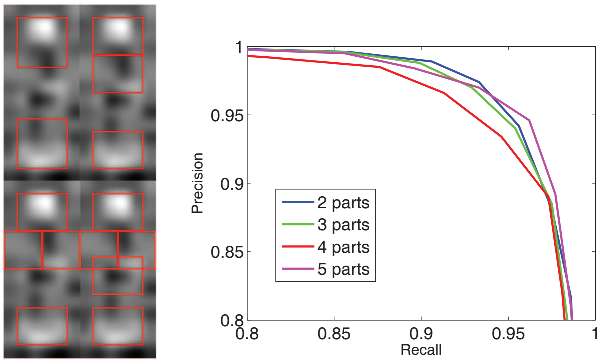


(a)

(b)

**Fig. 4**. Comparison of detection results: (a) Felzenszwalb's approach [6]; (b) Structure kernel.

We also evaluate the influence of part numbers to the performance of the structure kernel. We change the part numbers from 2 to 5, which is shown in Figure 5(a), and train classifiers where $\mathcal{K}_0$ is a linear kernel. The precision recall curve is shown in Figure 5(b) and the average precision values are listed in Table 2. These results indicate the structure kernel is fairly robust with different part numbers.



(a)

(b)

**Fig. 5**. (a) Different part configurations; (b) precision recall curves of different part configurations.

**Table 2**. Average precision of different part numbers

| 2 parts | 3 parts | 4 parts | 5 parts |
|---|---|---|---|
| 0.987 | 0.985 | 0.982 | 0.987 |

### 5. CONCLUSIONS AND FUTURE WORK

In this paper, we propose a novel positive definite kernel called "structure kernel", which measures the similarity of two part-based represented objects in both appearances and spatial configurations. Experimental results show that the structure kernel achieves high accuracy in object detection tasks and outperforms state-of-the-art approaches. Future work will concentrate on the automatic tuning of parameters to the structure kernel in training, so that the kernel can be more flexible to fit the properties of various object categories. Furthermore, we will evaluate the structure kernel on larger datasets, like the PASCAL VOC challenge.

## References

[1] V. Vapnik, "Statistical learning theory," *Wiley*, 1998.

[2] B. Scholkopf and A. J. Smola, "Learning with kernels," *MIT Press*, 2002.

[3] S. Lyu, "Mercer kernels for object recognition with local features," in *CVPR*, 2005, pp. 223–229.

[4] C. Wallraven and B. Caputo, "Recognition with local features: the kernel recipe," in *ICCV*, 2003, pp. 257–264.

[5] H. Sahbi, J. Audibert, and R. Keriven, "Context-dependent kernels for object classification," *PAMI*, vol. 33, no. 7, pp. 699–708, 2011.

[6] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part based models," *PAMI*, vol. 32, no. 9, pp. 1627–1645, 2010.

[7] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *CVPR*, 2005, pp. 886–893.