LEARNING LOCAL FEATURES FOR OBJECT CATEGORIZATION

Yi Ouyang, Ming Tang, Shi Chen, Jinqiao Wang, Hanqing Lu, Songde Ma

National Laboratory of Pattern Recognition,
Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China
{youyang, tangm, schen, jqwang, luhq, masd}@nlpr.ia.ac.cn

ABSTRACT

In this paper, for every local feature, we propose to learn its similar local features across all positive images, instead of using heuristic distance as similarity measure. Specifically, Multiple Instance Learning (MIL) is employed to simultaneously determine the similar points of a local feature and learn its corresponding discriminative function which can be regarded as some kind of similarity measure. For each local feature, a weak learner is constructed based on such similarity measure. Then AdaBoost selects the most discriminative local features and combines them to form a strong classifier. Experimental results show encouraging performance of our method.

Index Terms—Object categorization, MIL, AdaBoost

1. INTRODUCTION

Object categorization has drawn considerable attention and has been investigated intensively in decades. However, it remains a challenging problem in multimedia and computer vision community. The major difficulties lie in the complex background and large intra-class variation caused by change in viewpoint and scale, shape deformation, occlusion, etc.

A great many methods have been proposed to deal with this problem recently, and most of them are based on the local features (key points), e.g. Geometric Blur [1].

One popular strategy to organize the local features is to quantize them first. Zhang et al. [2] propose to cluster the local features to “visual words”, and then represent an image as a frequency histogram over them. Based on the hierarchical clustering, Grauman and Darrel [3] propose Pyramid Match Kernel (PMK) to measure the similarity between two point sets. In [3], the image could be regarded as being described using hierarchical visual words.

The quantization-based image representation is efficient. This efficiency, however, comes at a high cost: quantization errors are introduced during the clustering process, and the amount of discriminative information is considerably reduced [4]. Thus, many researchers propose to use local features directly for object categorization.

Zhang et al. [5] propose a KNN and SVM based method, and use the mean nearest distances to measure the similarity between two sets of key points. In their algorithm, every point is of the same importance when measuring similarities, while Frome et al. [6] propose to learn the importance of each local feature for each image.

For these approaches which are not based on the quantization, the similarity of the local features is always simply measured by the Euclidean distance [5, 6], which ignores the distribution of key points in the feature space and may fail to achieve the best discriminativity. As shown in Fig. 1, some points from the other kinds of objects or even from the background may be regarded more similar to point \( i \) than points representing the same part of the same kind of object as \( i \). In this paper, for every key point, we propose to...
learn its similar key points across all positive images, instead of using heuristic distance as similarity measure.

Although unexplored in object categorization, the similar idea of local feature learning has been investigated in object detection and tracking. Grabner et al. [7] propose to treat tracking as a matching problem of detected key points between successive frames. Background information is incorporated to learn the classifier-based key point descriptors. In this method, AdaBoost is utilized as point-based classifier, while, in [8], random tree is used instead.

However, it is not straightforward to extend the idea of the above two methods to object categorization. In [7, 8], the local feature learning aims to model the changes of one point under different circumstances, e.g. affine transformation. Thus to learn a classifier for one point, the positive training set could be constructed by heuristically applying some geometric transformations to the original image patch [8] or by adding the matched points in the successive frames during the tracking process [7]. Whereas in object categorization, for one point \( i \) in the image of category \( Ca \), we need to model its changes in other images of \( Ca \). Thus the positive training set should consist of point \( i \) and its similar points from other images of \( Ca \). The learning process incorporates the category information, and captures discriminative information of the group of local features which represent the similar part of the images in \( Ca \), as shown in Fig. 1. Note that the classifier for point \( i \) cannot be trained directly, because the similar points to \( i \) in other images are unknown. In this paper, Multiple Instance Learning (MIL) is employed to find the similar points and learn the classifier simultaneously.

Based on each learned point-based classifier, a weak classifier can be constructed to determine whether an image is a positive one or not through checking if it contains a similar point to \( i \). Considering the large intra-class variation and probable noise points from the background, AdaBoost is employed to select the most discriminative points to form a strong classifier.

The most related work to ours is [9]. The major difference is how to construct weak learner. As in [5, 6], [9] simply use Euclidean distance to measure the similarity of two points, while we propose to learn the similarity measure for each point.

2. ADABOOST FRAMEWORK

The discrete version of AdaBoost [10] defines a strong binary classifier \( H \)

\[
H(z) = \text{sgn}\left(\sum_{i=1}^{T} \alpha_i h_i(z)\right)
\]

using a weighted combination of \( T \) weak learners \( h_i \) with weights \( \alpha_i \). Each weak learner \( h_i \) may explore any feature \( f \) of the data \( z \).

In object detection [11], \( f \) is defined as histograms computed for rectangular image regions on the object. In object categorization, as the location of the object in the training image is unknown, the feature used in detection cannot be applied any more. In this paper, ignoring spatial information, we represent an image as an unordered set of local features. The formal definition of \( f(z) \) in Eq. (1) is as follows:

\[
f(z) = \{p_{z,i} | i = 1, 2, ..., n_z\}
\]

where \( p_{z,i} \) is the local feature of the point \( i \) in the image \( z \), and \( n_z \) is the number of key points extracted in \( z \).

3. CONSTRUCTION OF THE WEAK CLASSIFIER

3.1. Multiple Instance Learning Formulation

For point \( i \) in the image of category \( Ca \), it is reasonable to assume that there is some point representing the similar content to \( i \) in the rest images of \( Ca \), and there is no similar point in images from other categories. Denoting the images in \( Ca \) as positive bags and those in other categories as negative ones, we formulate the similarity measure learning can as a problem of Multiple Instance Learning. Specifically, mi-SVM [12] is employed in this paper to solve the problem.

Denote \( y_{z,i} \) to be the instance label of point \( p_{z,i} \) and \( Y_z \), the label of image \( z \). mi-SVM is formulated as follows:

\[
\min_{\{y_{z,i}\}, w, b} \frac{1}{2} \|w\|^2 + C \sum_{z,i} \bar{\xi}_{z,i}
\]

subject to:

\[
\sum_{i} y_{z,i} \frac{1}{2} \|y_{z,i}\|_1 \geq 1, \forall z \text{ s.t. } Y_z = 1
\]

\[
y_{z,i} = -1, \forall z \text{ s.t. } Y_z = -1
\]

\[
\forall i : y_{z,i} \left(\left\langle w, p_{z,i} \right\rangle + b\right) \geq 1 - \bar{\xi}_{z,i}, \xi_{z,i} \geq 0, y_{z,i} \in \{-1, 1\}
\]

It should be noticed that our problem is slightly different from the standard MIL. In this paper, each mi-SVM is trained for one point \( p_{z,i} \), denoted by \( mi-SVM_{z,i} \), and we have known that \( p_{z,i} \) is a positive instance.
3.2. Implementation Discussions

Before training an mi-SVM for point \( i \), the points that are very different from \( i \) could be filtered out to reduce the computational burden. This strategy enables the classifier to be learned only in the local feature space. Specifically, mi-SVM classifier is trained in the hyper-sphere centered at \( p_{z,i} \) with radius of \( r_{z,i} \) in the feature space.

Denote the set of positive images as \( pos \), and that of negative ones as \( neg \). Define the distance from point \( p_{z,j} \) to image \( z' \) as \( d_{z,j,z'} = \min_{j} \| p_{z,j} - p_{z,i} \| \), where \( \| \cdot \| \) represents \( L_2 \)-norm. In our algorithm, geometric blur [1] is used as local descriptor. And in practice, it is found quite robust and in majority cases the positive instance in the positive image \( z' \) is just

\[
\{ p_{z,j'} | j' = \arg\min_j \| p_{z,j} - p_{z,i} \| \}
\]

Based on this observation, \( r_{z,i} \) is set as follows:

\[
r_{z,i} = \text{mean}(d_{z,i,z'}) + \beta \times \text{var}(d_{z,i,z'})
\]

where \( \beta \) is a trade-off between efficiency and accuracy. The larger \( \beta \), the less probability that one positive instance will be filtered out before training, and the more points will be involved during solving the mi-SVM.

It is observed in experiments that there may be too many points falling into the hyper-sphere, especially when \( p_{z,i} \) is from the background or texture foreground. To be more efficient, within the hyper-sphere, at most \( k \) nearest points to \( p_{z,i} \) are selected for each positive image, and one for each negative image. Experiments demonstrate that this strategy can greatly reduce the computational burden and has no significant impact on the final results even if \( k = 1 \).

3.3. Construction of Weak Learners

Through the learned classifier of point \( i \), training images are projected to real values based on which a weak classifier is constructed. Denote \( \text{mi-SVM}_{z,i}(p_{z,i}) \in \mathbb{R} \) to be the output of \( \text{mi-SVM}_{z,i} \) with the input \( p_{z,i} \). The project function in Eq. (1) is constructed as follows:

\[
g_{z,i}(z') = \begin{cases} 
\text{max}(\text{mi-SVM}_{z,i}(p_{z,i})) & \exists p_{z,j'}, s.t. \| p_{z,j'} - p_{z,i} \| \leq r_{z,i} \\
-1 & \text{otherwise}
\end{cases}
\]

Based on each \( g_{z,i} \), one weak learner could be trained, and the optimal threshold in Eq. (1) is determined as in [11].

4. EXPERIMENTS

In our algorithm, local features are extracted randomly at edges [6] and described by geometric blur descriptor [1], which is capable of capturing the local shape and structure information, and is robust to the change of lighting and geometric deformation.

For each image, 800 key points are extracted. To be robust, two scales of descriptors are adopted (see details in [6]). To evaluate the proposed algorithm and compare with the method in [14], seven categories are chosen from the Caltech 101 database, which are airplanes, watch, leopards, motorbikes, faces, ketch and cars. And in all experiments \( \beta \) is set 2 and \( k \) is set 3 heuristically. All of our experiments are done on a server with 4 Quad-Core Intel Xeon E7320 (2.13GHz) processors and 16GB memory.

One-versus-all strategy is utilized to train classifiers. For each object category, 30 images are randomly selected as training data and 50 images as test data. Thus the total number of points in positive dataset is \( 30 \times 800 \). By removing almost duplicate points in feature space, about 10000 points are left to construct weak learners. It takes about 40 ms to train an mi-SVM for one point using a single core. Finally 100 most discriminative points are selected by AdaBoost to form a strong classifier. Each experiment is repeated for 10 times, and the average ROC equal error rate is reported in Tab. 1.

For comparisons, hierarchical clustering based Pyramid Match Kernel (PMK) [3] is implemented with PCA-SIFT [13] descriptors and partially contextual descriptors [14]. Our results are also compared with the results of the approach [9] (discussed in Sec. 1). As Opelt et al. [9] conduct experiments only on four categories, the rest categories in Tab. 1 are left black.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>airplanes</td>
<td>88.9</td>
<td>88.43</td>
<td>93.06</td>
<td>97.11</td>
</tr>
<tr>
<td>watch</td>
<td>--</td>
<td>82.57</td>
<td>83.05</td>
<td>96.91</td>
</tr>
<tr>
<td>leopards</td>
<td>--</td>
<td>99.75</td>
<td>99.00</td>
<td>96.81</td>
</tr>
<tr>
<td>motorbikes</td>
<td>92.2</td>
<td>88.90</td>
<td>93.06</td>
<td>97.73</td>
</tr>
<tr>
<td>faces</td>
<td>93.5</td>
<td>94.44</td>
<td>98.54</td>
<td>99.27</td>
</tr>
<tr>
<td>ketch</td>
<td>--</td>
<td>88.19</td>
<td>90.74</td>
<td>95.84</td>
</tr>
<tr>
<td>cars</td>
<td>83.0</td>
<td>95.20</td>
<td>96.62</td>
<td>99.11</td>
</tr>
<tr>
<td>Ave.</td>
<td>89.4</td>
<td>90.93</td>
<td>93.44</td>
<td>97.54</td>
</tr>
</tbody>
</table>

From Tab. 1, the error rates of our algorithm are lower than the three in all cases except for the category “leopards”. The probable reason for the relative higher performance of PMK on leopard category is that, for leopard, the dominant and discriminative feature is the body’s texture, and it occurs
frequently enough so that the quantization error is not as much as the other categories [4].

The average ROC equal error rates of the seven categories with the different number of weak learners are shown in Fig. 2. It can be observed that the curve tends to be stable after about 40 iterations. And using only seven weak learners, we are able to obtain comparable results with those by the three methods.

Figure 2. Average ROC equal error rate with different number of the weak learners.

Figure 3 shows the selected discriminative points in the first six iterations of AdaBoost for “cars” and “airplanes”, respectively. It is observed that all those key points are lying inside or just around the object. This phenomenon partly proves that our algorithm is capable of discovering some intrinsic characteristics of the images belonging to the same object category. For category “airplane”, it is noticed that all the selected key points are on the afterbody. Considering the large variance of the body of the different airplanes, the afterbody may be the most stable part and will be selected out first. Thus this result is reasonable.

Figure 3. Demonstration of the selected key points, the locations of which are at the cross points. (a) Selected points for cars. (b) Selected points for airplanes

4. CONCLUSIONS

In this paper, we propose a new method for object categorization with AdaBoost framework in which weak learners are constructed based on the similarity measure of the local features. The key contribution is that the similarity measure is based on the discriminative function through local feature learning, rather than using some heuristic distances. Preliminary experiments show promising performance. Further analysis on comparison with heuristic or weighted distances is undergoing.

5. ACKNOWLEDGEMENT

The research was supported by National Natural Science Foundation of China (Grant No. 60605004, 60835002, 60835004 and 60572057).

6. REFERENCES