International Journal of Computer Mathematics
Vol. 84, No. 9, September 2007, 1285–1297

Combinatorial and statistical methods for part selection for object recognition

ZHIPENG ZHAO†*, AKSHAY VASHIST†, AHMED ELGAMMAL†,
ILYA MUCHNIK†‡ and CASIMIR KULIKOWSKI†

†Department of Computer Science, Rutgers, The State University of New Jersey, Piscataway, NJ 08854, USA
‡Centre for Discrete Mathematics and Theoretical Computer Science (DIMACS), Rutgers, The State University of New Jersey, Piscataway, NJ 08854, USA

(Received 18 September 2006; accepted 24 November 2006)

In object recognition tasks, where images are represented as constellations of image patches, often many patches correspond to the cluttered background. In this paper, we present a two-stage method for selecting the image patches which characterize the target object class and are capable of discriminating between the positive images containing the target objects and the complementary negative images. The first stage uses a combinatorial optimization formulation on a weighted multipartite graph. The following stage is a statistical method for selecting discriminative patches from the positive images. Another contribution of this paper is the part-based probabilistic method for object recognition, which uses a common reference frame instead of reference patch to avoid possible occlusion problems. We also explore different feature representation using principal component analysis (PCA) and 2D PCA. The experiment demonstrates our approach has outperformed most of the other known methods on a popular benchmark dataset while approaching the best known results.

Keywords: Class recognition; Computer vision; Feature selection; Object detection; Pattern representation and modelling

AMS Subject Classifications: 62H20; 62H35; 68T10; 68T45

1. Introduction

Object detection and class recognition is a classical fundamental problem in computer vision which has been the subject of much research. This problem has two critical components: representation of the images (image features) and recognition of the object class using this representation which requires learning models of objects that relate the object geometry to image representation. Both the representation problem, which attempts to extract features capturing the essence of the object, and the subsequent classification problem are active areas of research and have been widely studied from various perspectives. The methods for recognition stage can be broadly divided into three categories: 3D model-based methods, appearance

*Corresponding author. Email: zhipeng@cs.rutgers.edu

International Journal of Computer Mathematics
ISSN 0020-7160 print/ISSN 1029-0265 online © 2007 Taylor & Francis
http://www.tandf.co.uk/journals
DOI: 10.1080/00207160601167045

template search-based methods, and part-based methods. 3D model-based methods (e.g. [1]) are successful when we can describe accurate geometric models for the object. Appearance based matching approaches are based on searching the image at different locations and different scales for best match for object 'template' where the object template can be learned from training data and act as a local classifier [2, 3]. Such approaches are highly successful in modelling objects with wide within-class appearance variations such as in face detection [2, 3] but they are limited when the within-class geometric variations are large, such as detecting a motorbike.

In contrast, object recognition based on dense local 'invariant' image features have recently shown a lot of success [4–12] for objects with large within-class variability in shape and appearance. In such approaches objects are modelled as a collection of parts or local features and the recognition is based on inferring the object class based on the similarity of parts' appearance and their spatial arrangement. Typically, such approaches find interest points using some operator such as [13] and then extract local image descriptors around such interest points. Several local image descriptors have been suggested and evaluated, such as Lowe's scale invariant features (SIFT) feature [5], entropy-based scale invariant features [10, 13] and other local features which exhibit affine invariance such as [14–16]. Other approaches that model objects using local features include graph-based approaches such as [17]. In this paper, we adopt a part-based method with a common reference frame. We also experiment with both principal component analysis (PCA) and 2D PCA [18] for image patch representation.

An important subtask in object recognition lies at the interface between feature extraction and their use for recognition. It involves deciding which extracted features are most suitable for improving recognition rate [7], because the initial set of features is large, and often features are redundant or correspond to clutters in the image. Finding such actual object features reduces the dimensionality of the problem and is essential to learn a representative object model to enhance the recognition performance. Weber *et al.* [7] suggested the use of clustering to find common object parts and to reject background clutter from the positive training data. In such an approach large clusters are retained as they are likely to contain parts coming from the object. A similar approach has been used in [19]. However, there is no guarantee that a large cluster will just contain only object parts. Since the success of recognition is based on using many local features, such local features (parts) typically correspond to low level feature rather than actual high level object parts. In this paper we introduce two complementary approaches to select discriminative object parts from a pool of parts extracted from the training images.

1.1 *Contributions*

The contribution of this paper is twofold. Firstly, we introduce a probabilistic Bayesian approach for recognition where the object model does not need a reference part [10]. Instead, object parts are related to a common reference frame. Secondly, we propose two approaches for unsupervised selection of discriminative parts that find features that best discriminate the positive and negative examples. The first is a combinatorial optimization approach which optimally finds the best subsets of features common to the positive examples and distant from the negative examples. The second is a statistical approach which finds features that best discriminate the positive and negative examples. Experimental results show that each of the approaches enhances the recognition rate significantly. Since the two approaches are complementary in the way they select features, combining the two approaches in a sequential manner enhances the results even further.

The organization of this paper is as follows. Section 2 describes our part-based probabilistic model, the recognition method and 2D PCA representation of the image patch. Our combinatorial and statistical methods for image patch selection are explained in section 3 and section 4, respectively, and section 5 presents the results of applying the proposed methods on a benchmark dataset. Section 6 is the conclusion.

2. Part-based probabilistic model

We model an object class as a constellation of image patches from the object, which is similar in spirit to [7], but we model their relative locations to a common reference frame. In doing this, we avoid the problem of not detecting the landmark patch. We assume objects from the same class should always have the same set of image patches detected and these image patches are similar both in their appearance and their relative location to the reference frame. The recognition of an object in an image will be a high probability event of detecting similar image patches sharing a common reference frame. In our work, we use the centroid as the reference frame and use the image patches simultaneously to build a probabilistic model for the object class and the centroid.

2.1 *Model structure*

The model structure is best explained by first considering recognition. Using m observed image patches v_k, $(k = 1, \dots, m)$, from an image V, the problem of estimating the probability $P(O, C|V)$ of object class O and its centroid C given V can be formulated as (assuming independence between the patches and using Bayes' rule):

$$P(O, C|V) = \frac{P(V|O, C)P(O, C)}{P(V)} = P(O, C)\prod_{k=1}^{m} \frac{P(v_k|O, C)}{P(v_k)}. \tag{1}$$

We wish to approximate the probability $P(v_k|O, C)$ as a mixture of Gaussian model using the observed patches from the training data. We simplify this by clustering all the patches selected from the training data into n clusters, A_i, $i = 1, \dots, n$ according to their appearance and spatial information, which is the 2D offset to the centroid C. We can decompose $P(v_k|O, C)$ as

$$\begin{aligned} P(v_k|O, C) &= \sum_{i=1}^{n} P(v_k|A_i)P(A_i|O, C) \\ &= \frac{\sum_{i=1}^{n} P(v_k|A_i)P(O, C|A_i)P(A_i)}{P(O, C)}. \end{aligned} \tag{2}$$

Substituting (2) in (1), we get

$$P(O, C|V) \propto \prod_{k=1}^{m} \frac{\sum_{i=1}^{n} P(v_k|A_i)P(O, C|A_i)P(A_i)}{P(v_k)}. \tag{3}$$

While performing recognition, $P(v_k)$ can be ignored. Assuming that $P(C)$ and $P(O)$ are independent, we have

$$P(O, C|V) \propto \prod_{k=1}^{m} \sum_{i=1}^{n} P(v_k|A_i)P(O|A_i)P(C|A_i)P(A_i). \tag{4}$$

2.2 *Learning*

The task of learning is to estimate each term in (4) from the training data. We concatenate the image patches' appearance and spatial vectors as features in the image patches clustering process. Since the resulting clusters contain similar features, we can assume image patches from one cluster will follow normal distribution in both appearance and spatial subspaces. By calculating the sample mean and sample covariance matrix of the subspaces of these clusters, we can approximate the probability of v_k and C for each cluster $A_i, i = 1, \ldots, n$. We use μ_i^v and μ_i^c to denote the sample means for v_k and C, respectively, and Σ_i^v and Σ_i^c to denote the sample covariances for v_k and C, respectively. Then for cluster A_i we have $P(v_k|A_i) \sim \mathrm{N}(v_k|\mu_i^v, \Sigma_i^v)$ and $P(C|A_i) \sim \mathrm{N}(C|\mu_i^c, \Sigma_i^c)$.

The rest of the terms in (4) can be approximated using the statistics from each of the cluster $A_i, i = 1, \ldots, n$. If the Cluster A_i has n_i points of which n_{ij} belong to Class O_j, we can estimate the following: $P(A_i) = n_i / \sum_{i=1}^{n} n_i$ and $P(O_j|A_i) = n_{ij}/n_i$[†].

2.3 *Recognition*

Recognition proceeds by first detecting and selecting image patches, and then evaluating the probability of the event of detecting object features sharing a common reference frame, as described in section 2.1. By calculating the probability and comparing it to a threshold, the presence or the absence of the object in the image may be determined.

Equation (4) can be interpreted as a probabilistic voting scheme where each patch casts a weighted vote for the object class and centroid given its similarity to each of the clusters. This formulation extends to handle scale variations by considering each pair of patches instead of each individual patch.

2.4 *Image feature representation*

The image patch feature concatenated from appearance and spatial information could be a high dimension vector. Usually PCA is applied to reduce the dimension while retaining much of the information.

Recently Yang *et al.* [18] have proposed 2D PCA for image representation. In contrast with PCA, 2D PCA is based on 2D image matrices rather than 1D vector such that the image matrix does not need to be transformed into a vector before feature extraction. Instead, the original image matrices can be used to directly construct the image covariance matrix. As a result, the size of the image covariance matrix using 2D PCA is much smaller than that of PCA. Thus, the 2D PCA method is better than PCA in the following ways: (1) it is easier to evaluate the covariance matrix accurately to calculate the eigenvectors; (2) it also takes less time because it deals with much smaller matrices. In this paper, we have experimented with both approaches to evaluate the advantage of using 2D PCA over traditional PCA in patch representation.

3. Combinatorial selection of characteristic image patches

In an object recognition task where an image is represented as a constellation of image patches, often many patches correspond to the cluttered background. If such patches are used to build

[†] It must be remarked that this model extends to modelling multiple object classes directly; however, since our problem consists of only one class, we have $P(O_j|A_i) = 1$.

the model for object class recognition, they will adversely affect the recognition rate. In this section, we formulate the problem of finding the set of image patches that can help in discriminating between image with and without the target object as an combinatorial optimization problem on a multipartite graph. We first introduce some notations which will help in formalizing this problem. Suppose we are given a set $V^+ = \{V_1^+, V_2^+, \ldots, V_p^+\}$ of p images (positive class) containing the instances of the target object, and a set $V^- = \{V_1^-, V_2^-, \ldots, V_n^-\}$ of n images (negative class) which do not contain the target object. Recall that any arbitrary image is represented as a set of m salient image patches, so the image ith from the positive class can be denoted as $V_i^+ = \{v_{i1}^+, v_{i2}^+, \ldots, v_{is}^+, \ldots, v_{im}^+\}$, where v_{is}^+ is the sth image patch. Further, we also use V^+ to denote the set of all patches in V_1^+ through V_p^+, i.e. $V^+ = \cup_{\ell=1}^{p} V_\ell^+$; similarly, $V^- = \cup_{\ell=1}^{n} V_\ell^+$. The usage will become clear from the context.

We are interested in finding the subset of image patches from the set V^+ which are very similar to each other and, at the same time, distant from those in the set V^-. Furthermore, while finding image patches that characterize the target object, it is best to focus on similarities between image patches across different instances of the target object, rather than similarities between patches from the same image although they may be very similar. These two informal requirements can be expressed in a multipartite graph representation of the similarities between image patches from different images, as shown in figure 1. The right part of this figure shows an undirected edge weighted vertex weighted multipartite graph, $G = (V^+, E, W, N)$, with p partite sets V_1^+ through V_p^+ so that, as described earlier, $V^+ = \cup_{\ell=1}^{p} V_\ell^+$. The edges in the set $E \subseteq \cup_{i \neq j} V_i^+ \times V_j^+$ represent similarity between the image patches from different images while the weight w_{ab} on the edge connecting the vertices corresponding to the patches a and b represents the strength of their similarity. Each vertex in V^+ is also associated with a weight $N : V^+ \to \mathbb{R}^+$ which reflects its aggregated similarity to image patches in V^-. For any vertex $i \in V^+$, its vertex weight $N(i)$ is calculated as $N(i) = \sum_{s \in V^-} m_{is}^2$, where m_{is} is the similarity between image patch i and the image patch s from a negative image.

We consider the situation where the negative images in the training set do not contain any instance of the target object, and the positive images contain exactly one instance of the target object. Of course, it is possible to model more complex situations where the positive images contain multiple instances of the target object. However, we have focused on modelling the simpler situation. We now formulate the optimization problem for finding the subset of image patches which are characteristic of positive images and distant from patches in the negative

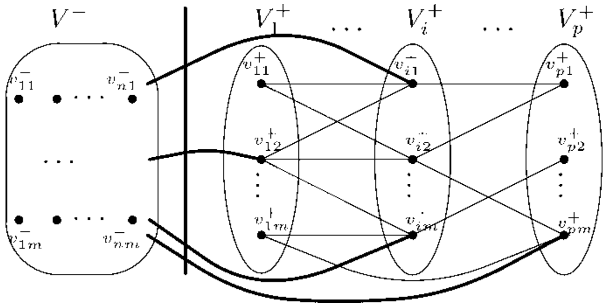

Figure 1. A multipartite graph representation for expressing similarity relationships between the image patches. Ellipse corresponding to V_i^+ represents the ith instance of target image, and the m points inside this ellipse represent the image patches from this image. The patches from the images that do not contain the target object are represented inside the oval V^- without distinguishing between the images of those patches. The straight lines connecting the image patches across different instances of images represent the weighted similarity between them, while the thick curved lines represent the aggregated (weighted) similarity between an image patch from the positive image to all image patches in the negative class. For visual clarity, weights are not shown on the edges.

images. In other words, we want to find a subset $H \subseteq V^+$ (so, $H = \cup_{\ell=1}^{p} H_\ell$, where $H_\ell \subseteq V_\ell^+$) of image patches from the positive images in which patches are very similar to each other and at the same time different from image patches in the negative images. To achieve this, any subset H is assigned a score $F(H)$ which measures the degree of similarity between the patches from different images in H and also their distinction from patches in V^-. This score is designed to be higher, as described later, for desirable subsets. The best subset, H^*, is the globally optimal solution for the following criterion:

$$H^* = \arg \max_{H \subseteq V^+} F(H) \tag{5}$$

The score $F(H)$ is defined using a linkage function $\pi(i, H)$ which measures the degree of similarity of the patch i to patches from the other images in H.

$$F(H) = \min_{i \in H} \pi(i, H) \tag{6}$$

Thus, the score $F(H)$ for the subset H is a linkage function value, $\pi(i, H)$, for the least similar patch in H. Then, the optimal solution H^* described in (5) corresponds to the subset of image patches where the similarity of the least similar patch is maximum.

The design of the linkage function is critical for a suitable problem formulation. It must be remarked that we only have the pairwise similarities between the image patches from different images and using this we must design the function $\pi(i, H)$. Also, recall that H is a multipartite subset, $H = \cup_{\ell=1}^{p} H_\ell$ where $H_\ell \subseteq V_\ell^+$ is a subset of patches from the image V_ℓ^+. If w_{ij} is the similarity value between the image patch i from the image $I(i)$ and the image patch j from the image $I(j)$, then the linkage function is defined as:

$$\pi(i, H) = \sum_{\substack{\ell=1 \\ \ell \neq I(i)}}^{p} \left(\sum_{j \in H_\ell} w_{ij}^2 - \sum_{k \in V_\ell^+ \setminus H_\ell} w_{ik}^2 \right) - \beta N(i) \tag{7}$$

where $\beta \in \mathbb{R}^+$ is a constant factor for scaling $N(i)$, the weight associated with the vertex (i), defined as the aggregated similarity of i to the patches from the negative images. This scaling factor β serves to account for any imbalance between the number of positive and negative instances of the target object. The first term $(\sum_{j \in H_\ell} w_{ij}^2)$ in the linkage function aggregates the similarity of the patch i from image $I(i)$ to patches from other images present in H. The second term $\left(\sum_{k \in V_\ell^+ \setminus H_\ell} w_{ik}^2\right)$ estimates how the patch i is related to patches not included in H_ℓ. A large positive value of the linkage function $\pi(i, H)$ indicates that i is very similar to patches in H and different from the patches in the negative images or the patches from the positive images not included in H. According to this definition of the linkage function, the optimal solution, H^*, corresponds to a collection of image patches from different positive images each of which is highly similar to each other (as the least similar patch is highly similar to other patches) and very different from the patches in the negative images. So, such a formulation indeed serves our purpose of selecting characteristic and discriminative image patches.

This combinatorial optimization problem has been studied in [20] and it has been shown that an efficient algorithm exists for finding the global optimal solution H^* if the linkage function $\pi(i, H)$ is monotone increasing. The monotone increasing property requires that the value of the linkage function for the vertex i can only increase when the second argument H increases in a set theoretic sense, i.e. monotone increasing linkage function satisfies the condition: $\pi(i, H) \leq \pi(i, H \cup \{k\})$ for all $i \in H$ and for all $k \in V^+ \setminus H$. Indeed the linkage function defined in (7) satisfies this property. Observe that the third term $\beta N(i)$ is the vertex weight for i and is independent of H, so it does not affect the monotonicity property. Consider the effect of augmenting the subset H, by including $k \notin H$, on the linkage function value for the element i: when k is included in H, the value w_{ik} is deducted from the second term and added to the first term. So, $\pi(i, H \cup \{k\}) - \pi(i, H) = 2w_{ik}^2 \geq 0$, or $\pi(i, H) \leq \pi(i, H \cup \{k\})$.

Algorithm 3.1 (Algorithm for finding $H^*(\,)$)

```
$t \leftarrow 1$;   $H_t \leftarrow V^+$;   $H^* \leftarrow V^+$;
$F(H^*) \leftarrow \min_{i \in V^+} \pi(i, V^+)$
while $(H_t \neq \emptyset)$
do
    $M_t \leftarrow \{\alpha \in H_t : \pi(\alpha, H_t) = \min_{j \in H_t} \pi(j, H_t)\}$;
    $F(H_t) \leftarrow \min_{j \in H_t} \pi(j, H_t)$;
    if $(H_t \setminus M_t) = \emptyset) \vee (\pi(i, H_t) = 0 \ \forall i \in H_t)$
    then
        output $H^*$ as the optimal set and
            $F(H^*)$ as the optimal value.
    else
        $H_{t+1} \leftarrow H_t \setminus M_t$;
        $t \leftarrow t + 1$;
        if $(F(H_t) > F(H^*))$
            then $\{H^* = H_t$;
```

The algorithm for solving this combinatorial optimization problem is given [20], and is described in the pseudocode form in Algorithm 3.1. This iterative algorithm begins by calculating $F(V^+)$ and finds the set M_1 containing the set of vertices from V^+ which have the minimum value of the linkage function, i.e. $M_1 = \{\alpha \in V^+ : \pi(\alpha, V^+) = \min_{j \in V^+} \pi(j, V^+)\}$. The vertices in the set M_1 are removed from V^+ and the set H_2 is constructed as $H_2 = V^+ \setminus M_1$. At this point, the second iteration begins with the calculation of $F(H_2)$ and finds the set M_2. At the iteration t, the algorithm considers the set H_t as the input, calculates $F(H_{t-1})$, finds the subset M_t such that $F(H_{t-1}) = \pi(j, H_{t-1})$, $\forall\, j \in M_t$, and removes this subset from H_{t-1} to produce $H_t = H_{t-1} \backslash M_t$. Finally, the algorithm terminates at the iteration T, when $H_T = \emptyset$ or when $\pi(i, H_T) = 0 \,\forall\, i \in H_T$. It outputs H^* as the subset H_j with the smallest j such that $F(H_j) \geq F(H_l) \,\forall\, l \in \{1, 2, \ldots, T\}$.

This problem formulation gives us one subset of similar image patches from the positive images and likely corresponds to some characteristic in the target object in those images. However, an object often has multiple salient characteristics, and these disjoint subsets of patches corresponding to different characteristics of the target object can be found by removing the optimal solution H^* from the set V^+ and solving the optimization problem on the reduced set $V^+ \setminus H^*$. Thus, sequentially solving this optimization problem until we get optimal solutions with large values allows us to find the desired groups of image patches.

A complexity analysis of the method can be found in [20]. It runs in $O(|E| + |V| \log |V|)$ time, where E and V are the set of edges and vertices, respectively, in the graph.

4. Statistical image patch selection

In the previous section we had focused on a combinatorial optimization formulation for finding subsets of patches characterizing the images from the positive class, and hopefully corresponding to salient regions in the target object. In this section, we formulate the same problem in a statistical framework by selecting those image patches from the positive images which consistently appear in multiple instances of the positive images but only rarely appear in the negative images (barring some hypothetical and pathological cases). Intuitively, if an individual image patch from a positive image performs well in recognizing the images of the target object, a combination of a number of such image patches is likely to enhance the overall performance. This is because the individual classifiers, although weak, can synergistically guide the combined classifier in producing statistically better results.

Our approach is different from the Boosting method [11]. Boosting is originally a way of combining classifiers and its use as feature selection is an overkill. In contrast, our statistical method does not boost the previous stage but filters out the over-represented and undesirable clusters of patches corresponding to background. In spirit, our approach is similar to [21]. We formalize this intuitive statistical idea in the following straightforward yet effective method for selecting the characteristic image patches.

We select an image patch $v \in V^+$ from the positive images V^+ in the training data if it is able to discriminate between the positive and negative images in the evaluation data, $V_e = \{V_e^+, V_e^-\}$ with a certain accuracy. A complete description of this method requires describing the classification method using a single image patch and the accuracy threshold. For classifying an image $\mathcal{V} \in V_e$ in the evaluation set, using a single image patch $v \in V^+$, we first calculate the distance, $D(\mathcal{V}, v) = \min_{\nu \in \mathcal{V}} d(\nu, v)$, between $\mathcal{V}$ and v defined as the Euclidean distance between v and the closest image patch from $\mathcal{V}$. For classifying the images in the evaluation data, we use a threshold, t on distance $D(\mathcal{V}, v)$; if $D(\mathcal{V}, v) < t$, the image $\mathcal{V}$ is predicted to contain the target object, otherwise not. Accordingly we can associate an error function, $\mathcal{E}r(\mathcal{V}, v, t)$ (defined below in (8)), which assumes a value 1 if and only if the classifier makes the mistake .

$$\mathcal{E}r(\mathcal{V}, v, t) = \begin{cases} 0, & \text{if } (D(\mathcal{V}, v) < t \ \wedge \ \mathcal{V} \in V_e^+) \vee \\ & \quad (D(\mathcal{V}, v) \geq t \ \wedge \ \mathcal{V} \in V_e^-) \\ 1, & \text{otherwise} \end{cases} \tag{8}$$

Clearly, the performance depends on the parameter t, so we find an optimal circular region of radius t_v around v which minimizes the error rate of the classifier on the evaluation data. Finally, only those image patches from the positive images are selected which have recognition rate above a threshold, θ. A description of this algorithm, in the form of a pseudocode, is given in Algorithm 4.1. This algorithm takes the positive image patches V^+, patches from the evaluation data V_e, and the threshold θ as input and outputs $\widehat{H} \subseteq V^+$, the subset of selected image patches.

ALGORITHM 4.1 (select patches, $\widehat{H}(V^+, V_e, \theta)$)

```
)$\widehat{H} \leftarrow \emptyset$;
for each $v \in V^+$
do
    for each $\mathcal{V} \in V_e$
        do $D(\mathcal{V}, v) = \min_{\nu \in \mathcal{V}} d(\nu, v)$;
    $t_v \leftarrow \arg\min_{t \in \mathbb{R}^+} \sum_{\mathcal{V} \in V_e} \mathcal{E}r(\mathcal{V}, v, t)$
    $err \leftarrow \frac{1}{|V_e|} \sum_{\mathcal{V} \in V_e} \mathcal{E}r(\mathcal{V}, v, t_v)$
    if $(err < \theta)$
        then $\widehat{H} \leftarrow \widehat{H} \cup \{v\}$
```

5. Experiment

5.1 *Data set*

The experiment was carried out using Caltech database.[†] This database contains four classes of objects: motorbikes, airplanes, faces, and car rear end which have to be distinguished from image in the background data set, also available in the database. Each object class is represented by 450 different instances of the target object, which were randomly and evenly split into training and testing images. Of the 225 positive images set aside for selecting the characteristic image patches, 175 were used as the training images and the remaining 50 were spared to be used as evaluation data. In addition, the evaluation data also consisted of 50 negative images. The combinatorial and the statistical methods used the training and evaluation images slightly differently – while the combinatorial method selected image patches by simultaneously analysing 175 positive (remaining 50 positive images from the evaluation data were not used in this method) and 50 negative images from the evaluation data, the statistical method selected patches from 175 positive images by judging their performance on 50 positive and 50 negative images in the evaluation data.

5.2 *Image patch detection and the intensity representation*

We use region based detector [13] for detecting informative image patches. We perform normalization for intensity and rescaled the image patches to 11 × 11 pixels, thus representing them as 121 dimension intensity vectors. Then we tried with both PCA and 2D PCA on these vectors to get a more compact 18 dimension intensity representation.

5.3 *Experimental setting*

We extracted 100 image patches for each of the 175 training images, and 100 evaluation images. Following this, we applied the combinatorial and statistical methods individually and in a combination for removing the image patches from the background.

For the combinatorial image patch selection, we converted the Euclidean distance, $d(i, j)$ between the features from the patches i and j from different images to the similarity value $w_{ij} = d_{\max} - d_{ij}$. The similarity values larger than an empirically calculated threshold were removed to convert the complete multipartite graph into a sparse graph containing 10% of the original edges. The same similarity threshold was used for considering similarity between patches from positive and negative images. We used $\beta = 3.0$ in the linkage function (7) to account for the imbalance in the number of positive images (175) and the negative images (50) used in the training data.

For statistical image patch selection, we built a simple classifier from each image patch in the training images and selected the one which led to a classifier with classification error rate less than 24%, an empirically calculated value.

We also used a sequential combination of the two methods. Figure 2 shows results from the three methods (statistical, combinatorial and their combination) for selecting image patches. The results show that both approaches are successful in removing a significant number of patches corresponding to background and the sequential combination of the methods performs the best.

[†] http://www.vision.caltech.edu/html-files/archive.html

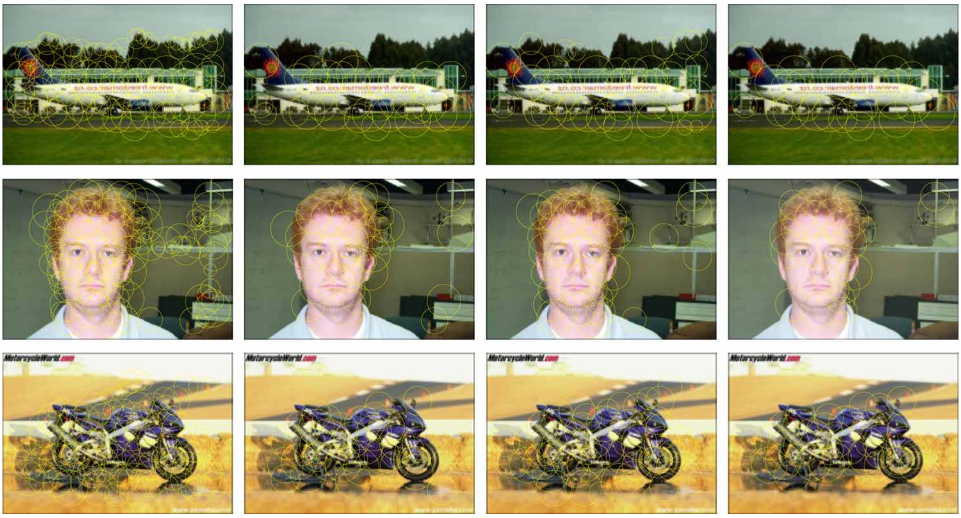

Figure 2. Image patch selection. The image patches are shown using yellow (pale grey) circles on the images. The first column shows the image patches extracted by Kadir and Bradys' [13] feature detector. The second and third columns show image patches selected by combinatorial and the statistical methods, respectively. The patches selected by the sequential combination of the method are shown in column four.

After the image patch selection process, we computed the centroid for each object in the image. We used 2D offset between the image patch and the object centroid as the spatial feature for the image patch and concatenated it with the intensity feature vector as the feature representation for each image patch. We then used k-means algorithm for clustering them into 70 clusters (this number was empirically chosen) and calculated the mean and covariance for them.

5.4 *Experimental results*

In the testing phase, we used Kadir and Bradys' [13] feature detector for extracting the image patches. Then we calculated the probability of the centroid of a possible object in the image as an indicator of its presence.

Figure 3 shows the computationally estimated frame for the object along with the image patches which contributed towards estimating this frame. Observe that the estimated frame was mainly voted by the image patches located on the object. It also shows some examples of misclassification. There are two major reasons for such misclassification. The first is the presence of multiple target objects in the image, as shown in the airplane example. In this scenario, there is no centroid which gets a strong probability estimation from the matched parts. The second is poor illumination conditions which seriously limits the number of initial image patches extracted from the object, as illustrated by the face example.

We compared our result to the state of the art results from [10] and [22]. Table 1 summarizes the recognition accuracy at the equal ROC points (point at which the true positive rate equals one minus the false positive rate) of our different approach: no part selection with PCA, statistical selection with PCA, statistical selection with 2D PCA, combinatorial selection with PCA, combination of combinatorial and statistical methods with PCA and results from other recent methods. This shows that the result from 2D PCA representation is similar to that from PCA. We also see that both the proposed methods perform well in recognition and their combination improves the recognition rates even further and yields better results, quite often

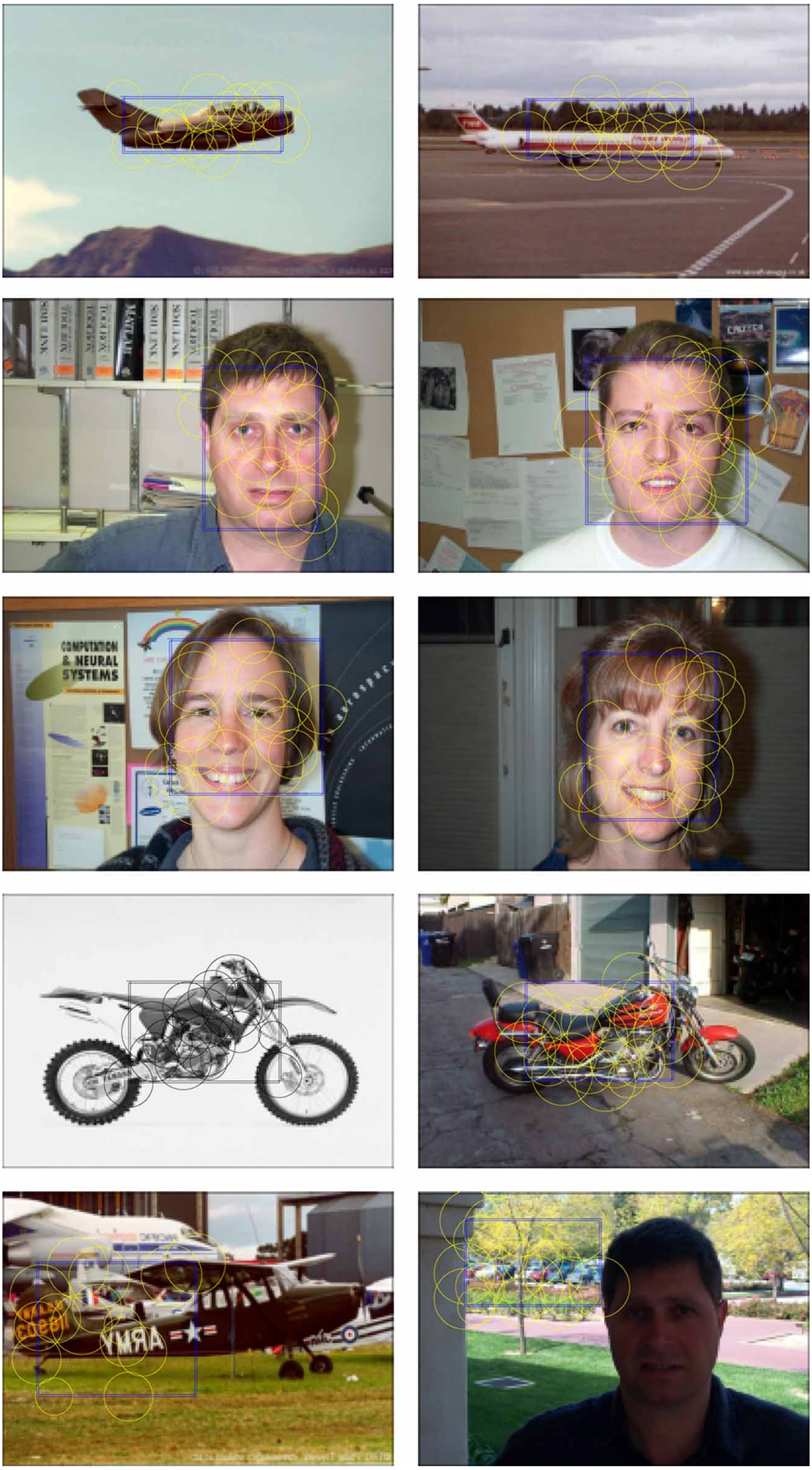

Figure 3. This figure demonstrates the estimation of object frame in some typical testing image using part based probabilistic model. The estimated centroid is indicated by a rectangle. All the image patches contributed to this estimation are indicated by yellow (pale grey) circles. The bottom row of the images are some misclassification examples.

Table 1. Equal ROC performance of our different approaches and other recent methods.

	No	Statistical	Statistical	Combinatorial	Combinatorial		
	selection	method	method	method	method	Fergus *et al.*	Opelt *et al.*
Dataset	with PCA	with 2D PCA	with PCA	with PCA	with PCA	[10]	[22]
Airplane	54.2	95.8	94.4	88.9	95.8	90.2	88.9
Motorbike	67.8	93.7	94.9	92.9	95.8	92.5	92.2
Face	62.7	97.3	98.4	97.6	98.9	96.4	93.5
Car (rear)	65.6	98.0	96.7	97.8	99.3	90.3	n/a

by a significant margin, than previous methods, which reports equal ROC performance using this data set.

6. Conclusion

We have presented a combinatorial and a statistical method for selecting informative image patches for part-based object detection and class recognition. Both of these methods when used alone and in combination yield competitive recognition rates, and surpass the performance of many existing methods. Although these methods have been demonstrated in the context of image patch selection, they are general methods suitable for selecting a subset of features in other applications. A natural extension of this method is by integrating the auxiliary information regarding spatial arrangement between image patches; one way for doing this is currently under investigation. In the future, we intend to further develop and disseminate this framework as a general method for selecting features by automatically determining various hyper-parameter, which are currently empirically calculated.

References

[1] Wolfson, H.J. and Rigoutsos, I., 1997, Geometric hashing: An overview. *IEEE Computational Science & Engineering*, **4**, 10–21.

[2] Viola, P. and Jones, M., 2002, Robust real-time object detection. *International Journal of Computer Vision*, **57**, 137–154.

[3] Schneiderman, H. and Kanade, T., 2000, A statistical method for 3D object detection applied to faces and cars. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2000*, pp. 45–51.

[4] Fischler, M. and Elschlager, R., 1973, The representation and matching of pictorial structures. *IEEE Transaction on Computers*, **22**, 67–92.

[5] Lowe, D.G., 1999, Object recognition from local scale-invariant features. *Proceedings of the International Conference on Computer Vision (ICCV)*, Corfu, pp. 1150–1157.

[6] Schmid, C. and Mohr, R., 1997, Local grayvalue invariants for image retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, **19**, 530–535.

[7] Weber, M., Welling, M. and Perona, P., 2000, Unsupervised learning of models for recognition. *European Conference on Computer Vision (ECCV) 2000*, pp. 18–32.

[8] Agarwal, S. and Roth, D., 2002, Learning a sparse representation for object detection. *European Conference on Computer Vision (ECCV) 2002*, pp. 113–130.

[9] Borenstein, E. and Ullman, S., 2002, Class-specific, top-down segmentation. *European Conference on Computer Vision (ECCV) 2002*, pp. 109–124.

[10] Fergus, R., Perona, P. and Zisserman, A., 2003, Object class recognition by unsupervised scale-invariant learning. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2003*, pp. 264–271.

[11] Torralba, A.B., Murphy, K.P. and Freeman, W.T., 2004, Sharing visual features for multiclass and multiview object detection. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2004*.

[12] Fergus, R., Perona, P. and Zisserman, A., 2005, A sparse object category model for efficient learning and exhaustive recognition. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2005*.

[13] Kadir, T. and Brady, M., 2001, Scale, saliency and image description. *International Journal on Computer Vision*, **45**(2), 83–105.

[14] Baumberg, A., 2000, Reliable feature matching across widely separated views. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2000*, pp. 774–781.

[15] Tuytelaars, T. and Van Gool, L.J., 2000, Wide baseline stereo matching based on local, affinely invariant regions. *The British Machine Vision Conference (BMVC) 2000*.

[16] Schaffalitzky, F. and Zisserman, A., 2002, Multi-view matching for unordered image sets, or how do I organize my holiday snaps? *European Conference on Computer Vision (ECCV) 2002*, pp. 414–431.

[17] Felzenszwalb, P.F. and Huttenlocher, D.P., 2005, Pictorial structures for object recognition. *International Journal of Computer Vision*, **61**, 55–79.

[18] Yang, J., Zhang, D., Frangi, A.F. and Yang, J.-Y., 2004, Two-dimensional PCA: a new approach to appearance-based face representation and recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, **26**, 131–137.

[19] Leung, T.K. and Malik, J., 1999, Recognizing surfaces using three-dimensional textons. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 1999*, pp. 1010–1017.

[20] Vashist, A., Kulikowski, C., and Muchnik, I., 2005, Ortholog clustering on a multipartite graph. *Proceedings of Algorithms in Bioinformatics (WABI). Lecture Notes in Computer Science* (Berlin: Springer-Verlag), Vol. 3629, pp. 328–340.

[21] Dorkó, Gy. and Schmid, C., 2003, Selection of scale-invariant parts for object class recognition. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2003*, pp. 634–640.

[22] Opelt, A., Fussenegger, M., Pinz, A. and Auer, P., 2004, Weak hypotheses and boosting for generic object detection and recognition. *European Conference on Computer Vision (ECCV) 2004*, pp. 71–84.