Local Metric Learning for Exemplar-Based Object Detection

Xinge You, Senior Member, IEEE, Qiang Li, Dacheng Tao, Senior Member, IEEE, Weihua Ou, and Mingming Gong

Abstract—Object detection has been widely studied in the computer vision community and it has many real applications, despite its variations, such as scale, pose, lighting, and background. Most classical object detection methods heavily rely on category-based training to handle intra-class variations. In contrast to classical methods that use a rigid category-based representation, exemplar-based methods try to model variations among positives by learning from specific positive samples. However, current existing exemplar-based methods either fail to use any training information or suffer from a significant performance drop when few exemplars are available. In this paper, we design a novel local metric learning approach to well handle exemplar-based object detection task. The main works are two-fold: 1) a novel local metric learning algorithm called exemplar metric learning (EML) is designed and 2) an exemplar-based object detection algorithm based on EML is implemented. We evaluate our method on two generic object detection data sets: UIUC-Car and UMass Fddb. Experiments show that compared with other exemplar-based methods, our approach can effectively enhance object detection performance when few exemplars are available.

Index Terms—Co-occurrence voting, exemplar metric learning (EML), object detection.

I. INTRODUCTION

OBJECT detection is one of the most fundamental and challenging topics in computer vision. Object detection plays a critical part in wide applications, including content-based image retrieval, video surveillance, object tracking, natural human–computer interface, and intelligent transport systems. However, it is difficult, owing to large intra-class variations. Particularly in real application systems, the imaging conditions can be extraordinarily complex. The performance of detection algorithms will drop as a result of view difference, illumination effect, object deformation, and occlusion. In this paper, most object detection algorithms are based on delicately designed structures like cascade [1] and deformable part model [2]. Simultaneously, those informative histogram features such as SIFT [3], SURF [4], LBP [5], and HOG [6], have also been widely used to reliably represent objects. Both of the above two aspects are aimed at learning robust category object detectors that can endure large intra-class variations. However, such a rigid category-based methodology may suffer from the problem of visual-incoherence in each object category [7].

Different from category-based methods, exemplar-based approaches treat positive samples specifically to avoid the visual-incoherence problem. Chum and Zisserman [8] proposed the exemplar-based classification model to coherently represent object categories. Frome et al. [9], [10] designed local distance function learning algorithms using triplets, and the focal image is actually an exemplar. Malisiewicz et al. [7] designed an ensemble of so-called exemplar-SVMs for object detection tasks on the PASCAL VOC 2007 data set [11]. In addition, one-exemplar, query-driven and image-to-class approaches can also be summarized to the exemplar-based methodology. Seo and Milanfar [12], [13] designed a novel feature extraction algorithm called LARK and applied LARK in object and action detection tasks by using one exemplar. Fu et al. [14], [15] proposed a locally adaptive subspace learning framework based on a query-driven strategy. Boiman et al. [16] used image-to-class distances to defend nearest neighbor classifier, which has achieved comparable performance with support vector machines (SVM) [17]. Wolf et al. [18] also used image-to-class distances to design a similarity measure based on a variant of linear discriminant analysis (LDA) [19], [20]. However, among those methods which can handle object detection task, most of them [8]–[10] are based on an almost balanced training scenario, and the number of exemplars is comparable with that of negative samples. These methods cannot be used in a training setting where only a fewer exemplars are available. Some other methods which can be used with this setting, either fail to use any training information [12] or encounter significant performance drops [7].

Metric learning is another very effective approach in handling intra-class variations and combining heterogeneous
features in recognition tasks [21]–[25]. The basic objective of metric learning is to shrink the distance of similar samples while enlarging the distance of dissimilar samples by the metric mapping. Most of the recent representative metric learning algorithms are mainly in a global sense, including probabilistic global distance metric learning (PGDM) [26], neighborhood component analysis [27], metric learning via large margin nearest neighbor (LMNN) [28], information-theoretic metric learning (ITML) [29], and metric learning by consistent empirical loss minimization [30]–[32]. To find a global consistent metric, millions of distance pairs are employed during learning. Unfortunately, global metric learning methodology is not suitable for the exemplar-based scenario for two reasons. First, exemplar’s local adaptiveness is totally neglected in training by most of the global metric learning algorithms. Second, a single well-trained global metric cannot fit well to all local exemplars simultaneously, because a global metric is usually a compromising product for all of the training data.

To tackle the mentioned problems involving existing exemplar-based methods and metric learning algorithms when only a few exemplars are available, this paper presents an exemplar metric learning (EML) approach to handle exemplar-based object detection. Compared with previous studies in exemplar techniques and metric learning, the main works of this paper are as follows.

1) We propose a novel EML algorithm in a local sense. More specifically, we learn a specific metric for each exemplar using the imbalanced training set based on hinge loss. Furthermore, to protect exemplar metric from undesirable distractions, we discard all exemplar-irrelevant distance pairs in training. In general, metric learning is usually considered in a cross-dimension sense to find an optimal distance metric for the whole feature space to well handle the problems resulting from naive Euclidean distance. Unfortunately, a single global metric learned by regular metric learning techniques is unable to play a reliable all-good role when used in the exemplar-based methodology. Different exemplars usually have large dimensional variations, which may lead to disastrous effects when using a global metric. Our proposed EML method is capable of local adaptiveness and is suited with exemplar-based object detection tasks.

2) We implement an exemplar-based object detection algorithm on the basis of the learned exemplar metrics. Firstly, the original redundant patchwise features, such as, LARK [12] and HOG [6], are projected onto local PCA basis that are obtained from the central exemplar. Local adaptiveness is thus strengthened which can help the exemplar to mine more discriminant information for detection. In addition, by applying previously learned exemplar metrics in feature representation and resemblance calculation, we get a reasonable score map between the exemplar and each sliding window of the testing image. Then a co-occurrence voting strategy is utilized to obtain a final map for nonmaxima suppression.

The rest of this paper is organized as follows. In Section II, we present the detailed model and algorithm of EML. Section III presents the exemplar-based object detection algorithm based on the learned exemplar metrics. Section IV reports the experimental results on UIUC-Car data set and UMass FDB face detection data set. Section V concludes this paper.

II. EXEMPLARY METRIC LEARNING

In this section, we first present the objective function of EML. Afterward, we present the algorithm to learn the exemplar specific metric from an imbalanced set of training data. The proposed learning method incorporates exemplar-based methodology into a local metric learning framework under the imbalanced training setting. On the one hand, the specific exemplar features are well exploited with the exemplar-based methodology. On the other hand, the local metric learning framework optimally specifies the distance weighting parameters of exemplar features. With the above description, the following presents the learning model and algorithm in details.

A. Objective Function

The traditional metric learning strategies learn a global metric M by exploiting a vast number of similar and dissimilar example pairs. It is generally accepted that global metric is more robust and capable of generalization. However, such a global metric may encounter deficiency of detection performance, which is a result of large intra-class variations. Most standard semantic categories (e.g., car, face, etc.) do not form coherent visual categories, thus treating visually different positives as a whole may results in weak detectors. To handle this problem encountered by global metric learning methods, we proposed the local metric learning method for visually specific exemplars. Such an exemplar metric can adaptively reflect the metric space of the local specific exemplar. In addition, the proposed method can also generalize well by combining relative amounts of diverse exemplar metrics.

First, we define the following patch index and feature matrices, \( p = 1, 2, \ldots, P \), \( P \) is the number of patches for each sample; \( Q = [q_1, q_2, \ldots, q_P] \) is the patchwise feature matrices of the exemplar. An instance of the feature matrix are shown in Figs. 1 and 2. \( X_i = [x_{i1}, x_{i2}, \ldots, x_{iP}] \) is the feature matrix of one positive training sample, \( i = 1, 2, \ldots, N_+ \), \( N_+ \) is the number of positive training samples; \( X_j = [x_{j1}, x_{j2}, \ldots, x_{jP}] \)

![Fig. 1. LARK feature and the feature after patch PCA of car image pos-0 from UIUC-Car dataset: FDI: feature dimensional index. PI: patch index.](image-url)
is the feature matrix of one negative training sample, \( j = 1, 2, \ldots, N_-, N_- \) is the number of negative training samples. It is an imbalanced training set because \( N_+ \) is usually much smaller than \( N_- \) in our experiment.

We redefine the Mahalanobis distance between matrix \( X_i \) and \( X_j \) as

\[
d_{L,M}(X_i, X_j) = \text{trace}(L(X_i - X_j)M(X_i - X_j)^T L^T) \tag{1}
\]

where \( L \in \mathbb{R}^{D \times D} \) is a patch projection matrix, \( D' \leq D \), and \( M \in \mathbb{R}^{P \times P} \) is symmetric and positive definite.

Given an exemplar with feature matrix \( Q \), the patch projection matrix is obtained adaptively from the exemplar, which we denote as \( LQ \). See also in Figs. 1 and 2 for the local patch PCA subspace obtained from the exemplar. Then, we obtain the patch-PCA projected version of \( Q, X_i, \) and \( X_j \) as

\[
\tilde{Q} = LQ, \quad \tilde{X}_i = LQX_i, \quad \tilde{X}_j = LQX_j.
\]

Thus, we obtain the exemplar-specific Mahalanobis distance as

\[
d_{LQ,MQ}(X_i, Q) = \text{trace}(LQ(X_i - Q)MQ(X_i - Q)^T L^T) = \text{trace}((X_i - \tilde{Q})MQ(X_i - \tilde{Q})^T) = d_{MQ}(\tilde{X}_i, \tilde{Q}). \tag{2}
\]

To simplify presentations, we still use \( Q, X_i, \) and \( X_j \) to denote the locally projected feature matrices in the rest of this paper. With these patch features, we aim at learning the distance calculation. By imposing different importance weights to patches, the objective function is constructed as

\[
\min_{MQ} \left\{ \sum_{i=1}^{N_+} L(-d_{MQ}(X_i, Q) + c_1) + \sum_{j=1}^{N_-} L(d_{MQ}(X_j, Q) + c_2) + \lambda \|MQ\|_F^2 \right\} \tag{3}
\]

where \( \| \cdot \|_F \) means the Frobenious norm, \( L(z) = \max(0, 1 - z) \), \( c_1, c_2 \) are bias terms, and \( MQ \in \mathbb{R}^{P \times P} \) is the exemplar metric. The first term penalizes misclassification of positive samples which are the same category with the exemplar. The second term penalizes misclassification of negative samples. Generally speaking, the parameter set is usually very large, thus we can put some special kinds of constraints on the metric.

There are several kinds of constraints for metric learning, for example, low-rank, sparsity, and diagonality. The low-rank constraint can be regarded as a kind of dimensional reduction method, which fundamentally shrinks the intrinsic dimension of the metric parameters. The sparsity constraint can well handle the over-fitting problem by setting some small parameters zero. However, the performance may be impaired if the sparsity constraint is too strong. In practice, the spatially corresponding feature importance weights are sufficient to excavate discriminant information of the whole distance. In addition, it is computationally friendly to optimize fewer parameters and implement the following detection algorithm. Assuming that the patch features are independent to each other, we consider the diagonal case \( MQ = \text{diag}(m_1, m_2, \ldots, m_P) \).

### B. Problem Reformulation and the Algorithm

Note that \( d_{MQ}(X_i, Q) \) is linear with respect to the elements of \( M \). Thus, we can rewrite \( d_{MQ} \) as

\[
d_{MQ}(X_i, Q) = \text{trace}(MQ(X_i - Q)(X_i - Q)^T) = \mathbf{m}^T \mathbf{d}_l \tag{4}
\]

where \( \mathbf{m} = [m_1, m_2, \ldots, m_s]^T \) is the vector form of the learned exemplar specific metric, \( S = P^2 \) is the number of parameters, and \( \mathbf{d}_l = [d_{l1}, d_{l2}, \ldots, d_{lS}]^T \) is the vector form of the matrix feature distances corresponding to exemplar metric parameters. Each element of \( \mathbf{d}_l \) is as

\[
d_{ls} = \|x_{lp} - q_{lp}\|_2^2 = (x_{lp} - q_{lp})^T(x_{lp} - q_{lp}) \tag{5}
\]

which represents the basic Euclidean distance of each patch feature pair. \( s = 1, 2, \ldots, S \) is the index of basic Euclidean distances, \( p = 1, 2, \ldots, P \) and \( t = 1, 2, \ldots, P \) are patch indices of \( l \)TH training sample \( X_i \) and the exemplar \( Q \), respectively. More specifically, \( d_{ls} \) is the pixel-corresponding patch distance when \( p = t \) and cross-pixel patch distance when \( p \neq t \).

Especially, when diagonality constraint is employed, we get the reduced vector \( \mathbf{m} = [m_1, m_2, \ldots, m_P]^T \) as the diagonal case of the exemplar metric. \( \mathbf{d}_l = [d_{l1}, d_{l2}, \ldots, d_{lP}]^T \) is also reduced as the vector form of the corresponding feature distance, which only reflects matrix feature distances from pixel-corresponding patches between one training sample and the exemplar. In the experiments, we employ this constraint on the exemplar metric to control the model complexity.

In addition, we can use an indicator variable to distinguish two types of misclassifications, \(-1\) for positive loss, \(1\) for negative loss, that is

\[
y_l = \begin{cases} 1, & \text{if sample } l \text{ is positive} \\ -1, & \text{otherwise} \end{cases}
\]

then, the objective function (3) can be rewritten as

\[
\min_{\mathbf{m}} \left\{ \sum_{l=1}^{N} L(y_l(\mathbf{m}^T \mathbf{d}_l + c)) + \lambda \|\mathbf{m}\|_2^2 \right\} \tag{7}
\]

where \( N \) is the total number of training samples and \( c \) is a bias term. Note that this is a standard SVM problem, which can be solved using LibSVM [33].
III. OBJECT DETECTION BASED ON THE LEARNED EXEMPLARY METRICS

The learned exemplar metric is applicable to object detection. Note that exemplar metric is capable of properly using the discriminant information extracted from original features. Thus, object detection algorithms will present a better detection performance on using exemplar metrics. More specifically, EML can successfully learn an adaptive local metric corresponding to the specific exemplar features, thus providing more accurate distance calculations and getting more precise resemblance maps in object detection. In addition, combining resemblance maps from different exemplars with co-occurrence voting, we can get more stable detection results. This voting strategy is implemented to handle problems in object detection such as large intra-class variations and complex backgrounds. For the following contents, we use LARK [12] to illustrate the whole framework of the proposed method.

A. Review of LARK Features

First of all, we briefly review LARK features [12] used in this paper. LARK features are constructed with so-called locally adaptive regression kernels. A local kernel $K(\cdot)$ is modeled by a radial-symmetric function as

$$K(z_l - z; H_l) = \frac{K(H_l^{-1} (z_l - z))}{\det(H_l)}, \quad l = 1, 2, \ldots, D \quad (8)$$

where $z_l = [z_{1l}, z_{2l}]^T$ is the spatial coordinates of pixels in a $\sqrt{D} \times \sqrt{D}$ size patch centered at $z$. $H_l = hC_l^{1/2} \in R^{2 \times 2}$ is defined as the steering matrix, where $h$ is a global smoothing parameter, the covariance matrix $C_l$ consists of gradient vectors within a local window centered at $z_l$. $K(\cdot)$ can be defined as a Gaussian function by using steering matrix, thus leading to a local steering kernel

$$K(z_l - z; H_l) = \frac{\sqrt{\det(C_l)}}{2\pi h^2} \exp\left\{\frac{-(z_l - z)^T C_l(z_l - z)}{2h^2}\right\}. \quad (9)$$

For pixels in the $\sqrt{D} \times \sqrt{D}$ size patch centered at $z$, their kernel values are column-stacked and then normalized into a $D$ dimensional kernel value vector. This vector can be considered as a feature representation of the patch centered at $z$. Applying the above process to each patch in an image, we obtain the original LARK features.

B. Feature Projection Representation With an Exemplar Metric

LARK features contain large amounts of useful image information [12], which can be directly used in object detection. Moreover, the detection performance will be improved after PCA dimensional reduction. In practice, when handling one certain exemplar, we use its specific PCA basis to project all original LARK features. Such a methodology will further strengthen the exemplar’s local adaptiveness. To simplify illustration, we still use the name of LARK after PCA dimensional reduction. The above-described feature extraction scheme is used in our experiments.

The extracted LARK features are very discriminant in detection [12]. However, in real detection tasks, simply treating features from different patches with the same importance in detection is inappropriate. Our proposed EML algorithm sufficiently mine feature importance of different patches using only several positive samples and large amount of negative ones, thus properly using the discriminability of LARK features. Fig. 3 shows the comparison between original LARK features and new features projected with exemplar metric.

Because LARK features of exemplar image after PCA dimensional reduction is still patchwise features, we simply presents these feature images corresponding to the first four PCA-dominant directions. The projected new features have weakened the impact of background features in exemplar image, which can be observed from the change of edge components in feature images. These edge components become smoother and foreground contour is more outstanding after EML projection.

C. Resemblance Map With Exemplar Metric

Feature projection representation obtained from above subsection can well be employed to object detection, we introduce resemblance map with exemplar metric in the following. Resemblance map usually refers to the matching degree between some sliding window of testing image and the template. It has two kinds of representations, distance and similarity scores. The smaller the distance score, the higher is the matching degree. Whereas similarity score is the opposite. We use the cosine similarity to calculate the resemblance map score in this paper. Thus, the higher score indicates the higher probability of object existence in that position. Fig. 4 completely shows the EML influence on feature distance calculation and resemblance map. The detection results are also provided for comparison.

Fig. 4 shows two testing image resemblance map comparison between LARK and LARK after EML projection. Both the methods used the same side-view car image as query. For the target image of test-10 of the UIUC-Car data set [34], EMLs resemblance map has two comparable outstanding areas, that is, both the two local maxima are significant. Whereas the LARKs resemblance map only has one significant local maximum and the other local maximum is really small in value. For the target image of test-25, LARK even gives the wrong resemblance map whereas our method can well reflect the proper response. When thresholding with the fixed alpha value 0.99, we get the detection result on the last column. The LARK method missed one positive and made a false positive, while our method gives all the three positives with no false
positives. From the above observation, we can conclude that, EML can well enhance the performance of LARK.

D. Scale Estimation Based on Maximum Likelihood

To cope with large-scale variations, we construct a multiscale pyramid of the target image $T$. This is a nonstandard pyramid as we reduce the target image size by steps of 10% so that a relatively fine quantization of scales is accounted. More specifically, we first generate images of all the different scales $T^0, T^1, \ldots, T^S$, where $S$ is the coarsest scale of the pyramid and the image $T^S$ is the most downsized one. Further we compute original LARK features for each scale $W_{T^0}, W_{T^1}, \ldots, W_{T^S}$ and projected them onto $L_Q$. $L_Q$ is the PCA subspace basis of exemplar features $W_Q$. After projection, we get dimensional reduced representations of each scaled image features as

$$F_{T^0} = L_Q W_{T^0}, F_{T^1} = L_Q W_{T^1}, \ldots, F_{T^S} = L_Q W_{T^S}. \quad (10)$$

Next, compute resemblance map of each scaled target image $f(\rho_i) = \rho_i^2 / (1 + \rho_i^2)$, where $\rho_i$ is cosine similarity of exemplar features and testing window features. This step gives $S + 1$ resemblance maps $RM_0, RM_1, \ldots, RM_S$, which construct the maximum likelihood function $p(f(\rho_i)|S_i)$. Note that $S_i$ is the scale at position $i$. In addition, we simply upscale resemblance maps to match the size of finest scaled image. Finally, the maximum likelihood estimate of the scale at each position is achieved by comparing the upscaled resemblance maps as

$$\hat{S}_i = \arg\max_{S_j} p(RM_{S_j}|S_i). \quad (11)$$

E. Co-occurrence Voting With Multiple Exemplars

Note that it is unavoidable to consider strategies for combining different resemblance maps from different exemplars in object detection tasks. In our proposed detection algorithm, we employ co-occurrence voting strategy based on exemplars, thus suppressing redundant responses. We first threshold each exemplar’s resemblance map to partly clear undesired small background responses. Then, all the resemblance maps of each scale are added into one resemblance map according to the same scale. Finally, the maximum likelihood scale estimation is implemented to aggregate a final resemblance map and corresponding scale factors of each location. Co-occurrence voting is presented in Fig. 5, each of the resemblance maps is an aggregated map for illustration clarity.

F. Nonmaxima Suppression on Resemblance Map

In the object detection field, the most famous strategy to combine multiple overlapping bounding boxes is nonmaxima suppression (NMS). The aim of NMS is to avoid false detection resulting from overlapping bounding boxes. In general, a bounding box is considered to be a true positive when it is 50% overlapped with the ground truth. However, the evaluator will only choose one bounding box as true positive when there are multiple detections around one ground truth position. Thus, other detections are all treated as false ones. In fact, NMS can also suppress false detections coming from background. The NMS strategy used in the proposed algorithm is to gradually search on the resemblance map. The first step is to find the threshold response value according to the statistical distribution and parameter $\alpha$. Next, a bounding box is reported around the global maximum when it is larger than the threshold value. Then set area inside the bounding box 0 and continue another global search. The newly found global maximum must satisfy two conditions: 1) the maximum is larger than threshold and 2) the overlap ratio between the bounding box around this maximum and each of previous
IV. EXPERIMENTS AND COMPARISONS

In this section, we evaluate the performance of the proposed method and compare with some other methods. We use two data sets for experiments, that is UIUC-Car data set [34] and UMass FDDB data set [35]. The first data set is for side-view car detection, and the second one is for face detection from complex backgrounds. The proposed algorithm generates a series of bounding boxes around objects of interest and compares them with the ground truth provided by data set. In addition, if the detected region is matched with the ground truth based on certain criteria specified by data set, we evaluate it as a correct detection. Otherwise, it is counted as a false positive. Eventually, we compute precision and recall defined as

\[
\text{Recall} = \frac{TP}{nP}; \quad \text{Precision} = \frac{TP}{TP + FP}
\]  

(12)

where \(TP\) is the number of true positives, \(FP\) is the number of false positives, \(nP\) is the total number of positives in the data set, and \(1 - \text{Precision} = \frac{FP}{TP + FP}\). Experimental results of the UIUC-Car data set will be presented as recall versus 1-precision curve and detection equal-error rate. However, the results of the FDDB data set will be presented as two kinds of receiver operating characteristic (ROC) curves, namely, recall versus false positive curves. Note that detection equal-error rate is a detection (recall) rate when a recall rate is the same as the precision rate.

A. Car Detection

The UIUC-Car data set [34] consists of training and test sets. The training set contains 550 side-view car images and 500 noncar images, both with the same size 100 \times 40\) pixels. There are two parts of the test set: the first for single-scale case and the second for multiscale case. The single-scale test set has 170 gray-scale images containing 200 side views of cars of size 100 \times 40\) pixels. The multiscale test set has 108 gray-scale images containing 139 cars at various sizes between 89 \times 36 and 212 \times 85\) pixels.

According to the evaluation criterion in [34], the proposed algorithm provides a series of bounding boxes to indicate locations and sizes of objects of interest. More specifically, if the bounding box detected by the proposed method lies within an ellipse of a certain size centered at the ground truth, we count it as a correct detection. Otherwise, it is treated as a false positive. Eventually, we accumulate the detecting results of all the testing images to compute precision and recall. Then, by changing the detection threshold, we present recall versus 1-precision curves and calculate equal-error rates. We compare our method with state-of-the-art methods with these two criterions in the following.
1) Single-Scale Test Set: In our experiments, we tested EML with two different patchwise features, LARK [12] and HOG [6]. We first describe the basic settings of the LARK feature extraction procedure. First, the LARK-adaptive templates are used to calculate $9 \times 9$ size LARK descriptors which are the same as the templates. Thus there will be a 81-D descriptor for each patch in exemplar and target images. Furthermore, according to the theoretical analysis and experimental observations in [12], we similarly calculate the projection basis from LARK features of the exemplar image for PCA dimensional reduction. Then, the final image feature representations are obtained by projecting the original LARK features onto this projection basis. The preserved energy of PCA is fixed with 70% in all our experiments. Note that, each exemplar has its own projection matrix and all training samples’ LARK features are projected according to the specific basis. Similar to the HOG feature extraction process, we used VLFeat (version 0.9.17) toolbox [36] to extract a 36-D descriptor for each patch in exemplar and target images. The cell size is fixed with $4 \times 4$. The other steps are the same as in the case of LARK.

To fairly compare with the method in [12], we used the same four exemplars employed in [12]. These four exemplars are pos-0, 19, 21, and 25 in UIUC-Car training set, which are shown in Fig. 6. Note that in our algorithm, we alternately choose one as exemplar and other three as positive training samples. We compared the detection results of each exemplar as query image and also presented the final co-occurrence voting results. Besides, the detection results of ESVM [7] using LARK features are also presented for comparison. Fig. 7 shows all the above recall versus 1-precision curves. We also conducted the above comparisons using HOG in Fig. 8.

The overall performance of the proposed method is shown by the recall versus 1-precision curves. To better show the effectiveness of our proposed algorithm, we provide more detection results of other state-of-the-art object detection methods for comparison in Fig. 9. It is worth noting that
these methods used all the 550 positive and 500 negative samples in the UIUC-Car training set. We also present the results of LARK, LARK+ESVM, and the proposed method LARK+EML for completeness. In contrast, LARK only uses four exemplars to vote without training. In addition, both ESVM and our method EML employed another 500
negative samples for training. From the comparison, we can conclude that our method can largely improve the detection performance of each individual exemplar which outperformed that of ESVM. The proposed algorithm also gets better results after co-occurrence voting.

2) Multiscale Test Set: For the multiscale test set, we construct a multiscale pyramid of the target image $T$: 12 scales with scale factors ranging from 0.4 to 1.5 and the step is 0.1. More specifically, we downscale the target image by steps of 10% up to only 40% of the original image size and upscale the target image by 10% to one and a half times bigger than the original image. Using such a big and fine range of scales, we not only deal with both cases of either the size of objects in the target images being bigger or smaller than the query, but also approximate the best scales of the objects. The rest of the process is similar to the single-scale case.

Similar to the single-scale test, we also use the same exemplars and training setting for the multiscale case. Both individual and voting results are shown in Figs. 10 and 11. The proposed algorithm shows even greater detection performance improvements than another two exemplar-based methods. The final voting results also show the effectiveness of our algorithm.

3) Equal-Error Rate Comparison and Analysis: Now we compare our algorithm with other methods from the aspect of equal-error rate in a more general sense. Because results may be influenced by different choice of exemplars, we construct random tests using the following experimental setting: 1) we first construct the query and negative sets with the first 50 positive and all 500 negative samples from UIUC-Car training set; 2) when given the number of exemplars $NQ$, conduct 100 groups of tests (the single-exemplar case only has 50 groups); 3) for each group, the $NQ$ exemplars are randomly chosen from the query set; and 4) all the three exemplar-based methods are tested using the above groups. Note that, all the training process are restricted just in the query and negative sets.

Table I shows the testing results on the UIUC-Car data set. Our algorithm and another two exemplar-based methods utilized 12 exemplars for the final co-occurrence voting. These equal-error rates are obtained by averaging all results of 100 random tests. See more comparison in the following analysis on varying number of exemplars. In contrast, the two global metric learning approaches and the other competing methods employed all the 550 positive and 500 negative samples for training.

### Table I

<table>
<thead>
<tr>
<th>Method</th>
<th>Single-scale</th>
<th>Multi-scale</th>
</tr>
</thead>
<tbody>
<tr>
<td>Agarwal et al. [34]</td>
<td>76.5%</td>
<td>39.6%</td>
</tr>
<tr>
<td>Fergus et al. [38]</td>
<td>88.5%</td>
<td>-</td>
</tr>
<tr>
<td>Fritz et al. [40]</td>
<td>88.6%</td>
<td>87.8%</td>
</tr>
<tr>
<td>LARK [12]</td>
<td>91.98%</td>
<td>87.65%</td>
</tr>
<tr>
<td>LARK+ITML [29]</td>
<td>91.69%</td>
<td>65.19%</td>
</tr>
<tr>
<td>LARK+LMNN [28]</td>
<td>97.82%</td>
<td>78.15%</td>
</tr>
<tr>
<td>LARK+ESVM [7]</td>
<td>96.31%</td>
<td>87.94%</td>
</tr>
<tr>
<td>LARK+EML</td>
<td>96.73%</td>
<td>90.76%</td>
</tr>
<tr>
<td>Kapoor/Winn [37]</td>
<td>94.0%</td>
<td>-</td>
</tr>
<tr>
<td>Mutch/Lowe [39]</td>
<td>99.94%</td>
<td>90.6%</td>
</tr>
<tr>
<td>Wu/Nevatia [41]</td>
<td>97.5%</td>
<td>93.5%</td>
</tr>
</tbody>
</table>

### Table II

<table>
<thead>
<tr>
<th>Exemplar Number</th>
<th>1</th>
<th>4</th>
<th>8</th>
<th>12</th>
</tr>
</thead>
<tbody>
<tr>
<td>Single-scale LARK [12]</td>
<td>80.63%</td>
<td>90.30%</td>
<td>91.65%</td>
<td>91.98%</td>
</tr>
<tr>
<td>LARK+ITML [29]</td>
<td>87.37%</td>
<td>91.69%</td>
<td>91.61%</td>
<td>91.49%</td>
</tr>
<tr>
<td>LARK+LMNN [28]</td>
<td>77.81%</td>
<td>97.82%</td>
<td>97.95%</td>
<td>97.86%</td>
</tr>
<tr>
<td>LARK+ESVM [7]</td>
<td>91.39%</td>
<td>95.40%</td>
<td>95.95%</td>
<td>96.31%</td>
</tr>
<tr>
<td>LARK+EML</td>
<td>94.20%</td>
<td>96.50%</td>
<td>96.73%</td>
<td>96.73%</td>
</tr>
</tbody>
</table>

### Table III

<table>
<thead>
<tr>
<th>Exemplar Number</th>
<th>1</th>
<th>4</th>
<th>8</th>
<th>12</th>
</tr>
</thead>
<tbody>
<tr>
<td>Single-scale HOG [6]</td>
<td>79.42%</td>
<td>85.40%</td>
<td>87.28%</td>
<td>87.50%</td>
</tr>
<tr>
<td>HOG+ITML [29]</td>
<td>66.48%</td>
<td>81.84%</td>
<td>86.23%</td>
<td>86.87%</td>
</tr>
<tr>
<td>HOG+LMNN [28]</td>
<td>79.62%</td>
<td>90.84%</td>
<td>92.73%</td>
<td>92.90%</td>
</tr>
<tr>
<td>HOG+ESVM [7]</td>
<td>85.54%</td>
<td>92.20%</td>
<td>93.13%</td>
<td>93.56%</td>
</tr>
<tr>
<td>HOG+EML</td>
<td>92.49%</td>
<td>93.72%</td>
<td>94.08%</td>
<td>94.23%</td>
</tr>
</tbody>
</table>

Similar to the single-scale test, we also use the same exemplars and training setting for the multiscale case. Both individual and voting results are shown in Figs. 10 and 11. The proposed algorithm shows even greater detection performance improvements than another two exemplar-based methods. The final voting results also show the effectiveness of our algorithm.

### Fig. 12

Face exemplars and some negative training images employed in face detection experiment.
training samples. For ITML [29] and LMNN [28], we used codes provided by the authors. Besides, the original features are dimensionally reduced to 200 dimensions using a global vector PCA before training.

It can be informed from the model that different number of exemplars will result in different detection results. To analyze its influence on EMLs performance, we test EML by changing the amount of exemplars. Note that each of the test is performed under the most beginning experimental setting. The detection result is the average of 100 random tests. See Tables II and III for the relationship between number of exemplars and equal-error rate. Note that we also provide the results of global metric learning algorithms for comparison.

The results in Tables II and III are consistent with viewpoint of [7], that is, each detector is quite specific to its exemplar and an ensemble of these exemplar-based detectors offers surprisingly good generalization. Therefore, we also employ this ensemble strategy to cover potential variations among positive samples.

In contrast, ITML [29] and LMNN [28] can get higher performance when the testing task is simpler, that is, the single-scale case. However, both of these two methods suffer a lot when tested on the multiscale test set. The reason why this phenomenon happens is, a single global metric learned by ITML [29] or LMNN [28] is unable to play as a reliable all-good role for exemplar-based object detection.

B. Face Detection

The face detection data set and benchmark (FDDB), is a data set of face regions designed for studying the problem of unconstrained face detection. This data set contains the annotations for 5171 faces in a set of 2845 images taken from the faces in the wild data set. Besides, it provides codes for matching detections and annotations and computing the resulting scores to generate the performance curves. They present two types of scoring the detections in an image: discrete score, and continuous score. Under the former criterion, if the ratio of the intersection of a detected region with an annotated face region is greater than 0.5, a score of 1 is assigned to the detected region, and 0 otherwise. On the other hand, for the second criterion, this ratio is used as the score for the detected region. Further details for the evaluation procedure can be found in the FDDB technical report [35].

Fig. 13 shows the testing results on FDDB data set which includes state-of-the-art detection algorithms. According to the standard setups of FDDB data set, we give both two kinds of ROC curves. The red solid lines represent face detection results of the proposed method. Although the result of our method falls a little far behind the results of Li et al. [42] and VJGPR [43], our method currently only employed four
positive samples and 500 negative samples for training and testing. The four positive samples are frontal face images with \(60 \times 60\) pixels, while all the negative samples are obtained from UIUC-Car training set by simple cropping and interpolation. Some training images are shown in Fig. 12. In contrast to the proposed method, Li et al. [42] and VJGPR [43] both need complex and time-consuming offline training process to get the final Adaboost classifier. The training process employed large amount of both positive and negative samples. Take Li et al. [42] for example, this method obtained 13,000 face images from three different databases as positive samples, and negative samples consist of 18,000 nonface images from another three different databases. It is worth noting that despite its simplicity, the proposed method even outperformed Subburaman and Marcel [44], Viola and Jones [1], and Mikolajczyk et al. [45].

Fig. 14 shows additional comparison results with other local approaches and the state-of-the-art metric learning algorithms. We should first note that both ITML [29] and LMNN [28] utilized another 400 aligned face images from ORL data set [46] for training. Interestingly, E SVM falls far behind the other two local methods. In addition, our method also get set [46] for training. Interestingly, E SVM falls far behind the state-of-the-art metric learning algorithms.

Li [42] and handle the model’s complexity.

V. Conclusion

We have presented a novel local metric learning method for exemplar-based object detection. Our method is capable of local adaptiveness during mining exemplar’s discriminant information and can be trained under an imbalanced setting. Our approach outperforms state-of-the-art exemplar-based detection methods when few exemplars are available and even performs on par with other vastly trained detection algorithms. This inspires us to consider a more difficult problem of general object detection in the future. One promising direction is to combine more amounts of diverse exemplars for category detection when handling general object categories with comparably larger intra-class variations.

Acknowledgment

The authors would like to thank the associate editor and the three reviewers for their constructive comments and suggestions.

References

Xinge You (SM’14) received the B.S. and M.S. degrees in mathematics from Hubei University, Wuhan, China, and the Ph.D. degree from the Department of Computer Science, Hong Kong Baptist University, Hong Kong, in 1990, 2000, and 2004, respectively.

He is a Professor with the Department of Electronics and Information Engineering, Huazhong University of Science and Technology, Wuhan, China. His research interests include wavelets and its applications, signal and image processing, pattern recognition, machine learning, and computer vision.

Qiang Li received the B.Eng. degree in electronics and information engineering and the M.Eng. degree in signals and information processing from Huazhong University of Science and Technology, Wuhan, China, in 2010 and 2013, respectively. He is currently working toward the Ph.D. degree from University of Technology, Sydney, Australia.

His research interests include probabilistic inference, probabilistic graphical models, image processing, video surveillance, and computer vision.

Dacheng Tao (M’07–SM’12) is a Professor of Computer Science with the Centre for Quantum Computation and Intelligent Systems and the Faculty of Engineering and Information Technology with the University of Technology, Sydney, Ultimo, NSW, Australia. He mainly applies statistics and mathematics for data analysis problems in data mining, computer vision, machine learning, multimedia, and video surveillance. He has authored and co-authored more than 100 scientific articles at top venues, including IEEE T-PAMI, T-NNLS, T-IP, NIPS, ICML, AISTATS, ICDM, CVPR, ICCV, ECCV; ACM T-KDD, Multimedia and KDD, with the Best Theory/Algorithm Paper Runner Up Award in IEEE ICDM’07 and the Best Student Paper Award in IEEE ICDM’13.

Weihua Ou was born in Hunan, China, in 1979. He received the M.S. degree in mathematics from Southeast University, Nanjing, China, in 2006. He is currently working toward the Ph.D. degree with Huazhong University of Science and Technology, Wuhan, China.

His research interests include pattern recognition, machine learning, and computer vision such as the application of sparse or low-rank representation in image processing and computer vision.