

# Discriminative Patch Selection using Combinatorial and Statistical Models for Patch-Based Object Recognition

Akshay Vashist<sup>1</sup>, Zhipeng Zhao<sup>1</sup>, Ahmed Elgammal<sup>1</sup>, Ilya Muchnik<sup>1,2</sup>, Casimir Kulikowski<sup>1</sup>

<sup>1</sup> Department of Computer Science, <sup>2</sup>DIMACS

Rutgers, The State University of New Jersey, Piscataway, NJ 08854, USA

{vashisht, zhipeng, elgammal, kulikows}@cs.rutgers.edu, muchnik@dimacs.rutgers.edu

## Abstract

*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 for object class recognition, they will adversely affect the recognition rate. In this paper, we present a two stage method for selecting 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 selection is done using a novel combinatorial optimization formulation on a weighted multipartite graph representing similarities between images patches across different instances of the target object. The following stage is a statistical method for selecting those images patches from the positive images which, when used individually, have the power of discriminating between the positive and negative images in the evaluation data. The individual methods have a performance competitive with the state of the art methods on a popular benchmark data set and their sequential combination consistently outperforms the individual methods and most of the other known methods while approaching the best known results.*

## 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 the 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 template search-based methods, and patch-based methods. 3D model-based methods (e.g. [22]) 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 the best match to an object “template” where the object template can be learned from training data and act as a local classifier [20, 16]. Such approaches are highly successful in modeling objects with wide within-class appearance variations such as in the case of face detection [20, 16] but they are limited when the within-class geometric variations are large, such as in detecting a motorbike.

In contrast, object recognition based on dense local “invariant” image features have shown a lot of success recently [8, 11, 15, 21, 1, 3, 6, 17, 7] for objects with large within-class variability in shape and appearance. In such approaches objects are modeled as a collection of patches or local features and the recognition is based on inferring object class based on similarity in patches’ appearance and their spatial arrangement. Typically, such approaches find interest points using some operator such as [9] 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) [11], entropy-based scale invariant features [9, 6] and other local features which exhibit affine invariance such as [2, 18, 14]. Other approaches that model objects using local features include graph-based approaches such as [5].

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 [21], because the initial set of features is large, and often features are redundant or correspond to the clutter 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. This is precisely the focus of this paper: selecting the “best” features from the already extracted image features that are both exclusive and well represented in different images of the target object.

Unsupervised selection of discriminative patches is a fundamental problem for learning object models. Weber *et al.* [21] suggested the use of clustering to find common object patches and to reject background clutter from the positive training data. In such an approach large clusters are retained as they are likely to contain patches on the target object. A similar approach has been used in [10]. However, there is no guarantee that a large cluster will contain only object patches. Since the success of recognition is based on using many local features, such local features typically correspond to low level features rather than actual high level object parts. In this paper we introduce two complementary approaches to select discriminative object patches from a pool of patches extracted from the training images.

**Contributions:** We introduce two novel approaches for unsupervised selection of discriminative patches that explicitly takes into account the contrast between positive and negative examples in the training data. 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. Finally, we use a probabilistic Bayesian approach for recognition where the object model does not need a reference patch [6]. Instead, object patches are related to a common reference frame.

The organization of this paper is as follows. Section 2 formulates the problem of finding distinctive image patches from the positive images as a combinatorial optimization problem and a statistical problem, which are our main foci and are described in sections 3 and 4, respectively. Section 5 describes our recognition method and section 6 presents the results of applying the proposed methods on a benchmark dataset. Section 7 is the conclusion.

## 2. Problem Formulation and Framework

The problem we address can be stated as: Given a pool of local features (patches) extracted from a set of labelled training images containing positive and negative images of the target class, how can we choose (in an unsupervised way) the best features representing the object. As feature extraction is not the primary focus of our investigation, we used the popular Kadir and Brady’s feature extractor [9] to get the initial set of image patches for representing an im-

age. Also we used a probabilistic method similar (in spirit) to [6, 12] for modeling the object class and for recognition. These choices allow us to focus on selecting the distinctive image patches from the positive class. The proposed selection algorithms are not tied by any means to the chosen feature extractor or recognition algorithm used in this paper and therefore can be used with any features and any recognition algorithm.

Undoubtedly, there can be many approaches for selecting a collection of image patches from images in the training data. Naively, it seems plausible to select patches from both the negative images and positive images, and classify a test image in the class to which it is closest. However, the space of negative images, devoid of any instance of the target image, is prohibitively large to allow any generalization on the negative class. So, one should rather train the classifier on the positive images using patches which are common to most of the positive images. This is based on the assumption that salience features of the target object will be present and captured from most of the positive images, and form a good representation for it. A potential side effect of focusing entirely on the positive images is the selection of undesirable patches corresponding to the background. A solution to this is by simultaneously considering the positive and negative images for selection the image patches representing both the saliencies of the target object while at the same time being exclusive/discriminative to the positive class. We present two approaches for realizing such a selection - a combinatorial approach and a statistical approach.

The combinatorial approach involves finding the subset of similar image patches shared in most of the positive images. To endow a discriminative power to the selected patches, we also consider their similarity to patches from the negative images. Thus, we wish to find such a subset of patches from the positive images where every patch is distant from the negative patches in the training data but highly similar to patches (in the selected subset) from other images in the positive images. Such selection is formulated as a combinatorial optimization problem on multipartite graph and is described in details in section 3.

Whereas the above described approach is a subset selection approach, the statistical approach analyzes an individual image patch from positive patches in an attempt to find patches which are both detectable and distinctive to the object class. This is achieved by determining if the patch has the power of discriminating between the positive and negative images in the evaluation data. Every patch from the positive training data is evaluated based on its performance in separating the positive and negative images in the evaluation data which was set aside from the training data a priori. If the image patch accurately predicts a significant number of evaluation images, it is selected. A detailed description of this procedure is provided in section 4.

The two approaches complement each other - apart from the obvious combinatorial and statistical nature of the formulation, the first does not involve any evaluation while evaluation is an integral part of the latter approach. Combinatorial selection mostly focuses on selecting patches which are over-represented in an ensemble of images of the target object, in contrast the statistical selection focuses on finding class-specific patches. From this perspective, combinatorial selection can be characterized as a method which has a low probability of losing a typical patch present in an image of the target object. On the other hand, the statistical selection is a method for eliminating, with high probability, patches which do not strongly belong to the target object. Due to their complementarity, one expects to gain by combining them. One way of combining them retaining advantages of both the methods is to initially use the combinatorial method for selecting the over-represented patches, and subsequently use the statistical method for filtering out the patches (from those selected at the first stage), which are not specific to the target object.

### 3. Combinatorial selection of characteristic image patches

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 a 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  $V^+ = \{V_1^+, V_2^+, \dots, V_p^+\}$  of  $p$  images (positive class) containing the instances of the target object, and a set  $V^- = \{V_1^-, V_2^-, \dots, V_n^-\}$  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  $i^{th}$  from the positive class can be denoted as  $V_i^+ = \{v_{i1}^+, v_{i2}^+, \dots, v_{is}^+, \dots, v_{im}^+\}$ , where  $v_{is}^+$  is the  $s^{th}$  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 conveniently expressed in a multipartite graph representation of the similarities between image patches from different images, as shown in Fig. 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  $N : V^+ \rightarrow \mathbb{R}^+$  which reflects its aggregated similarity to images 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.

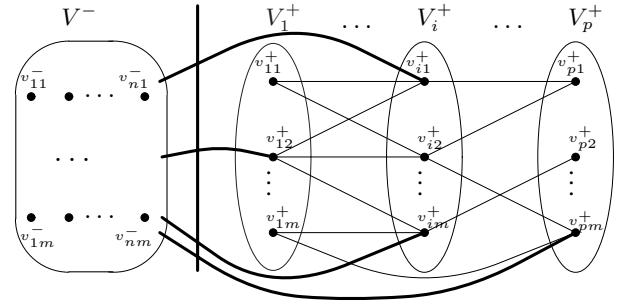


Figure 1. A multipartite graph representation for expressing similarity relationships between the image patches. Ellipse corresponding to  $V_i^+$  represents the  $i^{th}$  instance of target image, and the  $m$  points inside this ellipse represent the image patches from this image. The patches from 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 images 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 positive image to all image patches in the negative class. For visual clarity, weights are not shown on the edges.

We consider the situation where the negative the images in 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 modeling 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 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 score the  $F(H)$  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.

$$H^* = \arg \max_{H \subseteq V^+} F(H) \quad (1)$$

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  $H$ .

$$F(H) = \min_{i \in H} \pi(i, H) \quad (2)$$

Thus, the score  $F(H)$  for the subset  $H$  is linkage function value,  $\pi(i, H)$ , for the least similar patch in  $H$ . Then, the optimal solution  $H^*$  described in (1) 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, *i.e.*  $H = \cup_{\ell=1}^p H_\ell$  where  $H_\ell \subseteq V_\ell^+$  is a subset of patches from the image  $V_\ell^+$ . If  $w_{ij}$  is the similarity value between the image patch from  $i$  from the image  $I(i)$  and the image patch  $j$  from the image  $I(j)$ , then the linkage function is defined as:

$$\pi(i, H) = \sum_{\substack{\ell=1 \\ \ell \neq I(i)}}^p \left( \sum_{j \in H_\ell} w_{ij}^2 - \sum_{k \in V_\ell^+ \setminus H_\ell} w_{ik}^2 \right) - \beta N(i) \quad (3)$$

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  $\beta$  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 ( $\sum_{k \in V_\ell^+ \setminus H_\ell} w_{ik}^2$ ) estimates how the patch  $i$  is related to patches not included in  $H_\ell$ . 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 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 [19] 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 (3) 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 ← 1; Ht ← V+; H* ← V+;
F(H*) ← mini ∈ V+ π(i, V+)
while (Ht ≠ ∅)
  Mt ← {α ∈ Ht : π(α, Ht) = minj ∈ Ht π(j, Ht)};
  F(Ht) ← minj ∈ Ht π(j, Ht);
  if (Ht \ Mt = ∅) ∨ (π(i, Ht) = 0 ∀ i ∈ Ht)
    then { output H* as the optimal set and
           F(H*) as the optimal value.
  }
  else {
    Ht+1 ← Ht \ Mt;
    t ← t + 1;
    if (F(Ht) > F(H*))
      then { H* = Ht;
    }
  }

```

---

The algorithm for solving this combinatorial optimization problem is given [19], 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  $H_t = H_{t-1} \setminus M_t$ . 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  $F(H_j) \geq F(H_i) \forall i \in \{1, 2, \dots, T\}$ .

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, often an object has multiple salient characteristics, and these disjoint subset 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 [19]. 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, in isolation, those 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, barring a few pathological cases, the individual classifiers, although weak, can synergistically guide the combined classifier in producing statistically better results.

Our approach is different from the Boosting method [17]. 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 [4]. We formalize this intuitive statistical idea in the following straightforward yet effective method for selecting the characteristic image patches, as complementary to the combinatorial selection method, which is the main contribution of this paper.

We select an image patch  $v \in V^+$  from the positive images 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 description 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 con-

tain the target object, otherwise not. Accordingly we can associate an error function,  $\mathcal{E}r(\mathcal{V}, v, t)$  (defined below 4), which assumes a value 1 if and only if the classifier makes a mistake.

$$\mathcal{E}r(\mathcal{V}, v, t) = \begin{cases} 0, & \text{if } (D(\mathcal{V}, v) < t \wedge \mathcal{V} \in V_e^+) \vee \\ & (D(\mathcal{V}, v) \geq t \wedge \mathcal{V} \in V_e^-) \\ 1, & \text{otherwise} \end{cases} \quad (4)$$

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,  $\theta$ . 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  $\theta$  as input and outputs  $\hat{H} \subseteq V^+$ , the subset of selected image patches.

---

**Algorithm 4.1:** SELECT PATCHES,  $\hat{H}(V^+, V_e, \theta)$

---

```

 $\hat{H} \leftarrow \emptyset;$ 
for each  $v \in V^+$ 
  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  $\{ \hat{H} \leftarrow \hat{H} \cup \{v\}$ 

```

---

#### 5. Patches based probabilistic model

Following the selection of characteristic image patches from the positive images, we used a probabilistic method for object class recognition. The selected image patches were used, simultaneously, to build a probabilistic model for the object class and the object reference frame. We assumed that a correctly classified object should also have a good approximated reference frame. In our work, we use centroid as the reference frame. Using the  $m$  observed image patches  $v_k$ , ( $k = 1, \dots, m$ ), the problem of estimating the probability  $P(O, C|V)$  of object class  $O$  and its centroid  $C$  given the image  $V$  can be formulated as (assuming independence between the patches and using Bayes' rule):

$$P(O, C|V) = \frac{P(V|O, C)P(O, C)}{P(V)} = P(O, C) \prod_{k=1}^m \frac{P(v_k|O, C)}{P(v_k)} \quad (5)$$

We wish to approximate the probability  $P(v_k|O, C)$  as a mixture-of-Gaussians 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,

$A_i, i = 1, \dots, n$  and decompose  $P(v_k|O, C)$  as

$$\begin{aligned} P(v_k|O, C) &= \sum_{i=1}^n P(v_k|A_i)P(A_i|O, C) \\ &= \frac{\sum_{i=1}^n P(v_k|A_i)P(O, C|A_i)P(A_i)}{P(O, C)} \end{aligned} \quad (6)$$

Substituting (6) in (5), we get

$$P(O, C|V) \propto \prod_{k=1}^m \frac{\sum_{i=1}^n P(v_k|A_i)P(O, C|A_i)P(A_i)}{P(v_k)} \quad (7)$$

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

$$P(O, C|V) \propto \prod_{k=1}^m \sum_{i=1}^n P(v_k|A_i)P(O|A_i)P(C|A_i)P(A_i) \quad (8)$$

Since the clusters contain similar good features, we can assume that both the patch  $v_k$  and the centroid  $C$  from a cluster follow normal distribution. By calculating the sample mean and the sample covariance of these clusters, we can approximate the probability of  $v_k$  and  $C$  for each cluster  $A_i, i = 1, \dots, n$ . We use  $\mu_i^v$  and  $\mu_i^c$  to denote the sample means for  $v_k$  and  $C$ , respectively, and  $\Sigma_i^v$  and  $\Sigma_i^c$  to denote the sample covariances for  $v_k$  and  $C$ , respectively. Then for cluster  $A_i$  we have  $P(v_k|A_i) \sim N(v_k|\mu_i^v, \Sigma_i^v)$  and  $P(C|A_i) \sim N(C|\mu_i^c, \Sigma_i^c)$ . The rest of the terms in (8), can be approximated using the statistics from each of the cluster  $A_i, i = 1, \dots, n$ . If the Cluster  $A_i$  has  $n_i$  points of which  $n_{ij}$  belong to the Class  $O_j$ , we can estimate the following:  $P(A_i) = n_i / \sum_{i=1}^n n_i$  and  $P(O_j|A_i) = n_{ij}/n_i$ <sup>1</sup>.

Now we can calculate equation (8). The result will give us an estimate for the probability of finding an object class centroid. If it is larger than a threshold, it will indicate the presence of an instance of the object class in the image. Equation 8 can be interpreted as a probabilistic voting where each patch gives 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.

## 6. Experiment

### Data Set:

We applied the proposed image patch selection methods for recognizing images from the Caltech database (<http://www.vision.caltech.edu/html-files/archive.html>). This database contains four classes of objects: motorbikes, airplanes, faces, car rear end which have to be distinguished from image in the background data set, also available in the

<sup>1</sup>It must be remarked that this model can be extended for modeling multiple object classes directly, however, since our problem consists of only one class, we have  $P(O_j|A_i) = 1$ .

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 images patches by simultaneously analyzing 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.

### Image patch detection and the intensity representation:

We used region-based detector [9] for detecting informative image patches. We performed normalization for intensity and rescaled the image patches to  $11 \times 11$  pixels, and thus representing them as a 121 dimension intensity vectors. Then, PCA was applied on these vectors to get a more compact 18 dimension intensity representation.

### 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 were thresholded using an empirically calculated value 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 (3) 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.

After the image patch selection process, we computed

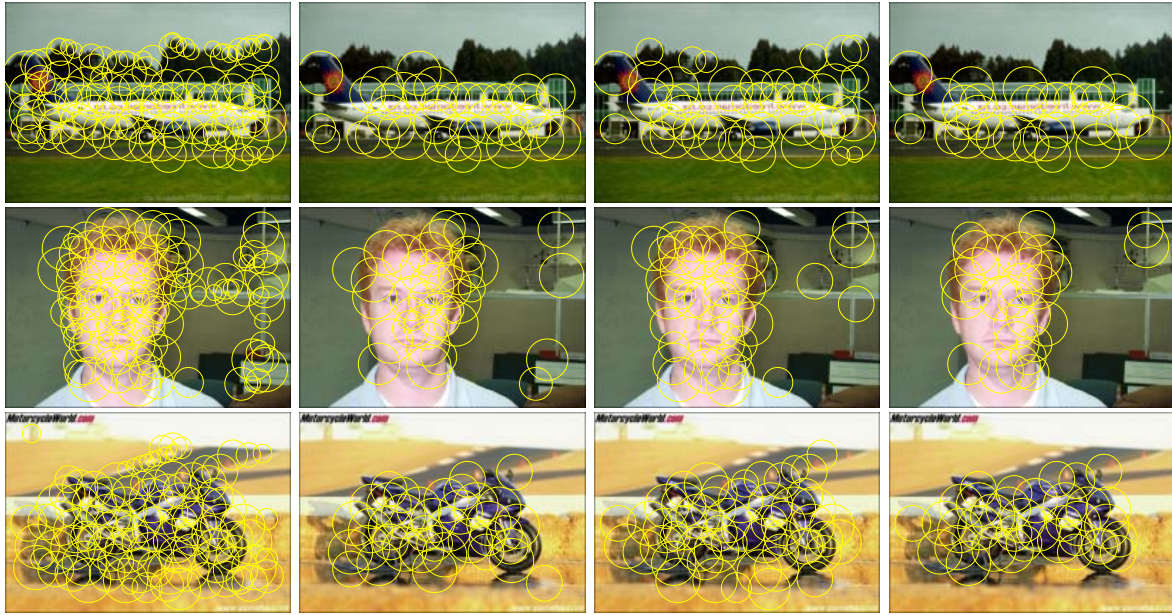


Figure 2. Image patch selection. The image patches are shown using a yellow circle on the images. The first column shows the image patches extracted by Kadir & Brady’s feature detector. The second and third columns show image patches selected by combinatorial and the statistical methods, respectively. The patches selected by the sequential combination of the method are shown in column four.

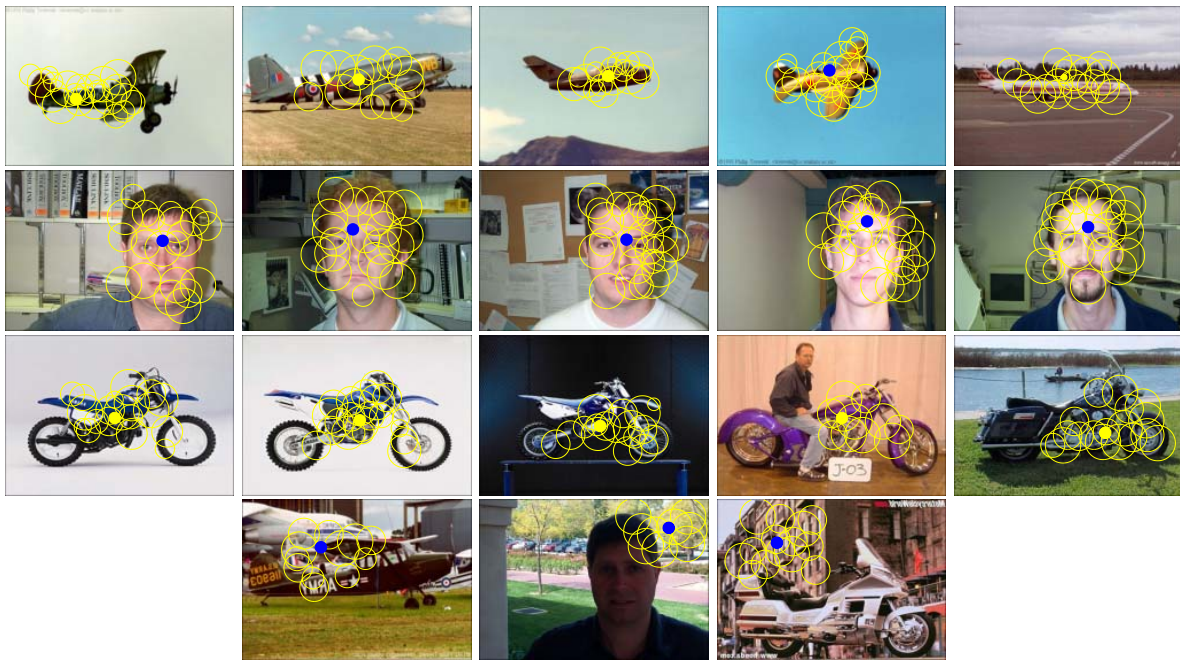


Figure 3. This figure demonstrates the estimation of object centroid in some typical testing image using the sequential combination of combinatorial and statistical approach. The estimated centroid is indicated by a dot with color contrast to the object. All the image patches contributed to this estimation are indicated by yellow circles. The bottom row of the images are some misclassification examples.

the centroid for each object in the image. We used a 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 statistics for them.

**Experimental Results:**

In the testing phase, we used Kadir & Brady’s feature detector for extracting the image patches. Then we calcu-

Dataset	No selection	combinatorial method	statistical method	combination	Fergus [6]	Opelt [13]
Airplane	54.2	88.9	94.4	95.8	90.2	88.9
Motorbike	67.8	92.9	94.9	95.8	92.5	92.2
Face	62.7	97.6	98.4	98.9	96.4	93.5
Car (rear)	65.6	97.8	96.7	99.3	90.3	n/a

Table 1. ROC equal error rates using different methods.

lated the probability of the centroid of a possible object in the image as an indicator of its presence.

Figure 3 shows the computationally estimated centroid for the object along with the image patches which contributed towards estimating this centroid. Observe that the estimated centroid was mainly voted by the image patches located on the object. It also shows some examples of misclassification. There are three 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 patches. 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. Finally, as shown in the motorbike example, when the background is cluttered the initial patches are extracted from all over the image leading and, thereby, confusing the estimator.

We compared our result to the state of the art results from [6] and [13]. Table 1 gives the ROC equal error rates of our different approach and results from other recent methods. This shows our approaches yield comparable or better performance. The results are shown for no selection, combinatorial method only, statistical method only and the sequential combination of combinatorial and statistical methods. These results are also compared to other recent methods reporting equal error rate using this data set. We see that both the proposed methods perform well and their combination improves the recognition rates even further and yielding better results, quite often by a significant margin, than previous methods.

## 7. Conclusion

We have presented a combinatorial and a statistical method for selecting informative image patches for patch-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 these methods is by integrating the auxiliary information regarding spatial arrangement between image patches; one way to

do 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-parameters, which are currently empirically calculated.

## Acknowledgments

Ilya Muchnik was supported by NSF grant CCF - 0325398.

## References

- [1] S. Agarwal and D. Roth. Learning a sparse representation for object detection. In *ECCV*, pages 113–130, 2002. 1
- [2] A. Baumberg. Reliable feature matching across widely separated views. pages 774–781. 1
- [3] E. Borenstein and S. Ullman. Class-specific, top-down segmentation. In *ECCV*, pages 109–124, 2002. 1
- [4] G. Dorkó and C. Schmid. Selection of scale-invariant parts for object class recognition. In *ICCV*, pages 634–640, 2003. 5
- [5] P. F. Felzenszwalb and D. P. Huttenlocher. Pictorial structures for object recognition. *IJCV*, 61(1):55–79, 2005. 1
- [6] R. Fergus, P. Perona, and A. Zisserman. Object class recognition by unsupervised scale-invariant learning. In *CVPR (2)*, pages 264–271, 2003. 1, 2, 8
- [7] R. Fergus, P. Perona, and A. Zisserman. A sparse object category model for efficient learning and exhaustive recognition. In *CVPR*, 2005. 1
- [8] M. Fischler and R. Elschlager. The representation and matching of pictorial structures, 1973. *IEEE Trans on Computer* 22(1): 67-92. 1
- [9] T. Kadir and M. Brady. Scale, saliency and image description. *IJCV*, 2001. 1, 2, 6
- [10] T. K. Leung and J. Malik. Recognizing surfaces using three-dimensional textons. In *ICCV (2)*, pages 1010–1017, 1999. 2
- [11] D. G. Lowe. Object recognition from local scale-invariant features. In *Proc. of the International Conference on Computer Vision ICCV, Corfu*, pages 1150–1157, 1999. 1
- [12] E. Murphy-Chutorian and J. Triesch. Shared features for scalable appearance-based object recognition. In *WACV/MOTION*, pages 16–21, 2005. 2
- [13] A. Opelt, M. Fussenegger, A. Pinz, and P. Auer. Weak hypotheses and boosting for generic object detection and recognition. In *ECCV (2)*, pages 71–84, 2004. 8
- [14] F. Schaffalitzky and A. Zisserman. Multi-view matching for unordered image sets, or "how do i organize my holiday snaps?". In *ECCV (1)*, pages 414–431, 2002. 1
- [15] C. Schmid and R. Mohr. Local grayvalue invariants for image retrieval. *IEEE PAMI*, 19(5):530–535, 1997. 1
- [16] H. Schneiderman and T. Kanade. A statistical method for 3d object detection applied to faces and cars. pages 45–51, 2000. 1
- [17] A. B. Torralba, K. P. Murphy, and W. T. Freeman. Sharing visual features for multiclass and multiview object detection. In *CVPR*, 2004. 1, 5
- [18] T. Tuytelaars and L. J. V. Gool. Wide baseline stereo matching based on local, affinity invariant regions. In *BMVC*, 2000. 1
- [19] A. Vashist, C. Kulikowski, and I. Muchnik. Ortholog clustering on a multipartite graph. In *Proceedings of Algorithms in Bioinformatics (WABI), LNCS*, volume 3629, pages 328–340, 2005. 4, 5
- [20] P. Viola and M. Jones. Robust real-time object detection. *International Journal of Computer Vision* 2002. 1
- [21] M. Weber, M. Welling, and P. Perona. Unsupervised learning of models for recognition. In *ECCV (1)*, pages 18–32, 2000. 1, 2
- [22] H. J. Wolfson and I. Rigoutsos. Geometric hashing: An overview. *IEEE Computational Science & Engineering*, 4(4):10–21, /1997. 1