Combining Low and High Level Features for Object Recognition

Ishani Chakraborty, Ahmed Elgammal

Rutgers University

Email addresses

Abstract

Object recognition algorithms usually identify: 1) point-based features and 2) global structure/geometric shape of an object to solve the recognition problem. In this paper, we propose a framework that integrates these two approaches and brings forth the advantages of both to achieve higher recognition rates. In our framework, the object features are represented using biologically motivated C2 descriptors [1], whereas the global structure information is represented with the skeleton structure of the object silhouette. A novel mechanism that integrates the information derived from different object silhouettes to produce a generic skeleton is implemented. We show that the integrated recognition approach using low level (C2 features) and high level object descriptors (skeletons) is robust under occlusions, for variability within specificity and is invariant to geometric changes. The concept of looking at these two disjoint approaches in a common framework is the first such attempt.

1. Introduction

Recognition of object categories, as opposed to recognition of specific objects, has recently gained a lot of attention in the computer vision community ([1], [6], [8]). The research community has looked at representing objects categories either as a collection of features (point to point matching schemes, [2] and [6]) or as shapes whose contours or other global structures are studied [3]. Very little emphasis is given to the fact that natural recognition (as might be expected in human vision) is indeed a combination of low-level feature identification and a high-level shape understanding [9]. In this paper we consider uniting these two methodologies to solve the recognition problem for object categories. To the best of our knowledge this work is the first of its kind.

As mentioned above, we consider solving the problem of recognizing specific object categories in images (where the input to the recognition system is an image, and its output is a label for the for the object class). Given an image that contains one of the above-mentioned objects, of similar form but different scale, orientation or visibility, we must recognize it as belonging to a correct object class.

Another inherent problem with recognizing generic objects is the variability within specificity. Here, various instances of a category may differ considerably in terms of appearance or pose. The algorithm should be robust enough to identify objects with slightly different part structures. Also, the object of interest has to be recognized in the context of multiple other objects, which may occlude each other.

To this end, we propose a framework and present an algorithm that integrates the knowledge derived from feature and shape information of the object. We show that the approach using the common framework is more adept and robust at solving the above problems. The intuition behind why such an approach would work better is discussed as follows.

We seek a robust mechanism to solve for the object recognition problem using template matching. In the point-to-point technique, matching between images is an unreliable method because low-level appearances can be deceptively similar at various points in images. Also, a single point in one image can match with multiple points in another image. So there has to be a method to streamline the match to reduce it to a one-to-one correspondence. The problem of finding correct correspondences can be resolved using shape matching technique. Here, the geometric shape is captured with its contour or the central-spine/skeleton and matched with a template. However, such a method is very sensitive to slight deviations in the shape and to geometric transformations. Thus both these methods when used individually have their own set of problems. A united framework proposed in this paper makes the system robust under various parameters.

The paper is organized as follows: In section 2, we give the overview of our object recognition framework. The framework consists of two parts, the low-level feature characterization and high level shape representation that is described in detail in section 3 and 4 respectively. In section 5 we provide the algorithm to integrate the two representations. Section 6 describes the matching algorithm. We experimentally evaluate our algorithm in section 7 and conclude in section 8.

2. The United Framework

In our framework, the object features are represented using biologically motivated C2 descriptors [1], whereas the global structure information is represented with the skeleton structure of the object silhouette. The underlying idea in our algorithm is to select point features along the object skeleton so as to create a feature point set under a structural context. The proposed method for object recognition is divided in two phases as listed below. The two phases are also illustrated in Figure 2.

1) Training phase for object models
   a) Extract key points from a segmented image.
   b) Find the skeleton of the segmented image
   c) Choose keypoints which are near/along the skeleton.
   d) Describe the selected keypoints using C2 feature descriptor.

   Each object is characterized by a set of feature points along a generic skeleton that encapsulate the information derived from a set of exemplars of the object.

2) Test Phase
   a) Extract keypoints in the query image.
   b) Describe the keypoints using feature descriptor.
   c) Compare the query image key-point cluster (keypoint and its neighbors) with the object models extracted from the test phase.

Our recognition system extends point-to-point matching by taking into account the appearance and structure of the
neighbourhood of each point. A set of neighbouring points in the image should match with contextually close points in the template. To measure if points are contextually close or not, we use the structure of the generic skeleton as a reference, and distances are measured around it. The formulation of a generic skeleton is a novel mechanism to integrate shape information from different instances of an object. We employ the method of PCA to derive this generic skeleton. By ensuring that neighbouring points on the image match with contextually similar points on the object, the problems that arise from point-to-point correspondences are resolved. This cluster matching technique has been described in detail in the later chapters.

3. Training Phase: Low Level Description

In this section we describe how keypoints are detected using Lowe’s detector and described using C2 features for low level representation.

3.1 Key-point Detection

To identify locations of interest in an image, we have employed Lowe’s scale-space extrema method [2]. It can be summarized in two stages:

1) Scale-space peak selection: The first stage of computation searches for potential interest points by scanning the image over location and scale. This is implemented by using a Difference-of-Gaussian (DoG) function to construct a Gaussian pyramid (Figure 1). Local peaks (maxima or minima locations detected by comparing a pixel to its neighbours) are identified in a series of DoG images constituting the pyramid and are labeled as potential keypoints.

2) Key-point localization: Once a keypoint candidate has been found, a detailed model ensures that nearby data fits with respect to location, scale and principal curvatures. This information allows points to be rejected if found unstable.

Key points of the image from this algorithm are invariant to translation, scaling and rotation and are negligibly affected by noise. As our feature description is based on the key-points that are detected at this stage, the above attributes facilitate for a robust low-level representation. Figure 3 shows the detected keypoints for a few object exemplars.

3.2 Key-point description

The detected key-points (as described in the previous section) are described using C2 features as introduced in the HMAX model [1]. The algorithm is described in Figure 4. This model follows from the quantitative design of object-recognition mechanism in primate cortex, wherein the first few hundreds milliseconds of visual processing follows a feed-forward hierarchy. In such a model, at each stage the receptive field of the neuron tends to get larger along with the complexity of their optimal stimuli. In its basic form, this model consists of four layers of computational units, the simple S units and complex C units. The S units combine their inputs with Gaussian like tuning to increase object selectivity and the C units pool their inputs through a maximum operation, thereby introducing invariance to scale and translation.

We consider image patches centered at the key-point locations and describe each key-point by their C2 features. The computation of feature vectors for small patches is consistent with the primate visual system. During the first few milliseconds of visual processing, the vision system is usually able to focus on a small part of an object, in which it moves its gaze between neighbouring points. This idea would also support the argument about how an image could be recognized even when it is partially occluded. Under occlusions, the visible portion is depicted as a collection of feature points, having a definite local and spatial configuration. Hence, recognition follows from identification of only the visible portion of the image. If a set of points is placed in a spatially coherent manner with respect to one another, then the visual system must recognize it.
4. Training Phase: High Level Shape Representation

In this section we describe how high level shapes are represented using skeletons. We also describe our mechanism to create generic skeletons that would be used as templates in the object recognition algorithm.

4.1 Skeletonization

To obtain a global shape representation, we use the notion of a skeleton (the "central-spine" representation of an object). Such a skeleton would extract the shape information from a silhouette of an object. In this paper we compute the skeleton by simulating the grassfire flow as a Hamilton-Jacobi equation [3].

The key idea is to measure the net outward flux of a vector field per unit area, and to detect locations where a conservation of energy principle is violated. Locations where the flux is negative and hence energy is lost, corresponds to sinks or skeletal points. Hence, a threshold on the divergence map yields a close approximation of the skeleton. However, for a continuous skeleton a homotopy-preserving algorithm is applied. The main idea there is to incorporate a homotopy preserving thinning process in a rectangular lattice, where the removal of points from the shape is guided by their divergence values. Such a skeleton is robust and accurate, has low computational complexity and preserves topology. Figure 5 illustrates the skeletons for silhouettes of objects.

4.2 Formulation of a generic skeleton

Any object would have deformations, scaling and occlusions in its parts that would lead to different skeletons. However we still expect objects to have a generic structure that we want to capture using a generic skeletal curve (in simple terms an average skeleton). In this section we discuss how the generic structure can be formed from skeletons of training images.

To illustrate this concept we consider the skeletonization of cow images. Figure 6 plots all the different skeletons generated from the training images. Clearly we see that although there is significant variation in the skeletal structure within different object parts, when looked at the global scale there is a distinct shape associated with the skeletons. The above illustration forms our intuition that each object category has a general shape that remains universal for all its prototypes. To exploit this geometrical structure of objects and unify them on a common frame of reference, we must map them on a single space. This map would provide us an anchorage on which we can position the feature points collected from all the prototypes of an object. To map the skeleton curves from the training images onto a common frame, we find the distance transform on the skeletal image points and perform Principal Component Analysis (PCA) on them. The principal eigenvector that is obtained represents a "generic-shape" representation for the object. Figure 7 illustrates the generic skeletons for a cow and a motorbike.

5. Integrating Low and High Level Features

The previous two sections describe how the low level and high level features are represented in our algorithm. In this section we discuss how the two sets of features are integrated. This is established by the following two steps:

1) We consider the union of feature points from the training images.
2) Next, we traverse the points on the generic skeleton and choose feature points which are within a given keypoint-selection-radius.
3) We term these selected keypoints as Structural keypoints forming the template for object categories.

This simple scheme to select the structural keypoints captures the important characteristics of the two different types of features. For example, in order to have good object recognition, using feature points, we need complex and large set of points. On the other hand generic shape representation is robust when it is simple. By choosing the union of features and only the single generic structure, we maintain the advantages of both the approaches.

We note here that the choice of radius for structural keypoint selection is an important parameter in our algorithm. By choosing a small radius very few feature points would be selected thus having a good structural context but losing on low-level details. On the other hand selecting a large radius would mean that the structural skeleton would be lost in the selected set of points. A detailed evaluation of the tradeoffs associated is part of future work.

6. Matching Algorithm

We now discuss the second phase of the algorithm where we use the templates created in the training phase and classify test/input images to object categories. The matching algorithm follows three distinct phases:
cluster-radius

The object of interest was segmented scaled. The training phase, we use eight exemplars for each object images of each class were tested for evaluation. For the motorbikes and planes from Caltech database [4]. 100 from TU Darmstadt database [5] and images of faces.

7. Algorithm Evaluation

We have evaluated the algorithm using images of cows from TU Darmstadt database [5] and images of faces, motorbikes and planes from Caltech database [4]. 100 images of each class were tested for evaluation. For the training phase, we use eight exemplars for each object class. The object of interest was segmented scaled. The cluster-radius was chosen as 10 units. The vector size of the C2 feature is 256. The recognition rates for the classes are given in Table 1. The recognition rate is dependent on the keypoint-selection-radius. With a larger radius, more keypoints are available for comparison and increases the recognition level. However the computation cost is significantly increased. To see how using the structural context with skeletons increases the recognition level, we find the object categories by randomly selecting the same number of keypoints as selected with radius 15 under structural context. We see that the recognition rate is significantly when kepoints are selected without structure. Thus can we achieve high recognition rates at low computation cost if we use the skeletons as reference for selecting keypoints.

8. Conclusions:

The recognition results presented here provide a preliminary analysis for the possibility of using a mixture model of low-level features and global shape information for object category recognition. It shows that a competitive performance can be achieved with few keypoints, whose relative positions with respect to each other are known. This method reduces the computational cost and recognizes objects even when geometrically changed.

To boost the recognition rate, we plan to address the problem of extraction of keypoints and their description. We realize that a more contextually relevant key-point extractor would help in making the system more robust and efficient. This would help in recognition from images with much larger viewpoint variation. Also a more number of exemplars per object can enhance the object model description.

9. References


<table>
<thead>
<tr>
<th>Dataset</th>
<th>w/o structure</th>
<th>with r = 15</th>
<th>with r = 20</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cows</td>
<td>58</td>
<td>78</td>
<td>82</td>
</tr>
<tr>
<td>Bikes</td>
<td>63</td>
<td>84</td>
<td>85</td>
</tr>
<tr>
<td>Faces</td>
<td>67</td>
<td>91</td>
<td>91</td>
</tr>
<tr>
<td>Planes</td>
<td>51</td>
<td>68</td>
<td>83</td>
</tr>
</tbody>
</table>

Table 1. Experimental results.