Training-Free, Generic Object Detection Using Locally Adaptive Regression Kernels

Hae Jong Seo, *Student Member*, *IEEE*, and Peyman Milanfar, *Fellow*, *IEEE*

Abstract—We present a generic detection/localization algorithm capable of searching for a visual object of interest without training. The proposed method operates using a *single* example of an object of interest to find similar matches, does not require prior knowledge (learning) about objects being sought, and does not require any preprocessing step or segmentation of a target image. Our method is based on the computation of local regression kernels as descriptors from a query, which measure the likeness of a pixel to its surroundings. Salient features are extracted from said descriptors and compared against analogous features from the target image. This comparison is done using a matrix generalization of the cosine similarity measure. We illustrate optimality properties of the algorithm using a naive-Bayes framework. The algorithm yields a scalar resemblance map, indicating the likelihood of similarity between the query and all patches in the target image. By employing nonparametric significance tests and nonmaxima suppression, we detect the presence and location of objects similar to the given query. The approach is extended to account for large variations in scale and rotation. High performance is demonstrated on several challenging data sets, indicating successful detection of objects in diverse contexts and under different imaging conditions.

Index Terms—Object detection, image representation, correlation and regression analysis.

1 Introduction

Analysis of visual objects in images is a very important component in computer vision systems which perform object recognition, image retrieval, image registration, and more. Areas where such systems are deployed are diverse and include such applications as surveillance (security), video forensics, and medical image analysis for computer-aided diagnosis, to mention just a few. In particular, the object recognition problem has attracted much attention recently due to the increasing demand for developing real-world systems.

Recognition is mainly divided into two parts: category recognition (classification) and detection/localization [1]. The goal of object category recognition is to classify a given object into one of the several prespecified categories, while object detection is to separate objects of interest from the background in a target image. In the current literature, a popular object recognition paradigm is *probabilistic constellation* [2] or *parts-and-shape models* [3] that represent not only the statistics of individual parts, but also their spatial layout. These are based on learning-based classifiers, which require an intensive learning/training phase of the classifier parameters and thus are called parametric methods. Object detection is also a critical part in many applications such as image retrieval, scene understanding, and surveillance system; however, it is still an open problem because the intraclass variations make generic detection very complicated, requiring various types of preprocessing steps. The sliding window scheme is usually used by taking the peak confidence values as an indication of the presence of an object in a given region. Most successful localization methods at the recent PASCAL VOC 2006 challenge [4] on object localization relied on this technique too, but these too still required a training phase. To make a real-time object detection system while achieving high detection rates, methods combining classifiers in a cascade [5], [6] have been proposed.

Recently, the recognition task with only one query (training-free) has received increasing attention [7], [8], [9], [10] for important applications such as automatic passport control at airports, where a single photo in the passport is the only example available. Another application is in image retrieval from the Web [2], [7]. In the retrieval task, a single probe or query image is provided by users and every gallery image in the database is compared with the single probe, posing an image-to-image matching problem. Recently, the face image retrieval task led to intensive activity in this area, culminating in the Face Recognition Grand Challenge (FRGC) [11]. More generally, by taking into account a set of images that represents intraclass variations, more robust object recognition can be achieved. Such sets may consist of observations acquired from a video sequence or by multiple still shots. In other words, classifying an unknown set of images into one of the training classes can be achieved through set-to-image or set-to-set matching [10] without an intensive training phase. As a successful example of set-to-image matching, Boiman et al. [12] very recently showed that a trivial nearest neighbor (NN)-based image classifier in the space of the local image descriptors such as SIFT [13] and local self-similarity [7] can even outperform the leading learning-based image classifiers such as SVM-KNN [14], pyramid match kernel (PMK) [15], and more.

• *The authors are with the University of California, Santa Cruz, 1156 High Street, Mailcode SOE2, Santa Cruz, CA 95064.*
E-mail: {rokaf, milanfar}@soe.ucsc.edu.

Manuscript received 2 Oct. 2008; revised 17 Feb. 2009; accepted 11 July 2009; published online 5 Aug. 2009.
Recommended for acceptance by B.S. Manjunath.
For information on obtaining reprints of this article, please send e-mail to: tpami@computer.org, and reference IEEECS Log Number TPAMI-2008-10-0662.
Digital Object Identifier no. 10.1109/TPAMI.2009.153.

1.1 Problem Specification

Inspired by this trend toward training-free image analysis, this paper addresses the generic detection/localization problem of searching for an object of interest (for instance, a picture of a face) within other "target" images with only a *single* "query" image. In order to avoid the disadvantages of learning-based methods, which require a large amount of training examples, can result in overfitting of parameters, and are generally slow in the training phase, we focus on a novel and sophisticated feature and a reliable similarity measure for comparing a collection of features.

In general, the target images may contain such similar objects (say, other faces) but these will typically appear in a completely different context and under different imaging conditions. Examples of such differences can range from rather simple optical or geometric differences (such as occlusion, differing viewpoints, lighting, and scale changes) to more complex inherent structural differences such as, for instance, a hand-drawn sketch of a face rather than a real face. As an example, we refer the reader to Fig. 3a. To date, many methods based on such features as histograms, gradients, and shape descriptors have been proposed to address this problem. We refer the interested reader to [16] and the references therein for a good summary.

1.2 Overview of the Proposed Approach

In this paper, our contributions to the object detection task are twofold. First, we propose using local regression kernels as descriptors, which capture the underlying local structure of the data exceedingly well, even in the presence of significant distortions. Second, we propose a novel approach to the detection problem using a nonparametric nearest neighbor classifier, along with a generalization of the cosine similarity to the matrix case. The origin and motivation behind the use of these local kernels is the earlier work on adaptive kernel regression for image processing and reconstruction [17]. In that work, localized nonlinear filters were derived which adapt themselves to the underlying structure of the image in order to very effectively perform denoising, interpolation, and deblurring [18]. The fundamental component of the so-called *steering* kernel regression method is the calculation of the local steering kernel (LSK), which essentially measures the local similarity of a pixel to its neighbors both geometrically and photometrically. The key idea is to robustly obtain local data structures by analyzing the photometric (pixel value) differences based on estimated gradients and use this structure information to determine the shape and size of a canonical kernel. Denoting the target image (T) and the query image (Q), we compute a dense set of local steering kernels from each. These densely computed descriptors are highly informative, but when taken together, they tend to be overcomplete (redundant). Therefore, we derive features by applying dimensionality reduction (namely, PCA) to these resulting arrays, in order to retain only the salient characteristics of the local steering kernels. Generally, T is bigger than the query image Q. Hence, we divide the target image T into a set of overlapping patches which are the same size as Q and assign a class to each patch (T_i). The feature vectors that belong to a patch are thought of as training examples in the corresponding class (See Fig. 2). The feature collections from Q and T_i form feature matrices $\mathbf{F}_Q$ and $\mathbf{F}_{T_i}$. We compare the feature matrices $\mathbf{F}_{T_i}$ and $\mathbf{F}_Q$ from the ith patch of T and Q to look for matches. Inspired in part by the many studies [19], [20], [21], [22], [23], [24] which took advantage of cosine similarity over the conventional euclidean distance, we employ and justify the use of "Matrix Cosine Similarity" as a similarity measure which generalizes the cosine similarity between two vectors [25], [26], [27] to the matrix case. We illustrate the optimality properties of the proposed approach using a naive Bayes framework, which leads to the use of the Matrix Cosine Similarity (MCS) measure. Furthermore, we indicate how this measure can be efficiently implemented using a nearest neighbor formulation. In order to deal with the case where the target image may not include any objects of interest or when there are more than one object in the target, we also adopt the idea of a significance test and nonmaxima suppression [28].

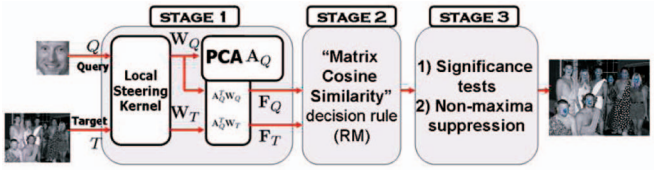


Fig. 1. System overview (there are broadly three stages).

Recently, Shechtman and Irani [7] introduced a related matching framework based on the so-called "local self-similarity" descriptor. It is worth mentioning that this (independently derived) local self-similarity measure is a special case of the local steering kernel and is also related to a number of other local data adaptive metrics, such as Optimal Spatial Adaptation (OSA) [29] and Nonlocal Means (NLM) [30], which have been used for restoration in the image processing community. While the local self-similarity descriptors in [7] were modeled as a function of a simple sum of squared difference (SSD) between a center image patch and surrounding image patches, local regression kernels are designed to have more sophisticated mechanisms to robustly obtain the local structure of images even in the presence of data uncertainty such as noise and blur. It is the aim of this paper to begin the process of applying the local regression kernel idea (in particular, the local *steering* kernel) to problems involving detection of similarity across images and, later, videos. It is worth noting that our contribution in this paper is intended in the same vein as the recent trend toward more extensive use of statistical signal processing and information theory, as nicely exemplified by the works [31]. Fig. 1 shows an overview of our proposed framework. The first stage consists of computing the normalized LSKs $\mathbf{W}_Q, \mathbf{W}_T$ and obtaining the salient feature matrices $\mathbf{F}_Q, \mathbf{F}_T$. In the second stage, we compare the feature matrices $\mathbf{F}_{T_i}$ and $\mathbf{F}_Q$ using the MCS measure. The final output is given after a sequence of significance tests, followed by nonmaxima suppression [28].

Before we begin a more detailed description, it is worthwhile to highlight some aspects of the proposed framework.

- Since the calculation of local regression kernels is stable in the presence of uncertainty in the data [17], our approach is robust even in the presence of noise. In addition, normalized local regression kernels

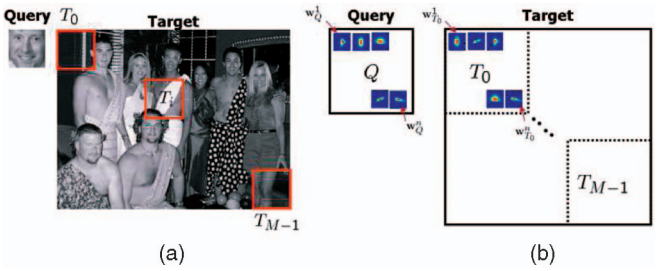


Fig. 2. (a) Given a query image Q, we want to detect/localize objects of interest in a target image T. T is divided into a set of overlapping patches. (b) Local steering kernels represent the geometric structure of underlying data.

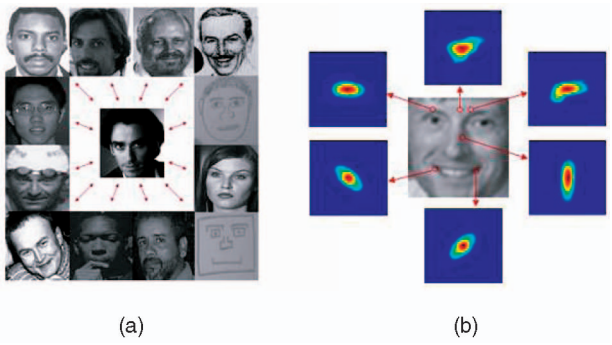


Fig. 3. (a) A face and some possibly similar images. (b) Examples of LSK in various regions.

provide a certain invariance to illumination changes (see Fig. 4.)

- The approach in [7], similar to selective feature techniques such as SIFT [16], filters out "noninformative" descriptors, while in our method we apply Principal Components Analysis (PCA) to a collection of LSKs in order to learn the most salient features of the data.
- While Shechtman and Irani [7] explicitly model local and global geometric relationship between features, we simply propose using Matrix Cosine Similarity, which is a generalized version of the cosine similarity that has been shown to outperform the conventional euclidean distance for subspace learning and classification tasks [19], [20], [21], [22], [23], [24]. We further propose "Canonical Cosine Similarity" to extend the proposed framework to the case of vector data such as a color image. As we shall see in Section 4.3, the Canonical Cosine Similarity is related to the concept of Canonical Correlation analysis [33].
- We employ nearest neighbor classification [12] to solve the object detection problem and show that, under the naive-Bayes assumption, theoretically optimal Bayes decision rule is approximated by the MCS measure. This is in the same spirit as [21], which shows that the Bayes decision rule can be deduced by the whitened cosine similarity under four strong assumptions.
- From a practical standpoint, it is important to note that the proposed framework operates using a single example of an image of interest to find similar matches, does not require any prior knowledge (learning) about objects being sought, and does not require any preprocessing step or segmentation of the target image.

The proposed framework is general enough to be extendable to 3D for such applications as action recognition [8], [10], suspicious behavior detection [34], etc., using an analogous 3D local steering kernel [35]. The discussion of this aspect of the ongoing work is outside the scope of this paper. The paper is organized as follows: In Section 2, we specify the algorithmic aspects of our object detection framework, using a novel feature (the local *steering* kernel) and a reliable similarity measure (the Matrix Cosine Similarity). Section 3 provides a theoretical formulation and justification of the proposed method. In Section 4, we extend the proposed method to more general scenarios, accounting for larger variations in scale and rotation, and for color images by introducing Canonical Cosine Similarity. In Section 5, we demonstrate the performance of the system with some experimental results, and finally, we conclude the paper in Section 6.

2 Technical Detail of the Framework

As outlined in the previous section, our approach to detect objects consists broadly of three stages. Below, we describe each of these steps in detail.

2.1 Extracting Features from the Local Steering Kernel Descriptors

The key idea behind local steering kernel is to robustly obtain the local structure of images by analyzing the photometric (pixel value) differences based on estimated gradients, and to use this structure information to determine the shape and size of a canonical kernel. The local kernel $K(\cdot)$ is modeled as a radially symmetric function:

$$K(\mathbf{x}_l - \mathbf{x}; \mathbf{H}_l) = \frac{K\big(\mathbf{H}_l{}^{-1}(\mathbf{x}_l - \mathbf{x})\big)}{\det(\mathbf{H}_l)}, l = 1, \ldots, P^2, \quad (1)$$

where $\mathbf{x}_l = [x_1, x_2]_l^T$ is the spatial coordinates, P^2 is the number of pixels in a local window ($P \times P$), and the so-called *steering* matrix is defined as

$$\mathbf{H}_l = h\mathbf{C}_l^{-\frac{1}{2}} \in \mathbb{R}^{(2\times 2)}, \quad (2)$$

where h is a global smoothing parameter and the matrix $\mathbf{C}_l$ is a covariance matrix estimated from a collection of spatial (x_1, x_2) gradient vectors within the local analysis window around a position $\mathbf{x}$. The *steering* matrix $\mathbf{H}_l$ modifies the shape and size of the local kernel in a way that roughly encodes the local geometric structures present in the image (See Fig. 3b for an example.) With such steering matrices, we choose a Gaussian function for $K(.)$, which leads to the following form for the LSKs:

$$K(\mathbf{x}_l - \mathbf{x}; \mathbf{H}_l) = \frac{\sqrt{\det(\mathbf{C}_l)}}{2\pi h^2} \exp\left\{-\frac{(\mathbf{x}_l - \mathbf{x})^T \mathbf{C}_l (\mathbf{x}_l - \mathbf{x})}{2h^2}\right\}. \quad (3)$$

We provide some discussion of this choice below, but for a more in-depth analysis, we refer the interested reader to [17]. In what follows, at a position $\mathbf{x}$, we will essentially be

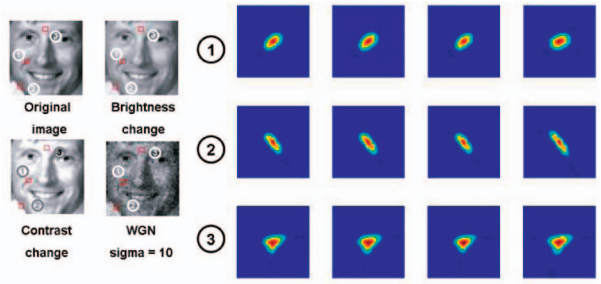


Fig. 4. Invariance and robustness of local *steering* kernel weights $W(\mathbf{x}_l - \mathbf{x}; 2)$ in various challenging conditions. Note that WGN means white Gaussian noise.

using (a normalized version of) the function $K(\mathbf{x}_l - \mathbf{x}; \mathbf{H}_l)$ as a function of $\mathbf{x}_l$ and $\mathbf{H}_l$ to represent an image's inherent local geometry. To be more specific, the local steering kernel function $K^j(\mathbf{x}_l - \mathbf{x}; \mathbf{H}_l)$ at a patch indexed by j is densely calculated and normalized as follows:

$$W_Q^j(\mathbf{x}_l - \mathbf{x}) = \frac{K_Q^j(\mathbf{x}_l - \mathbf{x}; \mathbf{H}_l)}{\sum_{l=1}^{P^2} K_Q^j(\mathbf{x}_l - \mathbf{x}; \mathbf{H}_l)}, \begin{cases} j = 1, \ldots, n, \\ l = 1, \ldots, P^2, \end{cases}$$
$$W_T^j(\mathbf{x}_l - \mathbf{x}) = \frac{K_T^j(\mathbf{x}_l - \mathbf{x}; \mathbf{H}_l)}{\sum_{i=1}^{P^2} K_T^j(\mathbf{x}_l - \mathbf{x}; \mathbf{H}_l)}, \begin{cases} j = 1, \ldots, n_T, \\ l = 1, \ldots, P^2, \end{cases} \quad (4)$$

where n and n_T are the number of patches where LSKs are computed in the query image Q and the target image T, respectively.[1] Next, we describe some key properties of the above.

Takeda et al. [17] showed that LSK based on the locally quadratic data model (regression order $N = 2$) consistently outperforms steering kernels based on the locally constant and the locally linear model (regression order $N = 0$ and $N = 1$) in their kernel regression framework for the tasks of image denoising and interpolation. They further provided the so-called "equivalent kernel" formulation, which is a computationally more efficient and intuitive solution to kernel regression. To simplify the notation, we describe the normalized local steering kernels with the regression order N as $W(\mathbf{x}_l - \mathbf{x}; N)$. We observe that 2nd order LSK $W(\mathbf{x}_l - \mathbf{x}; 2)$ provides better descriptive powers than zeroth order LSK $W(\mathbf{x}_l - \mathbf{x}; 0)$ and first order LSK $W(\mathbf{x}_l - \mathbf{x}; 1)$ even in complex texture regions or in the presence of moderate levels of noise. Normalization of this kernel function yields invariance to brightness change and robustness to contrast change as shown in Fig. 4. When large amounts of noise are present, the locally quadratic data model tends to be more sensitive to noise than the locally linear and the locally constant model. Hence, there is a trade-off between descriptive power of LSK and sensitivity to noise. Recently, Han and Vasconcelos [36] have proposed complex feature selection based on discriminant saliency for object classification. They showed that complex discriminant features tend to improve the performance of training-based image classifiers. Meanwhile, many studies [12], [37], [38] have shown that densely computed local image features give better results in classification tasks than key-point-based local image features, such as SIFT [13], which are designed for mainly invariance and compact coding. According to these studies, the distribution of the local image feature both in natural images as well as images of a specific object class follows a power-law (i.e., a long-tail) distribution [12], [37], [38]. In other words, the features are scattered out in a high-dimensional feature space and thus there basically exists no dense cluster in the feature space. In order to illustrate and verify that the normalized LSKs also satisfy this property as described in [7], [12] and follow a power-law distribution, we computed an empirical bin density (100 bins) of the normalized LSKs (using a total of 31,319 LSKs) densely computed from 60 images (from Shechtman's general object data set [7]) using the K-means clustering method. (See Fig. 5 for an example.)

1. Note that images here are gray scale (luminance channel only). In Section 4.3, we will deal with color images as well.

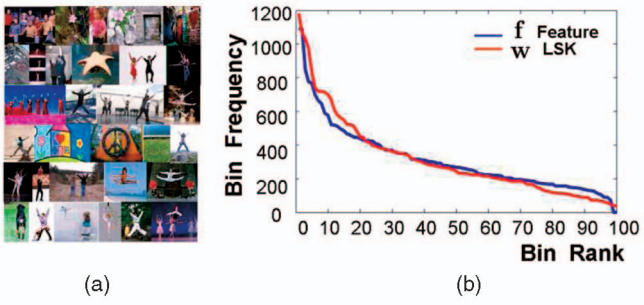


Fig. 5. (a) Some example images (Shechtman and Irani's object data set [7]) where LSKs were computed. (b) Plots of the bin density of LSKs and their corresponding low-dimensional features.

Boiman et al. [12] observed that while an ensemble of local features with little discriminative power can together offer a significant discriminative power, both quantization and informative feature selection on a long-tail distribution can lead to a precipitous drop in performance. Therefore, instead of any quantization and informative feature selection, we focus on reducing the dimension of densely computed LSKs using PCA to enhance the discriminative power and reduce computational complexity. It is worth noting that this approach was also taken by Ke and Sukthankar in [39], where PCA was applied to SIFT features, leading to enhanced performance. Ali and Shah [40] also applied PCA to derive salient kinematic features from optical flow in the action recognition task. This idea results in a new feature representation with a moderate dimension, which inherits the desirable discriminative attributes of LSK. The distribution of the resulting features sitting on the low-dimensional manifold also tends to follow a power-law distribution as shown in Fig. 5b and this attribute of the features will be utilized in applying a nearest neighbor approximation in the theoretical formulation in Section 3.

2.1.1 Feature Representation

In order to organize $W_Q^j(\mathbf{x}_l - \mathbf{x})$ and $W_T^j(\mathbf{x}_l - \mathbf{x})$, which are densely computed from Q and T, let $\mathbf{W}_Q, \mathbf{W}_T$ be matrices whose columns are vectors $\mathbf{w}_Q^j, \mathbf{w}_T^j$, which are column-stacked (rasterized) versions of $W_Q^j(\mathbf{x}_l - \mathbf{x}), W_T^j(\mathbf{x}_l - \mathbf{x})$, respectively:

$$\mathbf{W}_Q = \left[\mathbf{w}_Q^1, \ldots, \mathbf{w}_Q^n\right] \in \mathbb{R}^{P^2 \times n},$$
$$\mathbf{W}_T = \left[\mathbf{w}_T^1, \ldots, \mathbf{w}_T^{n_T}\right] \in \mathbb{R}^{P^2 \times n_T}. \quad (5)$$

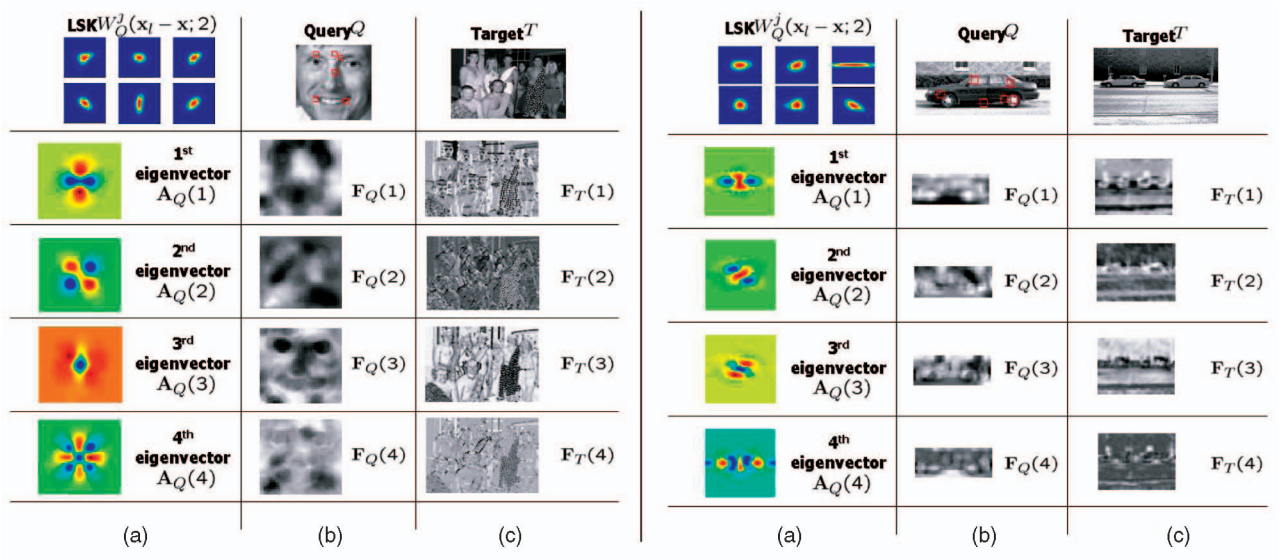


Fig. 6. Face and car examples: (a) $\mathbf{A}_Q$ learned from a collection of LSKs $\mathbf{W}_Q$, (b) feature row vectors of $\mathbf{F}_Q$ from query Q, (c) feature row vectors $\mathbf{F}_T$ from target image T. Eigenvectors and feature vectors were reshaped into image and upscaled for illustration purpose.

As described in Fig. 1, the next step is to apply PCA[2] to $\mathbf{W}_Q$ for dimensionality reduction and to retain only its salient characteristics. By applying PCA to $\mathbf{W}_Q$, we can retain the first (largest) d principal components,[3] which form the columns of a matrix $\mathbf{A}_Q \in \mathbb{R}^{P^2 \times d}$. Next, the lower dimensional features are computed by projecting $\mathbf{W}_Q$ and $\mathbf{W}_T$ onto $\mathbf{A}_Q$:

$$\begin{aligned} \mathbf{F}_Q &= \left[\mathbf{f}_Q^1, \ldots, \mathbf{f}_Q^n\right] = \mathbf{A}_Q^T \mathbf{W}_Q \in \mathbb{R}^{d \times n}, \\ \mathbf{F}_T &= \left[\mathbf{f}_T^1, \ldots, \mathbf{f}_T^{n_T}\right] = \mathbf{A}_Q^T \mathbf{W}_T \in \mathbb{R}^{d \times n_T}. \end{aligned} \tag{6}$$

Fig. 6 illustrates the principal components in $\mathbf{A}_Q$ and shows what the features $\mathbf{F}_Q, \mathbf{F}_T$ look like for some examples such as face and car.

2.2 Matrix Cosine as a Measure of Similarity

The next step in the proposed framework is a decision rule based on the measurement of a "distance" between the computed features $\mathbf{F}_Q, \mathbf{F}_{T_i}$. Earlier works, such as [19], [20], [24], have shown that correlation-based metrics outperform the conventional euclidean and Mahalanobis distances for the classification and subspace learning tasks. Motivated by the effectiveness of correlation-based similarity measure, we introduce Matrix Cosine Similarity for the matrix case and explore the idea behind this measure in this section. In general, "correlation" indicates the strength and direction of a linear relationship between two random variables. But the idea of correlation is quite malleable. Indeed, according to Rodgers and Nicewander [27], there are at least 13 distinct ways to look at correlation. However, we are interested in two main types of correlation: the Pearson's correlation coefficient, which is the familiar standard correlation coefficient, and the cosine similarity (so-called non-Pearson compliant). Note that the cosine similarity coincides with Pearson's correlation when each vector is centered to have zero mean. In several earlier papers, including [25], [26], it has been shown that Pearson correlation is less discriminating than the cosine similarity due to the fact that centered values are less informative than the original values and the computation of centered values is sensitive to zero or small values in the vectors. Since the discriminative power is critical in our detection framework, we focus on the cosine similarity. The cosine similarity is defined as the inner product between two normalized vectors as follows:

$$\rho(\mathbf{f}_Q, \mathbf{f}_{T_i}) = \left\langle \frac{\mathbf{f}_Q}{\|\mathbf{f}_Q\|}, \frac{\mathbf{f}_{T_i}}{\|\mathbf{f}_{T_i}\|} \right\rangle = \frac{{\mathbf{f}_Q}^T \mathbf{f}_{T_i}}{\|\mathbf{f}_Q\| \|\mathbf{f}_{T_i}\|} = \cos\theta_i \in [-1, 1], \tag{7}$$

where $\mathbf{f}_Q, \mathbf{f}_{T_i} \in \mathbb{R}^d$ are column vectors. The cosine similarity measure, therefore, focuses only on the angle (phase) information while discarding the scale information.

If we deal with the features $\mathbf{F}_Q, \mathbf{F}_{T_i}$ that consist of a set of vectors, "Matrix Cosine Similarity" can be defined as a natural generalization using the "Frobenius inner product" between two normalized matrices as follows:

$$\begin{aligned} \rho(\mathbf{F}_Q, \mathbf{F}_{T_i}) &= <\overline{\mathbf{F}}_Q, \overline{\mathbf{F}}_{T_i}>_F \\ &= \operatorname{trace}\left(\frac{\mathbf{F}_Q^T \mathbf{F}_{T_i}}{\|\mathbf{F}_Q\|_F \|\mathbf{F}_{T_i}\|_F} \right) \in [-1, 1], \end{aligned} \tag{8}$$

where $\overline{\mathbf{F}}_Q = [\frac{\mathbf{f}_Q^1}{\|\mathbf{F}_Q\|_F}, \ldots, \frac{\mathbf{f}_Q^n}{\|\mathbf{F}_Q\|_F}]$ and $\overline{\mathbf{F}}_{T_i} = [\frac{\mathbf{f}_{T_i}^1}{\|\mathbf{F}_{T_i}\|_F}, \ldots, \frac{\mathbf{f}_{T_i}^n}{\|\mathbf{F}_{T_i}\|_F}]$.

It is worth noting that this generalization is also known as "vector correlation" in the statistics literature [43]. Fu et al. [19] also applied a generalized cosine similarity to the tensor case for subspace learning and showed performance improvement in the task of image classification. Returning to our definition, if we look at (8) carefully, it is interesting to note that one can rewrite it as a weighted average of the cosine similarities $\rho(\mathbf{f}_Q, \mathbf{f}_{T_i})$ between each pair of corresponding feature vectors (i.e., columns) in $\mathbf{F}_Q, \mathbf{F}_{T_i}$ as follows:

2. It is worth noting that the use of the PCA here may not be critical in the sense that any unsupervised subspace learning method such as Kernel PCA, LLE [41], LPP [42] CDA [24], CPCA [19], and CEA [19] can be used.

3. Typically, d is selected to be a small integer such as 3 or 4 so that 80-90 percent of the "information" in the LSKs would be retained (i.e., $\frac{\sum_{i=1}^{d} \lambda_i}{\sum_{i=1}^{P^2} \lambda_i} \geq 0.8$ (to 0.9), where λ_i are the eigenvalues).

$$\rho(\mathbf{F}_Q, \mathbf{F}_{T_i}) = \sum_{\ell=1}^{n} \frac{\mathbf{f}_Q^{\ell\,T} \mathbf{f}_{T_i}^{\ell}}{\|\mathbf{F}_Q\|_F \|\mathbf{F}_{T_i}\|_F} = \sum_{\ell=1}^{n} \rho\left(\mathbf{f}_Q^{\ell}, \mathbf{f}_{T_i}^{\ell}\right) \frac{\|\mathbf{f}_Q^{\ell}\| \|\mathbf{f}_{T_i}^{\ell}\|}{\|\mathbf{F}_Q\|_F \|\mathbf{F}_{T_i}\|_F}. \quad (9)$$

The weights are represented as the product of

$$\frac{\|\mathbf{f}_Q^{\ell}\|}{\|\mathbf{F}_Q\|_F} \quad \text{and} \quad \frac{\|\mathbf{f}_{T_i}^{\ell}\|}{\|\mathbf{F}_{T_i}\|_F},$$

which indicate the relative importance of each feature in the feature sets $\mathbf{F}_Q, \mathbf{F}_{T_i}$. We see here an advantage of the MCS in that it takes care of the strength and angle similarity of vectors at the same time. Hence, this measure not only generalizes the cosine similarity, but also overcomes the disadvantages of the conventional euclidean distance, which is sensitive to outliers. We compute $\rho(\mathbf{F}_Q, \mathbf{F}_{T_i})$ over all of the target patches and this can be efficiently implemented by column-stacking the matrices $\mathbf{F}_Q, \mathbf{F}_{T_i}$ and simply computing the cosine similarity between two long column vectors as follows:

$$\begin{aligned} \rho_i &\equiv \rho(\mathbf{F}_Q, \mathbf{F}_{T_i}) \\ &= \sum_{\ell=1}^{n} \frac{\mathbf{f}_Q^{\ell\,T} \mathbf{f}_{T_i}^{\ell}}{\|\mathbf{F}_Q\|_F \|\mathbf{F}_{T_i}\|_F} \\ &= \sum_{\ell=1,j=1}^{n,d} \frac{f_Q^{(\ell,j)} f_{T_i}^{(\ell,j)}}{\sqrt{\sum_{\ell=1,j=1}^{n,d} |f_Q^{(\ell,j)}|^2} \sqrt{\sum_{\ell=1,j=1}^{n,d} |f_{T_i}^{(\ell,j)}|^2}} \\ &= \rho(\mathrm{colstack}(\mathbf{F}_Q), \mathrm{colstack}(\mathbf{F}_{T_i})) \in [-1, 1], \end{aligned} \quad (10)$$

where $f_Q^{(\ell,j)}, f_{T_i}^{(\ell,j)}$ are elements in lth vector $\mathbf{f}_Q^{\ell}$ and $\mathbf{f}_{T_i}^{\ell}$, respectively, and $\mathrm{colstack}(\cdot)$ means an operator which column-stacks (rasterizes) a matrix.

In Section 4, we will show that this idea enables us to further generalize the cosine similarity to a "Canonical Cosine Similarity," which is a corresponding version of the canonical correlation analysis (CCA) [33] for the vector data case where we have a set of features separately computed from multiple sources (for instance, color image (YCbCr or CIE L*a*b*) or a sequence of images). In a similar vein as Boiman et al. [12], we will show in Section 3 that a particular version of optimal naive Bayes decision rule can actually lead to the use of MCS measure.

The next step is to generate a so-called "resemblance map" (RM), which will be an image with values indicating the likelihood of similarity between Q and T. When it comes to interpreting the value of "correlation," it is noted in [44], [45] that $\rho_i^2 \in [0, 1]$ describes the proportion of variance in common between the two feature sets as opposed to ρ_i, which indicates a linear relationship between the two feature matrices $\mathbf{F}_Q, \mathbf{F}_{T_i}$. At this point, we can use ρ_i directly as a measure of resemblance between the two feature sets. However, the shared variance interpretation of ρ_i^2 has several advantages. In particular, as for the final test statistic comprising the values in the resemblance map, we use the *proportion* of shared variance (ρ_i^2) to that of the "residual" variance $(1 - \rho_i^2)$. More specifically, RM is computed using the mapping function f as follows:

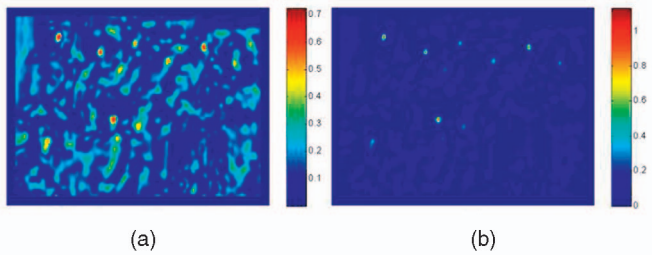


Fig. 7. (a) RM that consists of $|\rho_i|$. (b) RM that consists of $f(\rho_i)$. Note that Q and T are the same examples shown in Fig. 2.

$$\mathrm{RM}: f(\rho_i) = \frac{\rho_i^2}{1 - \rho_i^2}. \quad (11)$$

In Fig. 7, examples of a resemblance map (RM) based on $|\rho_i|$ and $f(\rho_i)$ are presented. Red represents higher resemblance. As is apparent from these typical results, qualitatively, the resemblance map generated from $f(\rho_i)$ provides better contrast and dynamic range in the result $(f(\rho_i) \in [0, \infty])$. More importantly, from a quantitative point of view, we note that $f(\rho_i)$ is essentially the Lawley-Hotelling Trace statistic [33], [46], which is used as an efficient test statistic for detecting correlation between two data sets. Furthermore, historically, this statistic has been suggested in the pattern recognition literature as an effective means of measuring the separability of two data clusters (e.g., [47]).

2.3 Significance Test and Nonmaxima Suppression

If the task is to find the most similar patch (T_i) to the query (Q) in the target image, one can choose the patch which results in the largest value in the RM (i.e., $\max f(\rho_i)$) among all of the patches, no matter how large or small the value is in the range of $[0, \infty]$. This, however, is not wise because there may not be *any* object of interest present in the target image. We are therefore interested in two types of significance tests. The first is an overall test to decide whether there is any sufficiently similar object present in the target image at all. If the answer is yes, we would then want to know how many objects of interest are present in the target image and where they are. Therefore, we need two thresholds: an overall threshold τ_o and a threshold τ to detect the possibly multiple objects present in the target image.

In a typical scenario, we set the overall threshold τ_o to be 0.96, which is about 50 percent of variance in common (i.e., $\rho^2 = 0.49$). In other words, if the maximal $f(\rho_i)$ is just above 0.96, we decide that there exists at least one object of interest. The next step is to choose τ based on the properties of $f(\rho_i)$. When it comes to choosing the τ, there is a need to be more careful. If we have a basic knowledge of the underlying distribution of $f(\rho_i)$, then we can make predictions about how this particular statistic will behave, and thus, it is relatively easy to choose a threshold which will indicate whether the pair of features from the two images are sufficiently similar. But, in practice, we do not have a very good way to model the distribution of $f(\rho_i)$. Therefore, instead of assuming a type of underlying distribution, we employ the idea of nonparametric testing. We compute an empirical PDF from all the given samples of $f(\rho_i)$ and we set τ so as to achieve, for instance, a 99 percent confidence level in deciding whether the given values are in

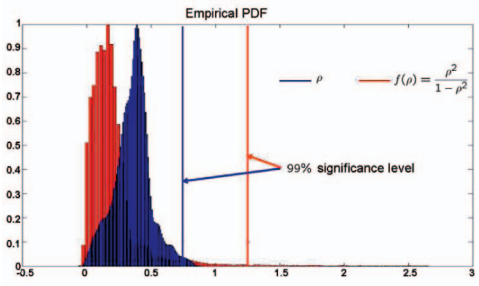


Fig. 8. Comparison of empirical PDF between ρ and $\frac{\rho^2}{1-\rho^2}$.

the extreme (right) tails of the distribution (see Fig. 8).[4] This approach is based on the assumption that, in the target image, most of the patches do not contain the object of interest, and therefore, the few matches will result in values which are in the tails of the distributions of $f(\rho_i)$.

After the two significance tests with τ_o, τ are performed, we employ nonmaxima suppression [28] for the final detection. We take the region with the highest $f(\rho_i)$ value and eliminate the possibility that any other object is detected within some radius[5] of the center of that region again. This enables us to avoid multiple false detections of nearby objects already detected. Then we iterate this process until the local maximum value falls below the threshold τ. Fig. 9 shows the graphical illustration of significance tests and the nonmaxima suppression idea.

3 Theoretical Justification

As explained in the previous section, the purpose of the proposed framework is to detect an object (or objects) of interest in the target image given a *single* query. In this section, we show that the naive-Bayes approach in a multiple hypothesis testing framework leads to the Matrix Cosine Similarity-based decision rule. It is worth noting that this idea is partly motivated by Boiman et al. [12] and Liu [21], who derived an optimal Bayes decision rule based on euclidean distance and the whitened cosine similarity, respectively, for the image classification task.

As described before, the target image T is divided into a set of overlapping patches and a class is assigned to each patch. Our task at hand is to figure out which class (i) the features from Q are most likely to have come from. Since we do not know the class-conditional pdf ($p(\overline{\mathbf{F}}_Q|class)$) of the normalized features extracted from Q, we set out to estimate it using a kernel density estimation method [48]. Once we have these estimates, we will show that the maximum likelihood (ML) decision rule boils down to computing and thresholding Matrix Cosine Similarity, which can be efficiently implemented using a nearest neighbor formulation.

By associating each patch (T_i) of the target image with a hypothesis, we now have the case where we wish to discriminate between M hypotheses ($\mathcal{H}_0, \ldots, \mathcal{H}_{M-1}$) as follows:

4. Yet another justification for using $f(\rho_i)$ instead of ρ_i is the observation that the empirical PDF of ρ_i is itself heavy-tailed, making the detection of rare events more difficult. The use of $f(\rho_i)$ instead tends to alleviate this problem (see Fig. 8).

5. The size of this "exclusion" region will depend on the application at hand and the characteristics of the query image.

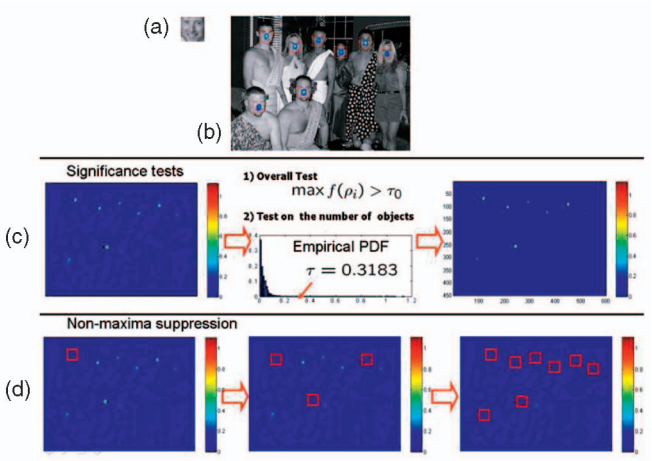


Fig. 9. (a) Query. (b) Target with detection. (c) Two significance tests. (d) Nonmaxima suppression [28].

$\mathcal{H}_0$: Q is similar to $T_0 \Leftrightarrow \overline{\mathbf{F}}_Q$ comes from class 0 ($\overline{\mathbf{F}}_{T_0}$) ,
$\mathcal{H}_1$: Q is similar to $T_1 \Leftrightarrow \overline{\mathbf{F}}_Q$ comes from class 1 ($\overline{\mathbf{F}}_{T_1}$),
$\vdots$ $\vdots$
$\mathcal{H}_{M-1}$: Q is similar to $T_{M-1} \Leftrightarrow \overline{\mathbf{F}}_Q$ comes from class $M-1$ ($\overline{\mathbf{F}}_{T_{M-1}}$).

The task at hand is to find the most likely hypothesis (or a correct class) given the query image Q. It is a well-known fact [47], [49] that maximizing a posteriori probability $P(\mathcal{H}_i|\overline{\mathbf{F}}_Q)$ minimizes Bayes risk (or the average classification error.) Assuming that the prior probabilities $P(\mathcal{H}_i)$ are equal, then the maximum a posterior (MAP) decision rule boils down to the M-ary ML decision rule:

$$\widehat{\mathcal{H}}_i = \arg\max_i P(\mathcal{H}_i|\overline{\mathbf{F}}_Q) = \arg\max_i p(\overline{\mathbf{F}}_Q|\mathcal{H}_i). \qquad (12)$$

Since we do not know the conditional probability density function $p(\overline{\mathbf{F}}_Q|\mathcal{H}_i)$ of features $\overline{\mathbf{F}}_Q$ given the features $\overline{\mathbf{F}}_{T_i}$ of the target patch T_i, we need to estimate it using a kernel density estimation method, which results in the naive or empirical Bayes approach.

3.1 Locally Data-Adaptive Kernel Density Estimation

The Parzen density estimator is a simple and generally accurate nonparametric density estimation method [48]. However, if the true conditional density that we want to model is close to a nonlinear lower dimensional manifold embedded in the higher dimensional feature space, a Parzen density estimator with an isotropic kernel is not the most appropriate method [50], [51], [52]. As explained before, the features $\overline{\mathbf{F}}_Q, \overline{\mathbf{F}}_{T_i}$ tend to generically come from long-tailed distributions, and as such, there are generally no tight clusters in the feature space. When we estimate a probability density at a particular point, for instance, $\overline{\mathbf{f}}_Q^\ell$, the isotropic kernel centered on that point will spread its density mass equally along all the feature space directions, thus, giving too much emphasis to irrelevant regions of space and too little along the manifold. Earlier studies [50], [51], [52] also pointed out this problem. This motivates us to use *a locally data-adaptive* version of the kernel density estimator

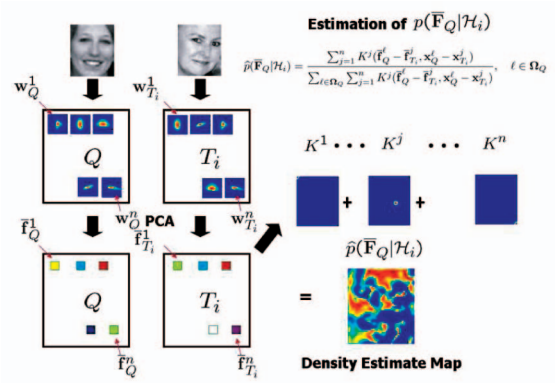


Fig. 10. The estimated conditional density $\widehat{p}(\overline{\mathbf{F}}_Q|\mathcal{H}_i)$ is a sum of kernels (weight functions) centered at the features $\overline{\mathbf{f}}_{T_i}$ in T_i, which belongs to the hypothesis $\mathcal{H}_i$. In the density estimate map, red value means a high conditional probability density $\widehat{p}(\overline{\mathbf{f}}_Q|\mathcal{H}_i)$ while blue value represents a low conditional probability density $\widehat{p}(\overline{\mathbf{f}}_Q|\mathcal{H}_i)$.

The estimated conditional density $\widehat{p}(\overline{\mathbf{F}}_Q|\mathcal{H}_i)$ is defined as a sum of kernels centered at the features $\overline{\mathbf{f}}_{T_i}$ in T_i which belong to the hypothesis $\mathcal{H}_i$. More specifically,

$$\widehat{p}(\overline{\mathbf{F}}_Q|\mathcal{H}_i) = \frac{\sum_{j=1}^{n} K^j\left(\overline{\mathbf{f}}_Q^{\ell} - \overline{\mathbf{f}}_{T_i}^{j}, \mathbf{x}_Q^{\ell} - \mathbf{x}_{T_i}^{j}\right)}{\sum_{\ell \in \mathbf{\Omega}_Q}\sum_{j=1}^{n} K^j\left(\overline{\mathbf{f}}_Q^{\ell} - \overline{\mathbf{f}}_{T_i}^{j}, \mathbf{x}_Q^{\ell} - \mathbf{x}_{T_i}^{j}\right)}, \quad \ell \in \mathbf{\Omega}_Q, \tag{13}$$

where K^j is a locally data adaptive kernel function, $\mathbf{\Omega}_Q$ is the query image domain consisting of $|\mathbf{\Omega}_Q|$ pixels, and $\mathbf{x}_Q^{\ell}, \mathbf{x}_{T_i}^{j}$ are column vectors denoting spatial coordinates of corresponding features $\overline{\mathbf{f}}_Q^{\ell}$ and $\overline{\mathbf{f}}_{T_i}^{j}$. A simple and intuitive choice of the K^j is to consider two terms for penalizing the spatial distance between the point of interest and its neighbors, and the photometric "distance" between the corresponding features $\overline{\mathbf{f}}_Q^{\ell}$ and $\overline{\mathbf{f}}_{T_i}^{j}$. More specifically, the kernel function is defined as follows:

$$\begin{aligned} K^j &= K_r^j\left(\overline{\mathbf{f}}_Q^{\ell} - \overline{\mathbf{f}}_{T_i}^{j}\right) K_s^j\left(\mathbf{x}_Q^{\ell} - \mathbf{x}_{T_i}^{j}\right) \\ &= \exp\left(-\frac{\operatorname{dist}\left(\overline{\mathbf{f}}_Q^{\ell}, \overline{\mathbf{f}}_{T_i}^{j}\right)}{2\sigma_r^2}\right) \exp\left(-\frac{\left\|\mathbf{x}_Q^{\ell} - \mathbf{x}_{T_i}^{j}\right\|^2}{2\sigma_s^2}\right), \quad \ell \in \mathbf{\Omega}_Q, \end{aligned} \tag{14}$$

where we define $\operatorname{dist}(\overline{\mathbf{f}}_Q^{\ell}, \overline{\mathbf{f}}_{T_i}^{j}) = \| \frac{\mathbf{f}_Q^{\ell}}{\|\mathbf{F}_Q\|_F} - \frac{\mathbf{f}_{T_i}^{j}}{\|\mathbf{F}_{T_i}\|_F} \|^2$, and σ_r, σ_s are parameters controlling the falloff of weights in photometric and spatial domains.

Inserting (14) into (13), the estimated conditional density $\widehat{p}(\overline{\mathbf{F}}_Q|\mathcal{H}_i)$ becomes

$$\widehat{p}(\overline{\mathbf{F}}_Q|\mathcal{H}_i) = \frac{1}{\beta}\sum_{j=1}^{n} \exp\left(-\frac{\operatorname{dist}\left(\overline{\mathbf{f}}_Q^{\ell}, \overline{\mathbf{f}}_{T_i}^{j}\right)}{2\sigma_r^2} - \frac{\left\|\mathbf{x}_Q^{\ell} - \mathbf{x}_{T_i}^{j}\right\|^2}{2\sigma_s^2}\right), \tag{15}$$

where $\beta = \sum_{\ell \in \mathbf{\Omega}_Q}\sum_{j=1}^{n} K^j(\overline{\mathbf{f}}_Q^{\ell} - \overline{\mathbf{f}}_{T_i}^{j}, \mathbf{x}_Q^{\ell} - \mathbf{x}_{T_i}^{j})$ is a normalization factor. Fig. 10 depicts how the conditional density function $\widehat{p}(\overline{\mathbf{F}}_Q|\mathcal{H}_i)$ is estimated, given Q and T_i.

In principle, all n features should be employed to obtain an accurate density estimate. However, this is too computationally time-consuming. Hence, as we describe next, we use an efficient approximation of this kernel density estimator.

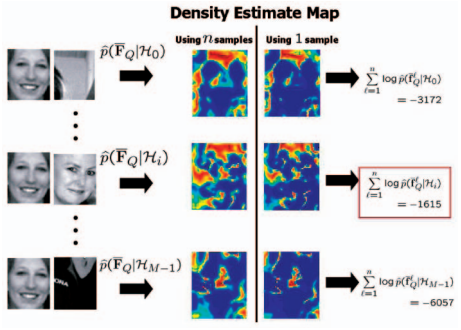


Fig. 11. The estimated conditional probability densities $\widehat{p}(\overline{\mathbf{F}}_Q|\mathcal{H}_i)$ using n samples and one sample are shown in the middle and the scores on the right side mean $\sum_{\ell=1}^{n} \log \widehat{p}(\overline{\mathbf{f}}_Q^{\ell}|\mathcal{H}_i)$. The higher this score is, the more likely it is that $\overline{\mathbf{F}}_Q$ comes from class i ($\overline{\mathbf{F}}_{T_i}$).

3.2 Approximation of Locally Data-Adaptive Kernel Density

Assuming that $\overline{\mathbf{f}}_Q^1, \overline{\mathbf{f}}_Q^2, \ldots, \overline{\mathbf{f}}_Q^n$ are i.i.d. given hypothesis $\mathcal{H}_i$, the ML decision rule can be rewritten by taking the log probability of the ML decision rule (12) as

$$\begin{aligned} \widehat{\mathcal{H}}_i &= \arg\max_i \log \widehat{p}(\overline{\mathbf{F}}_Q|\mathcal{H}_i) = \arg\max_i \log \widehat{p}(\overline{\mathbf{f}}_Q^1, \ldots, \overline{\mathbf{f}}_Q^n|\mathcal{H}_i) \\ &= \arg\max_i \sum_{\ell=1}^{n} \log \widehat{p}(\overline{\mathbf{f}}_Q^{\ell}|\mathcal{H}_i). \end{aligned} \tag{16}$$

What we do next is to estimate each local individual probability density $\widehat{p}(\overline{\mathbf{f}}_Q^{\ell}|\mathcal{H}_i)$ separately:

$$\widehat{p}(\overline{\mathbf{f}}_Q^{\ell}|\mathcal{H}_i) = \frac{1}{\beta'}\sum_{j=1}^{n} K^j\left(\overline{\mathbf{f}}_Q^{\ell} - \overline{\mathbf{f}}_{T_i}^{j}, \mathbf{x}_Q^{\ell} - \mathbf{x}_{T_i}^{j}\right), \quad \ell = 1, \ldots, n, \tag{17}$$

where $\beta' = \sum_{\ell=1}^{n}\sum_{j=1}^{n} K^j(\overline{\mathbf{f}}_Q^{\ell} - \overline{\mathbf{f}}_{T_i}^{j}, \mathbf{x}_Q^{\ell} - \mathbf{x}_{T_i}^{j})$ is a normalization factor. As nicely motivated in [12] and discussed in Section 2.1, since the distribution of the features on the low-dimensional manifold tends to follow a power-law, it should be sufficient to use just a few features in T_i to get a reasonable estimate of the conditional density $\widehat{p}(\overline{\mathbf{f}}_Q^{\ell}|\mathcal{H}_i)$. Therefore, we consider using a single (spatially nearest) neighbor for the approximation, which yields:

$$\begin{aligned} \widehat{p}(\overline{\mathbf{f}}_Q^{\ell}|\mathcal{H}_i) &\approx \exp\left(-\frac{1}{2\sigma_r^2}\operatorname{dist}\left(\overline{\mathbf{f}}_Q^{\ell}, \overline{\mathbf{f}}_{T_i}^{\ell}\right)\right), \quad \ell = 1, \ldots, n, \\ &= \exp\left(\frac{-\left(\frac{\|\mathbf{f}_Q^{\ell}\|^2}{\|\mathbf{F}_Q\|_F^2} + \frac{\|\mathbf{f}_{T_i}^{\ell}\|^2}{\|\mathbf{F}_{T_i}\|_F^2} - \frac{2\rho(\mathbf{f}_Q^{\ell}, \mathbf{f}_{T_i}^{\ell})\|\mathbf{f}_Q^{\ell}\|\|\mathbf{f}_{T_i}^{\ell}\|}{\|\mathbf{F}_Q\|_F\|\mathbf{F}_{T_i}\|_F}\right)}{2\sigma_r^2}\right). \end{aligned} \tag{18}$$

The approximate version of density estimator using one sample is compared to $\widehat{p}(\overline{\mathbf{F}}_Q|\mathcal{H}_i)$ estimated using all n samples in Fig. 11. Qualitatively, we observe that the resulting estimates are quite similar. More precisely, consistent with [12], we have verified that the use of the approximation takes little away from the performance of the overall algorithm.

Since $\log \widehat{p}(\overline{\mathbf{f}}_Q^\ell|\mathcal{H}_i)$ is approximately proportional to

$$-\left(\frac{\|\mathbf{f}_Q^\ell\|^2}{\|\mathbf{F}_Q\|_F^2}+\frac{\|\mathbf{f}_{T_i}^\ell\|^2}{\|\mathbf{F}_{T_i}\|_F^2}-2\rho(\mathbf{f}_Q^\ell,\mathbf{f}_{T_i}^\ell)\frac{\|\mathbf{f}_Q^\ell\|\|\mathbf{f}_{T_i}^\ell\|}{\|\mathbf{F}_Q\|_F\|\mathbf{F}_{T_i}\|_F}\right),$$

the ML decision rule becomes

$$\begin{aligned}\widehat{\mathcal{H}}_i &= \arg\max_i \sum_{\ell=1}^{n}\log\widehat{p}(\overline{\mathbf{f}}_Q^\ell|\mathcal{H}_i)\\ &\Rightarrow \arg\max_i \sum_{\ell=1}^{n}-\left(\frac{\|\mathbf{f}_Q^\ell\|^2}{\|\mathbf{F}_Q\|_F^2}+\frac{\|\mathbf{f}_{T_i}^\ell\|^2}{\|\mathbf{F}_{T_i}\|_F^2}-\frac{2\rho(\mathbf{f}_Q^\ell,\mathbf{f}_{T_i}^\ell)\|\mathbf{f}_Q^\ell\|\|\mathbf{f}_{T_i}^\ell\|}{\|\mathbf{F}_Q\|_F\|\mathbf{F}_{T_i}\|_F}\right)\\ &= \arg\max_i\left(-2+2\sum_{\ell=1}^{n}\frac{\mathbf{f}_Q^{\ell\,T}\mathbf{f}_{T_i}^\ell}{\|\mathbf{F}_Q\|_F\|\mathbf{F}_{T_i}\|_F}\right)\\ &= \arg\max_i \sum_{\ell=1}^{n}\frac{\mathbf{f}_Q^{\ell\,T}\mathbf{f}_{T_i}^\ell}{\|\mathbf{F}_Q\|_F\|\mathbf{F}_{T_i}\|_F}\\ &= \arg\max_i\left\langle\frac{\mathbf{F}_Q}{\|\mathbf{F}_Q\|_F},\frac{\mathbf{F}_{T_i}}{\|\mathbf{F}_{T_i}\|_F}\right\rangle_F.\end{aligned} \tag{19}$$

We can clearly see that the ML decision rule in (19) boils down to the computation of the Matrix Cosine Similarity, due to the relationship

$$\left\langle\frac{\mathbf{F}_Q}{\|\mathbf{F}_Q\|_F},\frac{\mathbf{F}_{T_i}}{\|\mathbf{F}_{T_i}\|_F}\right\rangle_F \approx \frac{2+\sum_{\ell=1}^{n}\log\widehat{p}(\overline{\mathbf{f}}_Q^\ell|\mathcal{H}_i)}{2}.$$

While the assumptions leading to the above conclusions may seem somewhat restrictive, in practice they appear to hold true, and they do provide a framework in which the proposed algorithm can be considered optimal in the naive Bayes sense. Indeed, as can be seen from the practical experimental results in Section 5, the range of applicability of the algorithm thus justified is quite wide. To summarize, the overall pseudocode for the algorithm is given in **Algorithm** 1.

Algorithm 1. Pseudocode for the nonparametric object detection algorithm

Q: Query image, T: Target image, τ_o: Overall threshold, α: Confidence level, P^2: Size of local steering kernel (LSK) window.

Stage 1: Feature representation

1) Construct $\mathbf{W}_Q, \mathbf{W}_T$ (a collection of normalized LSK associated with Q, T)

2) Apply PCA to $\mathbf{W}_Q$ and obtain projection space $\mathbf{A}_Q$ from its top d eigenvectors.

3) Project $\mathbf{W}_Q$ and $\mathbf{W}_T$ onto $\mathbf{A}_Q$ to construct $\mathbf{F}_Q$ and $\mathbf{F}_T$.

Stage 2: Compute Matrix Cosine Similarity

for every target patch T_i, where $i \in [0,\ldots,M-1]$ **do**

$\rho_i = <\frac{\mathbf{F}_Q}{\|\mathbf{F}_Q\|_F},\frac{\mathbf{F}_{T_i}}{\|\mathbf{F}_{T_i}\|_F}>_F$ and compute (RM) : $f(\rho_i)=\frac{\rho_i^2}{1-\rho_i^2}$.

end for

Then, find $\max f(\rho_i)$.

Stage 3: Significance tests and Non-maxima suppression

1) If $\max f(\rho_i) > \tau_o$, go to the next test. Otherwise, there is no object of interest in T.

2) Threshold RM by τ which is set to achieve 99 percent confidence level ($\alpha = 0.99$) from the empirical PDF of $f(\rho_i)$.

3) Apply nonmaxima suppression to RM until the local maximum value is below τ.

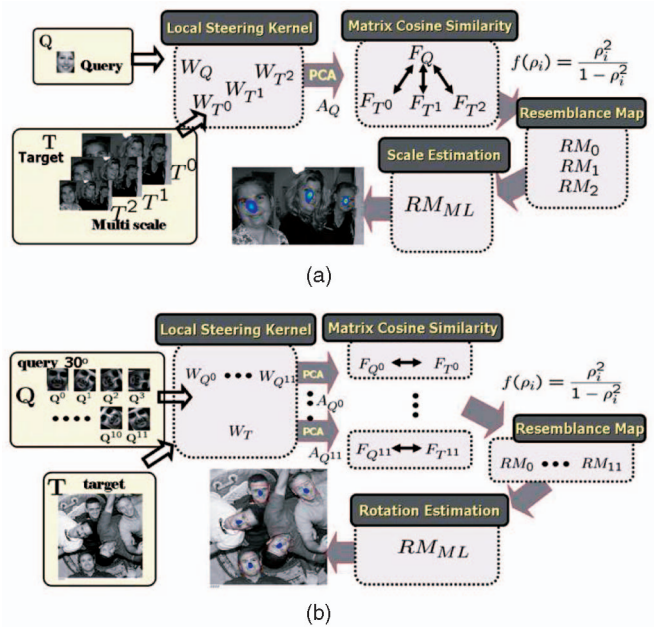


Fig. 12. (a) Block diagram of multiscale object detection system. (b) Block diagram of multirotation object detection system.

4 Handling Variations in Scale and Rotation and Color Images

Up to now, we only dealt with the detection of objects in a gray image at a single scale. Although our detection framework can handle modest scale and rotation variations by adopting a sliding window scheme, robustness to larger scale and rotation changes (for instance, above ± 20 percent in scale, 30 degrees in rotation) are desirable. Furthermore, the use of color images as input should be also considered from a practical point of view. In this section, the approach described in the previous sections for detecting objects at a single scale is extended to detect objects at different scales and at different orientations in an image. In addition, we deal with a color image by defining and using "Canonical Cosine Similarity."

4.1 Multiscale Approach

In order 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 10-15 percent so that a relatively fine quantization of scales is taken into account. Fig. 12a shows the block diagram of the multiscale approach. The first step is to construct the multiscale pyramid $T^0, T^1, \ldots, T^S$, where S is the coarsest scale of the pyramid. As shown in Fig. 12a, $\mathbf{F}_Q, \mathbf{F}_{T^0}, \mathbf{F}_{T^1}, \mathbf{F}_{T^2} (S=2)$ are obtained by projecting $\mathbf{W}_Q$ and $\mathbf{W}_{T^0}, \mathbf{W}_{T^1}, \mathbf{W}_{T^2}$ onto the principal subspace defined by $\mathbf{A}_Q$ as follows:

$$\begin{aligned}\mathbf{F}_Q &= \mathbf{A}_Q^T\mathbf{W}_Q, \quad \mathbf{F}_{T^0} = \mathbf{A}_Q^T\mathbf{W}_{T^0},\\ \mathbf{F}_{T^1} &= \mathbf{A}_Q^T\mathbf{W}_{T^1}, \quad \mathbf{F}_{T^2} = \mathbf{A}_Q^T\mathbf{W}_{T^2}.\end{aligned} \tag{20}$$

We obtain three resemblance maps, RM_0, RM_1, and RM_2, by computing $f(\rho_i)=\frac{\rho_i^2}{1-\rho_i^2}$. These resemblance maps represent the likelihood functions $p(f(\rho_i)|S_i)$, where S_i is the scale at the

ith point. However, the sizes of the respective resemblance maps RM_0, RM_1, and RM_2 are naturally different. Therefore, we simply upscale all of the resemblance maps by pixel replication so that they match the dimensions of the finest scale map RM_0. Next, the maximum likelihood estimate of the scale at each position is arrived at by comparing the upscaled resemblance maps as follows:[6]

$$\hat{S}_i = \arg\max_{S_i} p(\underline{\mathrm{RM}}|S_i). \tag{21}$$

4.2 Multirotation Approach

In order to cope with large rotations, we take a similar approach and generate rotated images (this time, of the query image Q) in roughly 30 degree steps. As seen in Fig. 12b, $\mathbf{F}_{Q^0}, \mathbf{F}_{Q^1}, \ldots, \mathbf{F}_{Q^{11}}$, and $\mathbf{F}_T$ are obtained by projecting $\mathbf{W}_{Q^0}, \ldots, \mathbf{W}_{Q^{11}}$, and $\mathbf{W}_T$ onto the principal subspace defined by $\mathbf{A}_{Q^0}, \ldots, \mathbf{A}_{Q^{11}}$. After computing $f(\rho_i) = \frac{\rho_i^2}{1-\rho_i^2}$ from 12 pairs by employing the sliding window scheme, we obtain 12 resemblance maps $\mathrm{RM}_0, \ldots, \mathrm{RM}_{11}$. We compute the maximum likelihood estimate of the best matching pattern accounting for rotation as follows:

$$\hat{R}_i = \arg\max_{R_i} p(\underline{\mathrm{RM}}|R_i). \tag{22}$$

4.3 Canonical Cosine Similarity

Now, we define Canonical Cosine Similarity (CCS) to extend the proposed framework with a single gray-scale query image to vector-valued images. In particular, suppose, at each pixel, the image has q values. As per the earlier discussion (Section 2.2), we generate q feature sets $\mathbf{F}_Q^\ell, \mathbf{F}_{T_i}^\ell$ $(\ell = [1, \ldots, q])$ by projecting $\mathbf{W}_Q^\ell, \mathbf{W}_{T_i}^\ell$ onto the subspaces $\mathbf{A}_Q^\ell$, respectively, and form the overall feature set as follows:

$$\mathbf{F}_I = \left[\mathrm{colstack}\left(\mathbf{F}_I^1\right), \ldots, \mathrm{colstack}\left(\mathbf{F}_I^q\right)\right] \in \mathbb{R}^{(d\times n)\times q}, I \in \{Q, F_i\}. \tag{23}$$

The key idea is to find vectors $\mathbf{u}_Q$ and $\mathbf{u}_{T_i}$ which maximally correlate two data sets ($\mathbf{F}_Q$, $\mathbf{F}_{T_i}$).

$$\begin{aligned}\mathbf{v}_I = \mathbf{F}_I\mathbf{u}_I &= u_{I_1}\mathrm{colstack}\left(\mathbf{F}_I^1\right)\\ &\quad + \cdots + u_{I_q}\mathrm{colstack}\left(\mathbf{F}_I^q\right) \in \mathbb{R}^{(d\times n)},\end{aligned} \tag{24}$$

where

$$\mathbf{u}_Q = \left[u_{Q_1}, \ldots, u_{Q_q}\right]^T \in \mathbb{R}^q \text{ and } \mathbf{u}_{T_i} = \left[u_{T_1}, \ldots, u_{T_q}\right]^T \in \mathbb{R}^q.$$

Then, the objective function we are maximizing is the cosine similarity as follows:

$$\begin{aligned}\rho = \max_{\mathbf{u}_Q, \mathbf{u}_{T_i}} \frac{\mathbf{v}_Q^T\mathbf{v}_{T_i}}{\|\mathbf{v}_Q\|\|\mathbf{v}_{T_i}\|} = \max_{\mathbf{u}_Q, \mathbf{u}_{T_i}} \frac{\mathbf{u}_Q^T\mathbf{F}_Q^T\mathbf{F}_{T_i}\mathbf{u}_{T_i}}{\|\mathbf{F}_Q\mathbf{u}_Q\|\|\mathbf{F}_{T_i}\mathbf{u}_{T_i}\|},\\ \text{such that } \|\mathbf{F}_Q\mathbf{u}_Q\| = \|\mathbf{F}_{T_i}\mathbf{u}_{T_i}\| = 1,\end{aligned} \tag{25}$$

where $\mathbf{u}_Q$ and $\mathbf{u}_{T_i}$ are called canonical variates. The above is inspired by CCA [33].

6. By $\underline{\mathrm{RM}}$ we mean a collection of RM indexed by i at each position.

The canonical cosine similarity ρ and canonical variates $\mathbf{u}_Q, \mathbf{u}_{T_i}$ can be obtained by solving the coupled eigenvalue problems as follows (derivation is given in the Appendix):

$$\begin{aligned}\left(\mathbf{F}_Q^T\mathbf{F}_Q\right)^{-1}\left(\mathbf{F}_Q^T\mathbf{F}_{T_i}\right)\left(\mathbf{F}_{T_i}^T\mathbf{F}_{T_i}\right)^{-1}\left(\mathbf{F}_{T_i}^T\mathbf{F}_Q\right)\mathbf{u}_Q &= \rho^2\mathbf{u}_Q,\\ \left(\mathbf{F}_{T_i}^T\mathbf{F}_{T_i}\right)^{-1}\left(\mathbf{F}_{T_i}^T\mathbf{F}_Q\right)\left(\mathbf{F}_Q^T\mathbf{F}_Q\right)^{-1}\left(\mathbf{F}_Q^T\mathbf{F}_{T_i}\right)\mathbf{u}_{T_i} &= \rho^2\mathbf{u}_{T_i}.\end{aligned} \tag{26}$$

The positive square root of eigenvalues ρ^2 is the "Canonical Cosine Similarity". If $\mathbf{F}_Q, \mathbf{F}_{T_i}$ are each composed of a single vector $(\mathrm{colstack}(\mathbf{F}_Q), \mathrm{colstack}(\mathbf{F}_{T_i}))$, the above equations reduce to

$$\frac{\left(\mathrm{colstack}(\mathbf{F}_Q)^T\mathrm{colstack}(\mathbf{F}_{T_i})\right)^2}{\|\mathrm{colstack}(\mathbf{F}_Q)\|^2\|\mathrm{colstack}(\mathbf{F}_{T_i})\|^2} = \rho^2,$$

which is just the squared cosine similarity defined earlier in Section 2.2.

Now, we take a closer look at the particular case of color images, where $q = 3$. A natural question here is whether we can gain more if we use the color information instead of using only the luminance channel as we have so far. The answer to this question is positive. There exist many color spaces such as RGB, YCbCr, CIE $\mathrm{L^*a^*b^*}$, etc. We observe that the CIE $\mathrm{L^*a^*b^*}$ color model provides the most discriminative information among all, as also observed by Shechtman and Irani [7]. We define the respective RM[7] as the summation of mapping function $f(\rho_i(\ell))$ of CCS $\rho_i(\ell)$ between a set of features, which are calculated from each channel $(\ell = 1, \ldots, q)$, where $\sum_{\ell=1}^{d_c} \frac{\rho_i^2(\ell)}{1-\rho_i^2(\ell)}$ (d_c is the number of canonical cosine similarity values $\rho_i(\ell)$ greater than zero). Also illustrated in the next section, the color approach based on CCS not only provides better discriminative power, but also gives more accurate localization results than the luminance channel only does.

5 Experimental Results

In this section, we demonstrate the performance of the proposed method with comprehensive experiments on three data sets, namely, the UIUC car data set [53], MIT-CMU face data set [54], and Shechtman's general object data set [7]. The proposed algorithm provides a series of bounding boxes around objects of interest using the criterion described in [53]. More specifically, if the detected region by the proposed method lies within an ellipse of a certain size centered around the ground truth, we evaluate it as a correct detection. Otherwise, it is counted as a false positive. Eventually, we compute *Precision* and *Recall* defined as

$$\mathrm{Recall} = \frac{TP}{nP}, \qquad \mathrm{Precision} = \frac{TP}{TP+FP}, \tag{27}$$

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 $1 - \mathrm{Precision} = \frac{FP}{TP+FP}$.

7. Again as mentioned before, note that $\sum_{\ell=1}^{d_c} \frac{\rho_i^2(\ell)}{1-\rho_i^2(\ell)}$ is analogous to the Lawley-Hotelling trace test statistic $\sum \frac{\rho_i^2}{1-\rho_i^2}$ that is used in the significance test of canonical variates in canonical correlation analysis [33], [46].

Fig. 13. (a) Examples of correct detections on the UIUC single-scale car test set [53]. (b) Examples of correct detections on the UIUC multiscale car test set. Confidence level α was set to 0.99 and RM only above the threshold τ corresponding to α is embedded in test images. Bounding boxes are drawn at the correct locations. In case of a multiple detection, a red bounding box indicates higher resemblance to Query than a blue bounding box.

Experimental results on each data set will be presented as recall versus (1-precision) curve and detection equal-error rate[8] in the following sections.

5.1 Car Detection

The UIUC car data set [53] consists of learning and test sets. The learning set contains 550 positive (car) images and 500 negative (noncar) images. The test set is divided into two parts: 170 gray-scale images containing 200 side views of cars of size 100×40, and 108 gray-scale images containing 139 cars at various sizes with a ratio between the largest and smallest cars of about 2.5. Since our method is training-free, we use only one query image at a time from the 550 positive examples.

5.1.1 Single-Scale Test Set

We compute LSK of size 9×9 as descriptors, as a consequence, every pixel in Q and T yields an 81-dimensional local descriptor $\mathbf{W}_Q$ and $\mathbf{W}_T$, respectively. The smoothing parameter h for computing LSKs was set to 2.1. We end up with $\mathbf{F}_Q, \mathbf{F}_T$ by reducing dimensionality from 81 to $d = 4$, and then we obtain RM by computing the MCS measure between $\mathbf{F}_Q, \mathbf{F}_{T_i}$. The threshold τ for each test example was determined by the confidence level $\alpha = 0.99$. Fig. 13a shows the output of the proposed method on single-scale test images.

We conducted an experiment by computing RM without performing PCA in order to verify that the use of dimensionality reduction step (PCA) plays an important role in extracting only salient features and improving the performance. We also repeated these experiments by changing the query image and computing precision and recall. In Fig. 14, recall-precision curves represent a performance comparison between the proposed method and the proposed method without PCA using five different query images. We can clearly see that the performance of our system is not terribly affected by a choice of the query images, but is quite consistent. Furthermore, PCA consistently contributes to a performance improvement. The detection equal-error rates comparison is provided in Table 1 as well.

To show an overall performance of the proposed method on five different query images, we summed up TP and FP over the entire experiment, then computed recall and precision at various steps of the threshold value τ according to the confidence level α. Note that, to the best of our knowledge, there are no other training-free methods evaluated on the UIUC data set [53], and thus, comparison is largely only made with the state-of-the-art training-based methods. The proposed method, which is training-free, performs favorably against the state-of-the-art training-based methods [53], [55], [56], which use extensive training, as shown in Fig. 15.

5.1.2 Multiscale Test Set

As explained in Section 4, we construct a multiscale pyramid of the target image T: five scales with scale factors 0.4, 0.6, 0.8, 1, and 1.2. More specifically, we reduce the target image size by steps of 20 percent up to 40 percent of the original size and upscale the target image by 20 percent so that we can deal with both cases of either the size of objects in the target images being bigger or smaller than the query. The rest of the process is similar to the single-scale

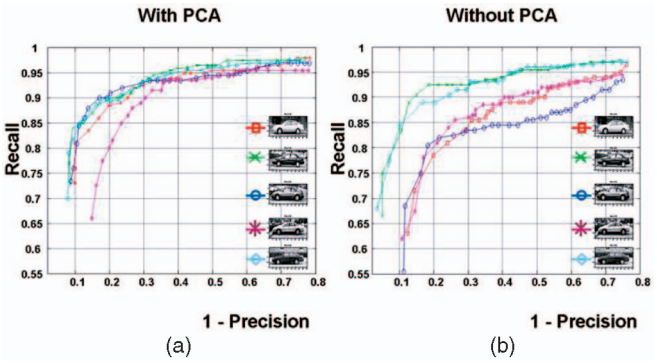


Fig. 14. (a) Recall versus 1-precision curves of the proposed method. (b) Recall versus 1-precision curves of the proposed method without PCA on the UIUC single-scale car test set [53] using five different query images.

8. Note that detection equal-error rate is a detection (recall) rate when a recall rate is the same as the precision rate.

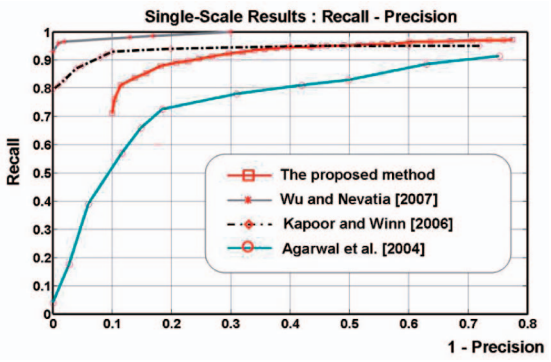


Fig. 15. Comparison of recall versus 1-precision curves between the proposed method and the state-of-the-art methods [53], [55], [56] on the UIUC single-scale test set [53].

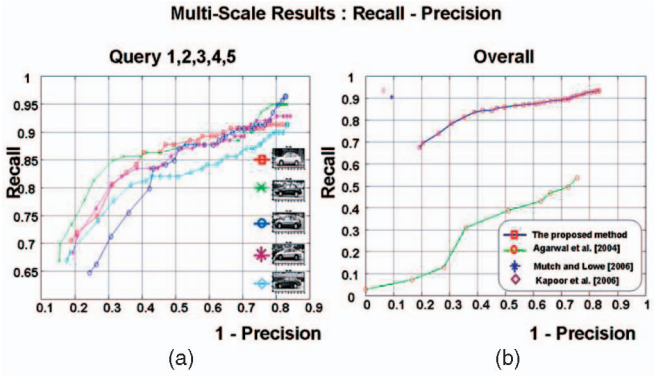


Fig. 16. (a) Recall versus 1-precision curve using five different query images. (b) Comparison of recall versus 1-precision curves between the proposed method and the state-of-the art methods [56], [57], [53] on the UIUC multiscale test set [53].

case. Fig. 16b shows examples of correct detections using τ corresponding to $\alpha = 0.99$.

The overall performance improvement of the proposed method (using five different query images) over Agarwal et al. [53] is even greater (over 30 percent) on the multiscale test set, as shown in Table 2 and Fig. 16. As for the interpretation of the performance on the UIUC car data set (both single-scale and multiscale cases), our methods show performance that is not far from the state-of-the-art training-based methods, except that it requires *no training* at all.

5.2 Face Detection

We showed the performance of the proposed method in the presence of a moderate scale variation (a ratio between the largest and smallest objects of about 2.5) in the previous section. In this section, we further evaluate our method on a more general scenario, where the scale ratio between the largest and smallest is over 10 and large rotations of objects may exist. Therefore, a test set is chosen from a subset of the MIT-CMU face data set [54]. The test set is composed of 43 gray-scale images[9] containing 149 frontal faces at various sizes and 20 gray-scale images[10] containing 30 faces with various rotations. A query face image of size 35×36 was employed as shown in Fig. 17, and images for a rotation experiment were resized so that faces are about the same size as the query face. Such parameters as the smoothing parameter (h), LSK size (P), confidence level (α) remain the same as the ones used in the UIUC car test sets. However, we increased scale steps for the multiscale pyramid up to 29 and rotation steps were set to 24 (i.e., rotate the query image by 15 degrees) to achieve an accurate rotation estimation. Figs. 17, 18, and 19 show that the proposed method is capable of detecting and localizing faces at distinct scale and rotation angle even in the presence of large variations in scale and rotation. We repeated the experiment by changing the query image. Fig. 20 shows recall versus 1-precision curves and (for the sake of completeness) corresponding receiver operating characteristic (ROC) curves with respect to two different queries. Note that, in the ROC curve, detection rate P_d and false alarm rate P_f are defined as $\frac{TP}{nP}$ $(= recall)$ and $\frac{FP}{FP+TN}$, respectively, where TN is the number of true negatives. As seen in Fig. 20, the performance of our method on this test set is consistent with the results in the UIUC car test sets. More specifically, the performance of the proposed method is little affected by the choice of similar query images and is quite stable.

5.3 General Object Detection

We have shown the performance of the proposed method on data sets composed of gray-scale images, which contain specific objects such as car and face. In this section, we have applied our method to a more difficult scenario, where general real-world images containing flowers, hearts, and human poses are considered. Furthermore, *rough hand-drawn sketches* are used as a query instead of real images. Shechtman and Irani's general object data set [7] consists of many challenging pairs of color images (60 pairs with queries such as flowers, hearts, peace symbols, face, and

Fig. 17. Detection results on the MIT-CMU multiscale test set [54]. α was set to 0.99. Hand-drawn faces on the white board were also detected using a real face query image.

9. The 43 images (from http://vasc.ri.cmu.edu/idb/html/face/index.html) are as follows: aerosmith-double.gif, blues-double.gif, original2.gif, audrey1.gif, audrey2.gif, baseball.gif, cfb.gif, cnn1714.gif, cnn2020.gif, cnn2600.gif, crimson.gif, ew-courtney-david.gif, gpripe.gif, hendrix2.gif, henry.gif, john.coltrane.gif, kaari1.gif, kaari2.gif, kaari-stef.gif, knex0.gif, lacrosse.gif, married.gif, police.gif, sarah4.gif, sarah_live_2.gif, tammy.gif, tori-crucify.gif, tori-entweekly.gif, tp.gif, voyager2.gif, class57.gif, trek-trio.gif, albert.gif, madaboutyou.gif, frisbee.gif, me.gif, speed.gif, ysato.gif, wxm.gif, torrance.gif, mona-lisa.gif, karen-and-rob.gif, and Germany.gif.

10. The 20 images (from http://vasc.ri.cmu.edu/idb/html/face/index.html) are as follows: 3.gif, 217.gif, 221.gif, af2206b.gif, am4945a.gif, am5528a.gif, am6227a.gif, bm5205a.gif, bm6290a.gif, boerli01. gif, cast1.gif, dole2.gif, jprc.gif, pict_6.gif, pict_28.gif, sbCelSte.gif, siggi.gif, tf5189a.gif, tf5581a.gif, and tm6109a.gif.

Fig. 18. Detection results on the MIT-CMU multiscale test set [54]. α was set to 0.99. Among 57 faces present, we detected 54 faces at a correct location with four false alarms.

Fig. 19. Detection results on the MIT-CMU multirotation test set [54]. α was set to 0.99.

human poses; see Fig. 5). In order to justify the usefulness of the MCS measure for this data set and to further verify the advantage of the CCS defined in Section 4.3 over the MCS measure, we begin with evaluating the proposed method on the luminance channel only. In Fig. 21, some examples of RM are shown. Figs. 22 and 23 show that the proposed method is able to detect and localize reliably.

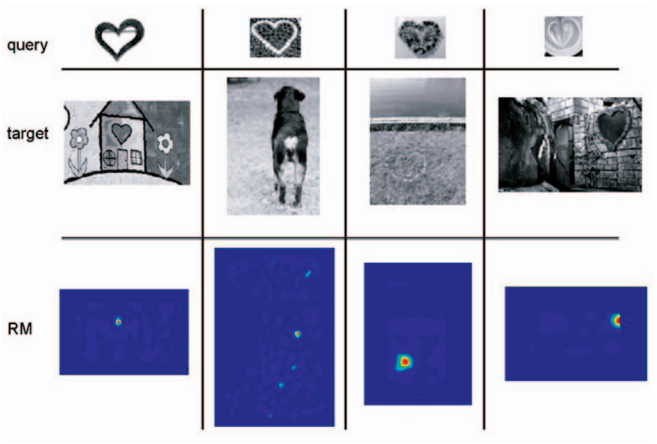


Fig. 21. Some examples of detection results with RMs in Shechtman and Irani's object test set [7]. RMs are shown in bottom row.

We further justify the use of LSKs by comparing the performance with the state-of-the-art local descriptors evaluated in [16] as was similarly done in [7]. We densely computed such local descriptors as *gradient location-orientation histogram* (GLOH) [16], *Shape Context* [59], and SIFT [13] using the implementation in [16]. By replacing LSKs with these descriptors but keeping the rest of the steps the same, we repeated the experiment on this test set. The precision-recall curve in Fig. 24 verifies that our LSKs have more discriminative power than other local descriptors. The proposed method is also evaluated on full CIE $\mathrm{L^*a^*b^*}$ data. If we look at recall rates in the range of $0 \leq (1\text{-precision}) \leq 0.1$ in Fig. 24, we can see that full CIE $\mathrm{L^*a^*b^*}$ data provide more information, and thus, CCS outperforms the MCS measure as also observed in [7]. Consistent with these results, it is worth noting that Shechtman and Irani [7] also showed that their local self-similarity descriptor clearly outperformed other state-of-the-art descriptors in their ensemble matching framework. However, the performance figures they provide are rather incomplete. Namely, they mentioned 86 percent detection rate without specifying either any precision rates or false alarm rates. Therefore, we claim that our proposed method is more general and practical than the training-free detection method in [7] .

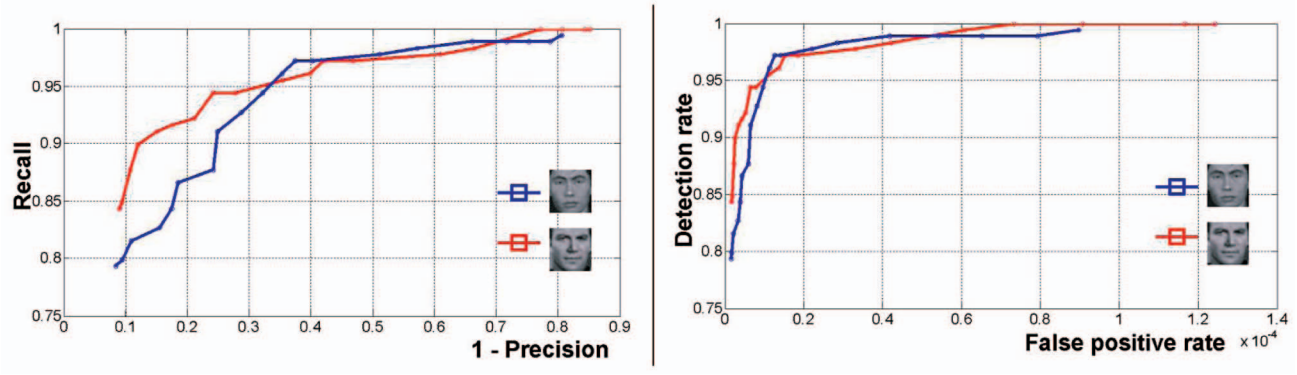


Fig. 20. Left: precision-recall curves. Right: ROC curves on the MIT-CMU test set [54] using two different query images. Note that detection rate P_d and false alarm rate P_f are defined as $\frac{TP}{nP}$ $(= recall)$ and $\frac{FP}{FP+TN}$, respectively, where TN is the number of true negatives.

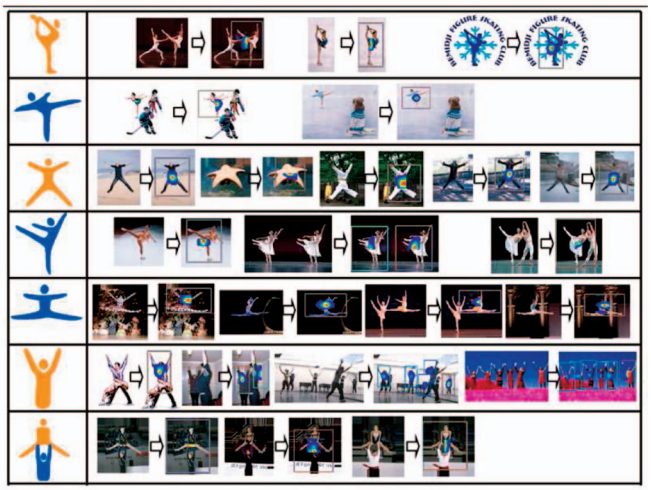


Fig. 22. Left: Hand-drawn sketch query (human poses). Right: Targets and examples of correction detections/localizations in Shechtman and Irani's object test set [7]. α was set to 0.98.

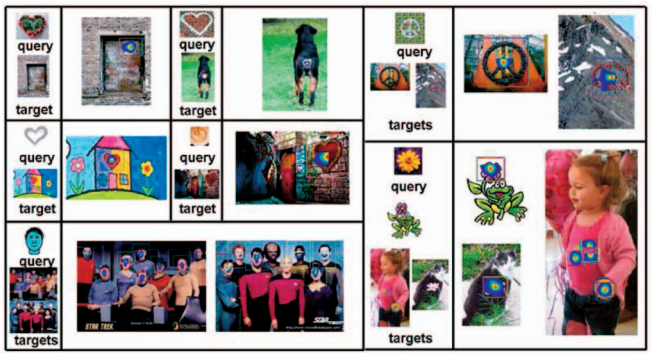


Fig. 23. Query: Hearts, hand-drawn face, peace symbol, and flower. Some targets and examples of correction detections/localizations in Shechtman and Irani's object test set [7] are shown. Some false positives appeared in a girl's T-shirt and candle. α was set to 0.98.

6 Conclusion and Future Directions

In this paper, we have proposed a novel and powerful training-free nonparametric object detection framework by employing LSKs, which well capture underlying data structure, and by using the MCS measure. We have justified the approach using a naive Bayes argument, which leads to the use of the MCS measure. The proposed method can automatically detect in the target image the presence, the number, as well as location of similar objects to the given query image. To deal with more general scenarios, accounting for large variations in scale and rotation, we further proposed multiscale and multirotation approach.

The CCS has proven to be more effective than MCS when vector-valued images are available though this requires further study. Challenging sets of real-world object experiments have demonstrated that the proposed approach achieves a high detection accuracy of objects of interest even in completely different context and under different imaging conditions. Unlike other state-of-the-art learning-based detection methods, the proposed framework operates using a *single* example of an image of interest to find similar matches, does not require any prior knowledge (learning) about objects being sought, and does not require any segmentation or preprocessing step of the target image. The proposed framework is general enough as to be extendable to 3D for such applications as action recognition, suspicious behavior detection, etc., using analogous 3D LSKs [35], [60]. Since the proposed method is designed with detection accuracy as a high priority, extension of the method to a large-scale data set requires a significant improvement of the computational complexity of the proposed method. Toward this end, we could benefit from an efficient searching method (coarse-to-fine search) and/or a fast nearest neighbor search method (e.g., vantage point tree [61]). Recently, large database-driven approaches [62], [63] have shown potential for nonparametric detection. For instance, [63] showed that with a database of 80 million images, even simple matching based on the SSDs can provide semantically meaningful classification performance for 32×32 images. Thus, we could use a fast indexing techniques such as spatial pyramid matching (SPM) [64] or GIST matching [65] in order to reduce the search space and rapidly and accurately limit the number of candidate images. Subsequently, we can apply the proposed method for more accurate detection. Additionally, for the proposed method to be feasible for scalable image retrieval, we may adopt the idea of encoding the features as proposed in [66], [67]. Interestingly, the detection framework proposed in our paper can also be useful for solving the bottom-up saliency detection problem [32] by computing a self-resemblance map between a center feature set (as a query) and

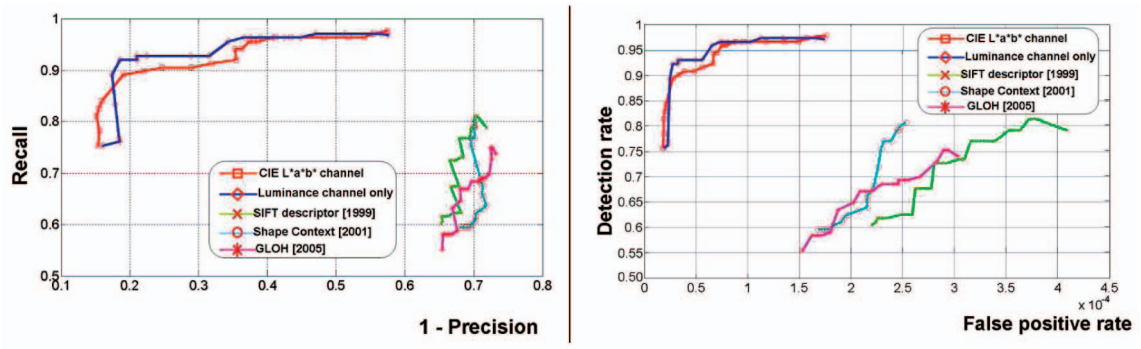


Fig. 24. Left: Comparison of recall versus 1-Precision curves between luminance channel only and CIE $\mathrm{L^*a^*b^*}$ channel on Shechtman and Irani's test set [7]. It is clearly shown that such descriptors as SIFT [13], GLOH [16], Shape Context [59] turn out to be inferior to LSKs in terms of discriminative power. Right: Comparison of ROC curves. Note that detection rate P_d and false alarm rate P_f are defined as $\frac{TP}{nP}$ $(= recall)$ and $\frac{FP}{FP+TN}$, respectively, where TN is the number of true negatives.

TABLE 1
Single-Scale Result: Detection Equal-Error Rates on the UIUC Single-Scale Car Test Set [53]

The proposed method w/o PCA	Query 1	Query 2	Query 3	Query 4	Query 5	Agarwal et al. [53] (1)	Wu and Nevatia [55]	Mutch and Lowe [57]
Detection rates	79.29 %	**88.12** %	81.11 %	80.41 %	87.11 %	77.08 %	97.5 %	99.94 %
The proposed method	Query 1	Query 2	Query 3	Query 4	Query 5	Agarwal et al. [53] (2)	Kapoor and Winn [56]	Lampert et al. [58]
Detection rates	85.26 %	**87.27** %	87.13 %	80.57 %	86.73 %	76.72 %	94.0 %	98.5 %

TABLE 2
Multiscale Result: Detection Equal-Error Rates on the UIUC Multiscale Car Test Set [53]

The proposed method	Query 1	Query 2	Query 3	Query 4	Query 5	Agarwal et al. [53]	Mutch and Lowe [57]	Kapoor and Winn [56]	Lampert et al. [58]
Detection rates	75.47 %	**77.66** %	70.21 %	75.00 %	74.22 %	43.77 $\sim$ 44.00 %	90.6 %	93.5 %	98.6 %

surrounding feature sets (as a target). We also expect to be able to apply the proposed method to other challenging medical/diagnostic problems such as change detection in medical imaging applications.

APPENDIX

The Lagrangian objective function to the minimization problem in (25) is

$$f(\lambda_Q, \lambda_T, \mathbf{u}_Q, \mathbf{u}_{T_i}) = \mathbf{u}_Q^T \mathbf{F}_Q^T \mathbf{F}_{T_i} \mathbf{u}_{T_i} - \frac{\lambda_Q}{2}\left(\mathbf{u}_Q^T \mathbf{F}_Q^T \mathbf{F}_Q \mathbf{u}_Q - 1\right) - \frac{\lambda_{T_i}}{2}\left(\mathbf{u}_{T_i}^T \mathbf{F}_{T_i}^T \mathbf{F}_{T_i} \mathbf{u}_{T_i} - 1\right). \tag{28}$$

Taking derivatives with respect to $\mathbf{u}_Q$ and $\mathbf{u}_{T_i}$, we obtain

$$\frac{\partial f}{\partial \mathbf{u}_Q} = \mathbf{F}_Q^T \mathbf{F}_{T_i} \mathbf{u}_{T_i} - \lambda_Q\left(\mathbf{F}_Q^T \mathbf{F}_Q \mathbf{u}_Q\right) = 0, \tag{29}$$

$$\frac{\partial f}{\partial \mathbf{u}_{T_i}} = \mathbf{F}_{T_i}^T \mathbf{F}_Q \mathbf{u}_Q - \lambda_{T_i}\left(\mathbf{F}_{T_i}^T \mathbf{F}_{T_i} \mathbf{u}_{T_i}\right) = 0. \tag{30}$$

We premultiply $\mathbf{u}_{T_i}^T$ to (30) and also premultiply $\mathbf{u}_Q^T$ to (29). By subtracting these two equations, we have

$$\mathbf{u}_Q^T \mathbf{F}_Q^T \mathbf{F}_{T_i} \mathbf{u}_{T_i} - \lambda_Q\left(\mathbf{u}_Q^T \mathbf{F}_Q^T \mathbf{F}_Q \mathbf{u}_Q\right) - \mathbf{u}_{T_i}^T \mathbf{F}_{T_i}^T \mathbf{F}_Q \mathbf{u}_Q - \lambda_{T_i}\left(\mathbf{u}_{T_i}^T \mathbf{F}_{T_i}^T \mathbf{F}_{T_i} \mathbf{u}_{T_i}\right) = 0, \tag{31}$$

where $(\mathbf{u}_Q^T \mathbf{F}_Q^T \mathbf{F}_{T_i} \mathbf{u}_{T_i})^T = \mathbf{u}_{T_i}^T \mathbf{F}_{T_i}^T \mathbf{F}_Q \mathbf{u}_Q$ is a scalar.

Enforcing the constraints

$$\left(\mathbf{u}_Q^T \mathbf{F}_Q^T \mathbf{F}_Q \mathbf{u}_Q\right)^T = \left(\mathbf{u}_{T_i}^T \mathbf{F}_{T_i}^T \mathbf{F}_{T_i} \mathbf{u}_{T_i}\right)^T = 1,$$

we are led to the conclusion that $\lambda_Q = \lambda_{T_i}$. We define $\rho = \lambda_Q = \lambda_{T_i}$. Assuming that $\mathbf{F}_{T_i}^T \mathbf{F}_{T_i}$ is invertible from (30),

$$\mathbf{u}_{T_i} = \frac{\left(\mathbf{F}_{T_i}^T \mathbf{F}_{T_i}\right)^{-1} \mathbf{F}_{T_i} \mathbf{F}_Q \mathbf{u}_Q}{\rho}, \tag{32}$$

and so, plugging into (29), we have

$$\frac{\left(\mathbf{F}_Q^T \mathbf{F}_{T_i}\right)\left(\mathbf{F}_{T_i}^T \mathbf{F}_{T_i}\right)^{-1}\left(\mathbf{F}_{T_i} \mathbf{F}_Q\right)\mathbf{u}_Q}{\rho} = \rho\left(\mathbf{F}_Q^T \mathbf{F}_Q\right)\mathbf{u}_Q. \tag{33}$$

Assuming that $\mathbf{F}_Q^T \mathbf{F}_Q$ is also invertible, we are left with

$$\left(\mathbf{F}_Q^T \mathbf{F}_Q\right)^{-1}\left(\mathbf{F}_Q^T \mathbf{F}_{T_i}\right)\left(\mathbf{F}_{T_i}^T \mathbf{F}_{T_i}\right)^{-1}\left(\mathbf{F}_{T_i}^T \mathbf{F}_Q\right)\mathbf{u}_Q = \rho^2 \mathbf{u}_Q. \tag{34}$$

Similarly, we have

$$\left(\mathbf{F}_{T_i}^T \mathbf{F}_{T_i}\right)^{-1}\left(\mathbf{F}_{T_i}^T \mathbf{F}_Q\right)\left(\mathbf{F}_Q^T \mathbf{F}_Q\right)^{-1}\left(\mathbf{F}_Q^T \mathbf{F}_{T_i}\right)\mathbf{u}_{T_i} = \rho^2 \mathbf{u}_{T_i}. \tag{35}$$

ACKNOWLEDGMENTS

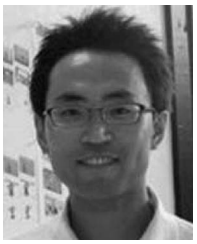
This work was supported in part by US Air Force Office of Scientific Research Grant FA 9550-07-01-0365.

REFERENCES

[1] J. Ponce, M. Hebert, C. Schmid, and A. Zisserman, *Toward Category-Level Object Recognition*. Springer, 2007.
[2] R. Fergus, L. Fei-Fei, P. Perona, and A. Zisserman, "Learning Object Categories from Google's Image Search," *Proc. 10th IEEE Int'l Conf. Computer Vision*, vol. 2, pp. 1816-1823, Oct. 2005.
[3] L. Fei-Fei, R. Fergus, and P. Perona, "Learning Generative Visual Models from Few Training Examples: An Incremental Bayesian Approach Tested on 101 Object Categories," *Proc. IEEE Conf. Computer Vision and Pattern Recognition Workshop Generative-Model Based Vision*, 2004.
[4] M. Everingham, A. Zisserman, C.K.I. Williams, and L. van Gool, "The PASCAL Visual Object Classes Challenge 2006 (VOC2006) Results," http://www.pascal-network.org/challenges/VOC/voc2006/results.pdf, 2009.
[5] P. Viola and M. Jones, "Robust Real-Time Object Detection," *Int'l J. Computer Vision*, vol. 57, no. 2, pp. 137-154, 2004.
[6] H. Masnadi-Shirazi and N. Vasconcelos, "High Detection-Rate Cascades for Real-Time Object Detection," *Proc. IEEE Int'l Conf. Computer Vision*, pp. 1-6, 2007.
[7] E. Shechtman and M. Irani, "Matching Local Self-Similarities across Images and Videos," *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, pp. 1-8, June 2007.
[8] E. Shechtman and M. Irani, "Space-Time Behavior-Based Correlation—or—How to Tell If Two Underlying Motion Fields Are Similar without Computing Them?" *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 29, no. 11, pp. 2045-2056, Nov. 2007.
[9] C. Yeo, P. Ahammad, K. Ramchandran, and S.S. Sastry, "High-Speed Action Recognition and Localization in Compressed Domain Videos," *IEEE Trans. Circuits and Systems for Video Technology*, vol. 18, no. 8, pp. 1006-1015, Aug. 2008.
[10] T. Kim, S. Wong, and R. Cipolla, "Tensor Canonical Correlation Analysis for Action Classification," *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, pp. 1-8, 2007.

[11] P. Phillips, P. Flynn, T. Scruggs, K. Bowyer, C. Jin, K. Hoffman, J. Marques, M. Jaesik, and W. Worek, "Overview of the Face Recognition Grand Challenge," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* pp. 947-954, 2005.
[12] O. Boiman, E. Shechtman, and M. Irani, "In Defense of Nearest-Neighbor Based Image Classification," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* pp. 1-8, June 2008.
[13] D. Lowe, "Distinctive Image Features from Scale-Invariant Keypoints," *Int'l J. Computer Vision,* vol. 20, pp. 91-110, 2004.
[14] H. Zhang, A. Berg, M. Maire, and J. Malik, "SVM-KNN: Discriminative Nearest Neighbor Classification for Visual Category Recognition," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* vol. 2, pp. 2126-2136, 2006.
[15] K. Grauman and T. Darrell, "The Pyramid Match Kernel: Efficient Learning with Sets of Features," *J. Machine Learning Research,* vol. 8, pp. 725-760, 2007.
[16] K. Mikolajczyk and C. Schmid, "A Performance Evaluation of Local Descriptors," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 27, no. 10, pp. 1615-1630, Oct. 2005.
[17] H. Takeda, S. Farsiu, and P. Milanfar, "Kernel Regression for Image Processing and Reconstruction," *IEEE Trans. Image Processing,* vol. 16, no. 2, pp. 349-366, Feb. 2007.
[18] H. Takeda, S. Farsiu, and P. Milanfar, "Deblurring Using Regularized Locally-Adaptive Kernel Regression," *IEEE Trans. Image Processing,* vol. 17, no. 4, pp. 550-563, Apr. 2008.
[19] Y. Fu, S. Yan, and T.S. Huang, "Correlation Metric for Generalized Feature Extraction," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 30, no. 12, pp. 2229-2235, Dec. 2008.
[20] Y. Fu and T.S. Huang, "Image Classification Using Correlation Tensor Analysis," *IEEE Trans. Image Processing,* vol. 17, no. 2, pp. 226-234, Feb. 2008.
[21] C. Liu, "The Bayes Decision Rule Induced Similarity Measures," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 29, no. 6, pp. 1086-1090, June 2007.
[22] C. Liu, "Clarification of Assumptions in the Relationship between the Bayes Decision Rule and the Whitened Cosine Similarity Measure," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 30, no. 6, pp. 1116-1117, June 2008.
[23] D. Lin, S. Yan, and X. Tang, "Comparative Study: Face Recognition on Unspecific Persons Using Linear Subspace Methods," *Proc. IEEE Int'l Conf. Image Processing,* vol. 3, pp. 764-767, 2005.
[24] Y. Ma, S. Lao, E. Takikawa, and M. Kawade, "Discriminant Analysis in Correlation Similarity Measure Space," *Proc. Int'l Conf. Machine Learning,* vol. 227, pp. 577-584, 2007.
[25] J.W. Schneider and P. Borlund, "Matrix Comparison, Part 1: Motivation and Important Issues for Measuring the Resemblance between Proximity Measures or Ordination Results," *J. Am. Soc. Information Science and Technology,* vol. 58, no. 11, pp. 1586-1595, 2007.
[26] P. Ahlgren, B. Jarneving, and R. Rousseau, "Requirements for a Cocitation Similarity Measure, with Special Reference to Pearson's Correlation Coefficient," *J. Am. Soc. Information Science and Technology,* vol. 54, no. 6, pp. 550-560, 2003.
[27] J. Rodgers and W. Nicewander, "Thirteen Ways to Look at the Correlation Coefficient," *The Am. Statistician,* vol. 42, no. 1, pp. 59-66, 1988.
[28] F. Devernay, "A Non-Maxima Suppression Method for Edge Detection with Sub-Pixel Accuracy," Technical Report RR-2724, Institut National de Recherche en Informatique et en Automatique, 1995.
[29] C. Kervrann and J. Bourlanger, "Optimal Spatial Adaptation for Patch-Based Image Denoising," *IEEE Trans. Image Processing,* vol. 15, no. 10, pp. 2866-2878, Oct. 2006.
[30] A. Buades, B. Coll, and J.M. Morel, "Nonlocal Image and Movie Denoising," *Int'l J. Computer Vision,* vol. 76, no. 2, pp. 123-139, 2008.
[31] M. Vasconcelos and N. Vasconcelos, "Natural Image Statistics and Low-Complexity Feature Selection," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 31, no. 2, pp. 228-244, Feb. 2009.
[32] H.J. Seo and P. Milanfar, "Static and Space-Time Visual Saliency Detection by Self-Resemblance," *J. Vision,* vol. 9, no. 12, pp. 1-27, 2009, http://journalofvision.org/9/12/15.doi:10.1167/9.12.15.
[33] M. Tatsuoka, *Multivariate Analysis.* Macmillan, 1988.
[34] O. Boiman and M. Irani, "Detecting Irregularities in Images and in Video," *Int'l J. Computer Vision,* vol. 74, pp. 17-31, Aug. 2007.
[35] H. Takeda, P. van Beek, and P. Milanfar, "Spatio-Temporal Video Interpolation and Denoising Using Motion-Assisted Steering Kernel (MASK) Regression," *Proc. IEEE Int'l Conf. Image Processing,* pp. 637-640, 2008.
[36] S. Han and N. Vasconcelos, "Complex Discriminant Features for Object Classification," *Proc. Int'l Conf. Image Processing,* pp. 1700-1703, 2008.
[37] T. Tuytelaar and C. Schmid, "Vector Quantizing Feature Space with a Regular Lattice," *Proc. IEEE Int'l Conf. Computer Vision,* pp. 1-8, Oct. 2007.
[38] F. Jurie and B. Triggs, "Creating Efficient Codebooks for Visual Recognition," *Proc. IEEE Int'l Conf. Computer Vision,* vol. 1, pp. 604-610, 2005.
[39] Y. Ke and R. Sukthankar, "PCA-SIFT: A More Distinctive Representation for Local Image Descriptors," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* pp. 506-513, 2004.
[40] S. Ali and M. Shah, "Human Action Recognition in Videos Using Kinematic Features and Multiple Instance Learning," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 32, no. 2, pp. 288-303, Feb. 2010.
[41] S. Roweis and L. Saul, "Nonlinear Dimensionality Reduction by Locally Linear Embedding," *Science,* vol. 290, no. 5500, pp. 2323-2326, 2000.
[42] X. He, S. Yan, Y. Hu, P. Niyogi, and H.-J. Zhang, "Face Recognition Using Laplacianfaces," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 27, no. 3, pp. 328-340, Mar. 2005.
[43] Y. Escoufier, "Operator Related to a Data Matrix: A Survey," *Proc. 17th Symp. Computational Statistics,* pp. 285-297, 2006.
[44] R.J. Rummel, *Applied Factor Analysis.* Northwestern Univ. Press, 1970.
[45] P. Horst, *Matrix Algebra for Social Scientists.* Holt, Rinehart, and Winston, 1963.
[46] T. Calinski, M. Krzysko, and W. Wolynski, "A Comparison of Some Tests for Determining the Number of Nonzero Canonical Correlations," *Comm. Statistics, Simulation and Computation,* vol. 35, pp. 727-749, 2006.
[47] R. Duda, P. Hart, and D. Stark, *Pattern Classification,* second ed. John Wiley and Sons, Inc., 2000.
[48] B. Silverman, *Density Estimation for Statistics and Data Analysis,* Monographs on Statistics and Applied Probability 26. Chapman & Hall, 1986.
[49] S. Kay, *Fundamentals of Statistical Signal Processing, Volume 1: Estimation Theory.* Prentice Hall, 1993.
[50] P. Vincent and Y. Bengio, "Manifold Parzen Windows," *Advances in Neural Information Processing Systems,* vol. 15, pp. 825-832, MIT Press, 2003.
[51] T. Brox, B. Rosenhahn, and H.-P.S.D. Cremers, "Nonparametric Density Estimation with Adaptive Anisotropic Kernels for Human Motion Tracking," *Proc. Second Workshop Human Motion,* pp. 152-165, 2007.
[52] Y. Bengio, H. Larochelle, and P. Vincent, "Non-Local Manifold Parzen Windows," *Advances in Neural Information Processing Systems,* vol. 18, pp. 115-122, MIT Press, 2005.
[53] S. Agarwal, A. Awan, and D. Roth, "Learning to Detect Objects in Images via a Sparse, Part-Based Representation," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 26, no. 11, pp. 1475-1490, Nov. 2004.
[54] H. Rowley, S. Baluja, and T. Kanade, "Neural Network-Based Face Detection," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 20, no. 1, pp. 22-38, Jan. 1998.
[55] B. Wu and R. Nevatia, "Simultaneous Object Detection and Segmentation by Boosting Local Shape Feature Based Classifier," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* pp. 1-8, June 2007.
[56] A. Kappor and J. Winn, "Located Hidden Random Fields: Learning Discriminative Parts for Object Detection," *Proc. European Conf. Computer Vision,* vol. 3954, pp. 302-315, May 2006.
[57] J. Mutch and D.G. Lowe, "Multiclass Object Recognition with Sparse, Localized Features," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* pp. 11-18, 2006.
[58] C.H. Lampert, M.B. Blaschko, and T. Hofmann, "Beyond Sliding Windows: Object Localization by Efficient Subwindow Search," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* pp. 1-8, 2008.
[59] S. Belongie, J. Malik, and J. Puzicha, "Shape Matching and Object Recognition Using Shape Contexts," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 24, no. 4, pp. 509-522, Apr. 2002.

[60] H. Takeda, P. Milanfar, M. Protter, and M. Elad, "Super-Resolution without Explicit Subpixel Motion Estimation," *IEEE Trans. Image Processing,* vol. 18, no. 9, pp. 1958-1975, Sept. 2009.
[61] N. Kumar, L. Zhang, and S. Nayar, "What Is a Good Nearest Neighbors Algorithm for Finding Similar Patches in Images," *Proc. European Conf. Computer Vision,* pp. 364-378, 2008.
[62] B.C. Russell, A. Torralba, K.P. Murphy, and W.T. Freeman, "Labelme: A Database and Web-Based Tool for Image Annotation," *Int'l J. Computer Vision,* vol. 70, no. 13, pp. 157-173, 2008.
[63] A. Torralba, R. Fergus, and W.T. Freeman, "80 Million Tiny Images: A Large Data Set for Non-Parametric Object and Scene Recognition," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 30, no. 11, pp. 1958-1970, Nov. 2008.
[64] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* pp. 2169-2178, 2006.
[65] A. Oliva and A. Torralba, "Modeling the Shape of the Scene: A Holistic Representation of the Spatial Envelope," *Int'l J. Computer Vision,* vol. 42, no. 3, pp. 145-175, 2001.
[66] D.M. Chen, S.S. Tsai, V. Chandrasekhar, G. Takacs, J. Singh, and B. Girod, "Tree Histogram Coding for Mobile Image Matching," *Proc. IEEE Data Compression Conf.,* Mar. 2009.
[67] V. Chandrasekhar, G. Takacs, D.M. Chen, S.S. Tsai, J.P. Singh, and B. Girod, "Transform Coding of Image Feature Descriptors," *Proc. Conf. Visual Comm. and Image Processing,* Jan. 2009.



Hae Jong Seo received the BS and MS degrees in electrical engineering from Sungkyunkwan University, Seoul, Korea, in 2005 and 2006, respectively. He is currently working toward the PhD degree in electrical engineering at the University of California, Santa Cruz. His research interests are in the domain of image processing (denoising, interpolation, deblurring, and super-resolution) and computer vision (visual object recognition). He is a student member of the IEEE.

Peyman Milanfar received the BS degree in electrical engineering and mathematics from the University of California, Berkeley, in 1988, and the MS, EE, and PhD degrees in electrical engineering from the Massachusetts Institute of Technology, Cambridge, in 1990, 1992, and 1993, respectively. Until 1999, he was a senior research engineer at SRI International, Menlo Park, California. He is currently a professor of electrical engineering at the University of California, Santa Cruz. He was a consulting assistant professor of computer science at Stanford University, California, from 1998 to 2000, where he was also a visiting associate professor in 2002. His technical interests include statistical signal and image processing and inverse problems. He won a US National Science Foundation CAREER award. He is an associate editor for the *IEEE Transaction on Image Processing* and was an associate editor for the *IEEE Signal Processing Letters* from 1998 to 2001. He is a member of the Signal Processing Society's Image, Video, and Multidimensional Signal Processing (IVMSP) Technical Committee. He is a fellow of the IEEE.