NONPARAMETRIC DETECTION AND RECOGNITION OF VISUAL OBJECTS FROM A SINGLE EXAMPLE

Hae Jong Seo and Peyman Milanfar

Electrical Engineering Department
University of California at Santa Cruz
1156 High Street, Santa Cruz, CA, 95064

ABSTRACT

We present a generic detection/localization algorithm capable of searching for a 2- or 3-D visual object of interest without training. The proposed method operates using a single example (query) of an object of interest to find similar matches; does not require prior knowledge (learning) about objects being sought; and does not require any pre-processing step or segmentation of a target image/video. 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. State of the art performance is demonstrated on several challenging datasets, indicating successful detection of visual objects in diverse contexts and under varying imaging conditions.

Index Terms— Visual object detection and recognition, image representation, correlation and regression analysis

1. INTRODUCTION

The central problem in computer vision is "visual object recognition": namely, the ability to automatically categorize an image or video of interest (a query) as either coming from a known category (classification), or as being significantly similar to an already seen visual target (detection/localization). In particular, the 2-D object recognition problem (including face, pedestrian, and vehicle recognition) and human action recognition problem have attracted much attention recently due to the increasing demand for developing real-world surveillance systems. Visual object recognition is considered to be a very difficult problem because objects can typically appear in completely different context and under different imaging conditions. Examples of such differences can be wide ranging, but include differing view points, occlusion, lighting, and scale changes as shown in Fig. 1.

For the last few decades, learning-based methods for recognizing visual objects have made impressive progress. Typically, learning-based approaches involve generative or discriminative models for each category based on many training examples. In other words, these methods are mostly parametric, relying on visual object models, such as constellation [1], template matching [2], bags of words [3], or shape models [4], etc. For specific object classes, in particular faces, pedestrians and cars, detectors based on the combination of low-level features combined with modern machine learning techniques have been shown effective. However, in order to achieve good accuracy, these systems require a large number of manually labeled training data, typically hundreds or thousands of example images for each class to be learned. Furthermore, 2-D object recognition methods were not directly applicable to 3-D action recognition task, and thus, completely separate approaches have been proposed for action recognition tasks. Indeed, even in terms of evaluation of performance, different criteria and methodologies have been employed in 2-D and 3-D.

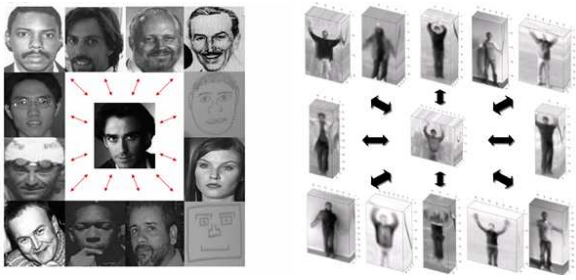

Fig. 1. (a) A face and some possibly similar images (b) A waving action and some possibly similar actions

Recently, the recognition task with only one query (training-free) has received increasing attention [5, 6, 7, 8] 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 [1, 5]. 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. As a successful example of image-to-image matching, Boiman et al. [9] showed that a rather simple nearest-neighbor (NN) based image classifier in the space of the local image descriptors is efficient and even outperforms the leading learning-based image classifiers such as SVM-KNN [10], pyramid match kernel (PMK) [11]. Ac-

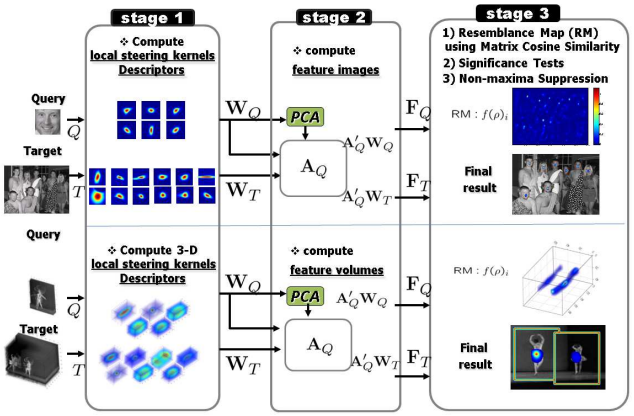

Fig. 2. System overview [14]. Top: Object detection framework, Bottom: action detection framework. (There are broadly three stages.)

tion recognition methods such as those in [6, 12, 7] which aim at recognizing actions based solely on one query support these ideas as well.

This paper addresses the generic detection/localization problem of searching for an object of interest (for instance a picture of a face or a ballet turning action) within a "target" with only a *single* "query". Denoting the target (T), and the query (Q), we compute a dense set of local descriptors from each. These densely computed descriptors are highly informative, but taken together tend to be over-complete (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 Q. Hence, we divide the target 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 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 i^{th} patch of T to Q to look for matches. We employ "Maxtrix Cosine Similarity" to measure the similarity between feature sets. 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 non-maxima suppression [13]. Fig. 2 shows the overview of the proposed system.

The detail of this paper can be found in two journal papers[1] [14, 15] which were accepted for and submitted to IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) respectively. In the next section, 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"). In Section 3, we demonstrate the performance of the system with some experimental results, and we conclude this paper in Section 4.

[1] http://users.soe.ucsc.edu/~milanfar/research/computer-vision.html

2. VISUAL OBJECT DETECTION IN 2-D AND 3-D

2.1. Local Descriptors

2.1.1. Local Steering Kernel (2-D LSK)

The key idea behind local steering kernels is to robustly obtain the local structure of images by analyzing the radiometric (pixel value) differences based on estimated gradients, and use this structure information to determine the shape and size of a canonical kernel. The local steering kernel is modeled as

$$K(\mathbf{x}_l - \mathbf{x}_i) = \frac{\sqrt{\det(\mathbf{C}_l)}}{h^2} \exp\left\{\frac{(\mathbf{x}_l - \mathbf{x}_i)^T \mathbf{C}_l (\mathbf{x}_l - \mathbf{x}_i)}{-2h^2}\right\}, \quad (1)$$

where $l \in \{1, \cdots, P\}$, P is the number of pixels in a local window; h is a global smoothing parameter. The matrix $\mathbf{C}_l \in \mathbb{R}^{2\times 2}$ is a covariance matrix estimated from a collection of spatial gradient vectors within the local analysis window around a position $\mathbf{x}_l = [x_1, x_2]_l^T$. More specifically, the covariance matrix $\mathbf{C}_l$ can be first naively estimated as $\mathbf{J}_l^T \mathbf{J}_l$ with

$$\mathbf{J}_l = \begin{bmatrix} z_{x_1}(\mathbf{x}_1), & z_{x_2}(\mathbf{x}_1) \\ \vdots & \vdots \\ z_{x_1}(\mathbf{x}_P), & z_{x_2}(\mathbf{x}_P)) \end{bmatrix},$$

where $z_{x_1}(\cdot)$ and $z_{x_2}(\cdot)$ are the first derivatives along x_1-, and x_2- axes. For the sake of robustness, we compute a more stable estimate of $\mathbf{C}_l$ by invoking the singular value decomposition (SVD) of $\mathbf{J}_l$ with regularization as [16, 14]

$$\mathbf{C}_l = \gamma \sum_{q=1}^{2} a_q^2 \mathbf{v}_q \mathbf{v}_q^T \in \mathbb{R}^{(2\times 2)}, \quad (2)$$

with

$$a_1 = \frac{s_1 + \lambda'}{s_2 + \lambda'} \quad a_2 = \frac{s_2 + \lambda'}{s_1 + \lambda'} \quad \gamma = \left(\frac{s_1 s_2 + \lambda''}{P}\right)^{\alpha}, \quad (3)$$

where λ' and λ'' are parameters[2]that dampen the noise effect and restrict γ and the denominators of a_q's from being zero. The singular values (s_1, s_2) and the singular vectors $(\mathbf{v}_1, \mathbf{v}_2)$ are given by the compact SVD of $\mathbf{J}_l = \mathbf{U}_l \mathbf{S}_l \mathbf{V}_l^T = \mathbf{U}_l \mathrm{diag}[s_1, s_2]_l [\mathbf{v}_1, \mathbf{v}_2,]_l^T$. Fig. 3 illustrates that how covariance matrices and LSK values are computed in an edge region.

2.1.2. Space-Time Local Steering Kernel (3-D LSK)

Now, we introduce the time axis to the data model so that $\mathbf{x}_l = [x_1, x_2, t]_l^T$: x_1 and x_2 are the spatial coordinates, t is

[2] These parameters are used for regularization purpose. They are set and fixed for all experiment.

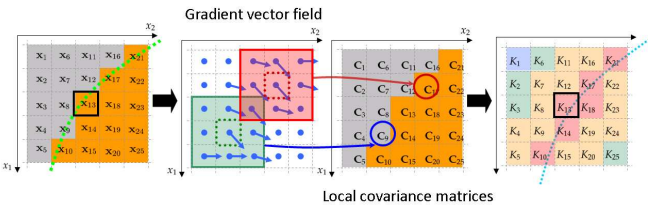

Fig. 3. Graphical description of how LSK values centered at pixel of interest $\mathbf{x}_{13}$ are computed in an edge region. Note that each pixel location has its own $\mathbf{C}$ computed from gradient vector field within a local window illustrated as green and red one.

the temporal coordinate. In this setup, the covariance matrix $\mathbf{C}_l$ can be naively estimated as $\mathbf{J}_l^T\mathbf{J}_l$ with

$$\mathbf{J}_l = \begin{bmatrix} z_{x_1}(\mathbf{x}_1), & z_{x_2}(\mathbf{x}_1), & z_t(\mathbf{x}_1) \\ \vdots & \vdots & \vdots \\ z_{x_1}(\mathbf{x}_P), & z_{x_2}(\mathbf{x}_P), & z_t(\mathbf{x}_P) \end{bmatrix},$$

where $z_{x_1}(\cdot), z_{x_2}(\cdot)$, and $z_t(\cdot)$ are the first derivatives along x_1-, x_2-, and $t-$ axes, and P is the total number of samples in a *space-time* local analysis window (or cube) around a sample position at $\mathbf{x}_i$. As similarly done in 2-D case, $\mathbf{C}_l$ is estimated by invoking the singular value decomposition (SVD) of $\mathbf{J}_l$ with regularization as [17, 15]:

$$\mathbf{C}_l = \gamma \sum_{q=1}^{3} a_q^2 \mathbf{v}_q \mathbf{v}_q^T \in \mathbb{R}^{(3\times 3)}, \tag{4}$$

with

$$a_1 = \frac{s_1+\lambda'}{\sqrt{s_2 s_3}+\lambda'}, a_2 = \frac{s_2+\lambda'}{\sqrt{s_1 s_3}+\lambda'},$$
$$a_3 = \frac{s_3+\lambda'}{\sqrt{s_1 s_2}+\lambda'}, \gamma = \left(\frac{s_1 s_2 s_3+\lambda''}{P}\right)^{\alpha}, \tag{5}$$

where λ' and λ'' are parameters that dampen the noise effect and restrict γ and the denominators of a_q's from being zero. As mentioned earlier, the singular values (s_1, s_2, and s_3) and the singular vectors ($\mathbf{v}_1, \mathbf{v}_2$, and $\mathbf{v}_3$) are given by the compact SVD of $\mathbf{J}_l = \mathbf{U}_l\mathbf{S}_l\mathbf{V}_l^T = \mathbf{U}_l \text{diag}[s_1, s_2, s_3]_l [\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3]_l^T$.

Then, the covariance matrix $\mathbf{C}_l$ modifies the shape and size of the local kernel in a way which robustly encodes the space-time local geometric structures present in the video

In what follows, at a position $\mathbf{x}_i$, we will essentially be using (a normalized version of) the function $K(\mathbf{x}_l - \mathbf{x}_i)$. To be more specific, the local steering kernel function $K(\mathbf{x}_l - \mathbf{x}_i)$ is calculated at every pixel location and normalized as follows

$$W_I^i = \frac{K(\mathbf{x}_l - \mathbf{x}_i)}{\sum_{l=1}^{P} K(\mathbf{x}_l - \mathbf{x}_i)}, \quad i = 1, \cdots, M, \tag{6}$$

where I can be Q or T for query or target.

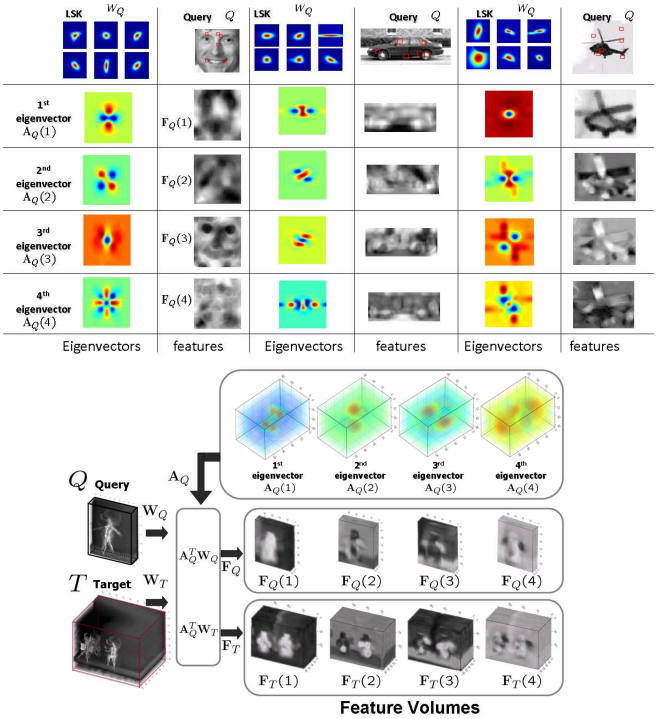

Fig. 4. Top: face, car, and helicopter examples, $\mathbf{A}_Q$ is learned from a collection of 2-D LSKs $\mathbf{W}_Q$, and Feature row vectors of $\mathbf{F}_Q$ are computed from $\mathbf{W}_Q$. Bottom: ballet action : $\mathbf{A}_Q$ is learned from a collection of 3-D LSKs $\mathbf{W}_Q$, and Feature row vectors of $\mathbf{F}_Q$ and $\mathbf{F}_T$ are computed from query Q and target video T respectively. Eigenvectors and feature vectors were transformed to volume and up-scaled for illustration purposes.

2.2. Feature Representation

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

$$\begin{aligned} \mathbf{W}_Q &= [\mathbf{w}_Q^1, \cdots, \mathbf{w}_Q^n] \in \mathbb{R}^{P^2\times n}, \\ \mathbf{W}_T &= [\mathbf{w}_T^1, \cdots, \mathbf{w}_T^{n_T}] \in \mathbb{R}^{P^2\times n_T}. \end{aligned} \tag{7}$$

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. Applying PCA to $\mathbf{W}_Q$ we can retain the first (largest) d principal components 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 $\mathbf{A}_Q$:

$$\begin{aligned} \mathbf{F}_Q &= [\mathbf{f}_Q^1, \cdots, \mathbf{f}_Q^n] = \mathbf{A}_Q^T\mathbf{W}_Q \in \mathbb{R}^{d\times n}, \\ \mathbf{F}_T &= [\mathbf{f}_T^1, \cdots, \mathbf{f}_T^{n_T}] = \mathbf{A}_Q^T\mathbf{W}_T \in \mathbb{R}^{d\times n_T}. \end{aligned} \tag{8}$$

Fig. 4 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, car, helicopter, and ballet turning action.

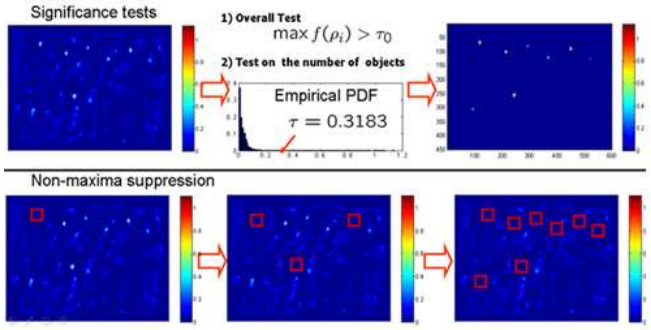

Fig. 5. Left: Non-parametric thresholding of resemblance map (RM) yields reliably similar objects which are accurately localized in the target image

2.3. Resemblance Map and Significance Testing

The next step in the proposed framework is a decision rule based on the measurement of a "distance" between the computed features $\mathbf{F}_Q, \mathbf{F}_{T_i}$. Motivated by the effectiveness of correlation-based similarity measures, we use "Matrix Cosine Similarity (MCS)" for the matrix case. The "Matrix Cosine Similarity" is defined as a natural generalization using the "Frobenius inner product" between two normalized matrices as follows:

$$\rho(\mathbf{F}_Q, \mathbf{F}_{T_i}) = <\overline{\mathbf{F}}_Q, \overline{\mathbf{F}}_{T_i}>_F = \text{trace}\Big(\frac{\mathbf{F}_Q^T \mathbf{F}_{T_i}}{\|\mathbf{F}_Q\|_F \|\mathbf{F}_{T_i}\|_F}\Big) \in [-1, 1], \quad (9)$$

where $\overline{\mathbf{F}}_Q = \frac{\mathbf{F}_Q}{\|\mathbf{F}_Q\|_F}$ and $\overline{\mathbf{F}}_T = \frac{\mathbf{F}_T}{\|\mathbf{F}_T\|_F}$.

The next step is to generate a so-called "resemblance map" (RM), which will be an image of pixels indicating the likelihood of similarity between Q and T at each pixel position. 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 function $f(\cdot)$ as follows:

$$\text{RM} : f(\rho_i) = \frac{\rho_i^2}{1 - \rho_i^2}, \quad i = 0, \cdots, M - 1. \quad (10)$$

From a quantitative point of view, we note that $f(\rho_i)$ is essentially the Lawley-Hotelling Trace statistic [18], which is used as an efficient test statistic for detecting correlation between two data sets.

Next, we employ a two-step significance test as shown in Fig 5. The first is an overall threshold (τ_0) on the RM to decide whether there is any sufficiently similar object or action present in the target at all. If the answer is yes at sufficiently high confidence, we would then want to know how many objects or actions of interest are present in the target and where they are. Therefore, we need two thresholds: an overall threshold [3] τ_o as mentioned above, and a threshold [4] τ to detect the (possibly) multiple occurrences of the same object or action in the target.

After the two significance tests with τ_o, τ are performed, we employ the idea of non-maxima suppression [13] for the final detection. We take the region with the highest $f(\rho_i)$ value and eliminate the possibility that any other object or action is detected within some radius of the center of that region again. This enables us to avoid multiple false detections of nearby objects or actions already detected. Then we iterate this process until the local maximum value falls below the threshold τ.

3. EXPERIMENTAL RESULTS

In this section, we show experimental results on several challenging datasets such as the general object dataset [5], the human action dataset [6], the Weizmann action dataset [19], and the KTH action dataset [20]. Our method detects the presence and location of objects (actions) similar to the given query and provides a series of bounding boxes (cubes) with resemblance map embedded around detected objects (actions). Note that no background/foreground segmentation is required in the proposed method. This method can also handle modest amount of variations in rotation (up to ± 15 degrees), and spatial and temporal scale change (up to $\pm 20\%$).

3.1. General object detection

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 $\alpha = 0.99$. Irani's general object dataset [5] consists of many challenging pairs of color images (60 pairs with queries such as flowers, hearts, peace symbols, face, and human poses.) Figs. 6 and 7 show qualitative results.

In order to further justify the use of LSKs, we compare the quantitative performance with state-of-the-art local descriptors evaluated in [22] as similarity done in [5]. More specif-

[3] In a typical scenario, we set the overall threshold τ_o to be, for instance, 0.96 which is about 50% 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 or action of interest.

[4] We employ the idea of nonparametric testing. We compute an empirical probability density function (PDF) from M samples $f(\rho_i)$ and we set τ so as to achieve, for instance, a 99 % ($\alpha = 0.99$) significance level in deciding whether the given values are in the extreme (right) tails of the distribution. This approach is based on the assumption that in the target, most patches do not contain the object or action of interest (in other words, object or action of interest is a relatively rare event), and therefore, the few matches will result in values which are in the tails of the distribution of $f(\rho_i)$.

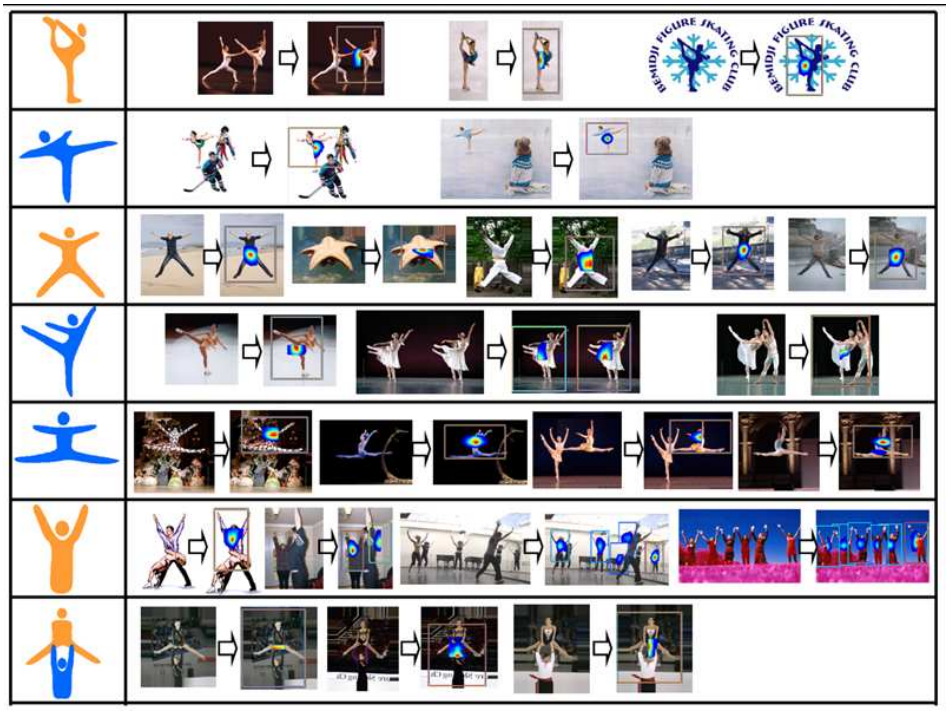

Fig. 6. Left: hand-drawn sketch query (human poses) Right: targets and examples of correction detections/ localizations in Shechtman's object test set [5]. α was set to 0.98.

Fig. 7. Query: hearts, hand-drawn face, peace symbol and flower. Some targets and examples of correction detections/ localizations in Shechtman's object test set [5] are shown. Some false positives appeared in a girl's T-shirt and candle. α was set to 0.98.

ically, we compare ROC curves. We densely computed such local descriptors as *gradient location-orientation histogram* (GLOH) [22], *Shape Context* [23], and SIFT [21] using the implementation in [22]. By replacing LSKs with these descriptors, but keeping the rest of the steps the same, we repeated the experiment on this test set. The ROC curve in Fig. 8 verifies that our LSKs have more discriminative power than other local descriptors. The proposed method is also evaluated on full CIE L*a*b* data. If we look at detection rates in the range of $0 \leq$ false positives rate $\leq 0.5 \times 10^{-4}$ in Fig. 8, we can see that full CIE L*a*b* data provide more information as also observed in [5]. Consistent with these results, it is worth noting that Shechtman and Irani [5] also showed that their local self-similarity descriptor clearly outperformed

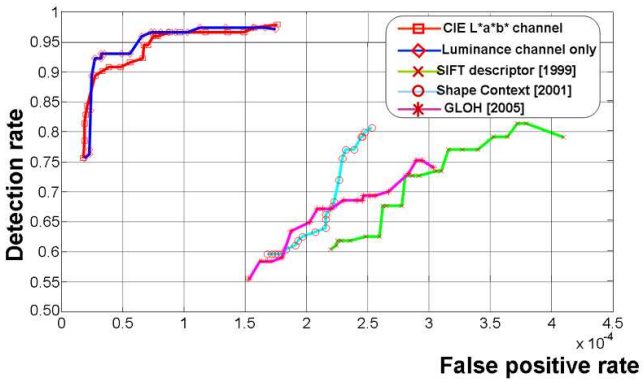

Fig. 8. Comparison of ROC curves between luminance channel only and CIE L*a*b* channel on the Shechtman's test set [5]. It is clearly shown that such descriptors as SIFT [21], GLOH [22], Shape Context [23] turn out to be inferior to LSKs in terms of discriminative power.

other state-of-the-art descriptors in their ensemble matching framework. However, the performance figures they provide are rather incomplete. Namely, they mentioned 86% 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 [5] .

3.2. Human Action Detection

In this section, we show action detection experiment results. Once given Q and T (typically Q of 60×70 pixels and T of 180×360 pixels), we blur and downsample both Q and T by a factor of 3 in order to reduce the time-complexity. We then compute 3-D LSK of size 3×3 (space) $\times 7$ (time) as descriptors so that every space-time location in Q and T yields a 63-dimensional local descriptor $\mathbf{W}_Q$ and $\mathbf{W}_T$ respectively. We end up with $\mathbf{F}_Q, \mathbf{F}_T$ by reducing dimensionality from 63 to $d = 4$ and then, we obtain RM by computing the MCS measure between $\mathbf{F}_Q, \mathbf{F}_T$. The threshold τ for each test example was determined by the 99 percent confidence level.

Fig. 9 shows the results of searching for instances of walking people in a target beach video (460 frames of 180×360 pixels). The query video contains a very short walking action moving to the right (14 frames of 60×70 pixels) and has a background context which is not the beach scene. In order to detect walking actions in either directions, we used two queries (Q and its mirror-reflected version) and generated two RMs. By voting the higher score among values from two RMs at every space-time location, we arrived at one RV which includes correct locations of walking people in the correct direction. Fig. 9 (a) shows a few sampled frames from Q. In order to provide better illustration of T, we divided T into 3 non-overlapping sections. Fig. 9 (b) and (c) represent each part of T and its corresponding RV respectively. Red color represents higher resemblance while blue color denotes lower resemblance values. Fig. 9 (d) and (e) show a few frames from T and RMs superimposed on T respectively.

3.3. Action Category Classification

Our baseline algorithm is designed for detecting actions in videos, but this method can also be extended to action classification. We conducted an extensive set of experiments to evaluate the action classification performance of the proposed method on the Weizmann action dataset ([19]) and the KTH action dataset ([20]).

3.3.1. Weizmann Action Data Set

The Weizmann action dataset contains 10 actions (bend, jumping jack, jump forward, jump in place, jump sideways, skip, run, walk, wave with two hands, and wave with one hand) performed by 9 different subjects. The testing was performed in a "leave-one-out" setting, *i.e.*, for each run the videos of 8 subjects are labeled and the videos of the remaining subject are used for testing (query). We classify each testing video as one of the 10 action types by 3-NN (nearest neighbor) as similarly done in [12]. The results are reported as the average of nine runs. We were able to achieve a recognition rate of 96% for all ten actions. The recognition rate comparison is provided in Table 1 as well. The proposed method which is training-free performs favorably against state-of-the-art methods [24, 25, 26, 27, 28, 29] which largely depend on training. We further provide the results using 1-NN and 2-NN for comparison in Table 1.

Table 1. Comparison of average recognition rate on the Weizmann dataset ([19])

Our Approach (1-NN)	Juenjo *et al.* ([25])	Liu *et al.* ([26])
90%	95.33%	90%
Our Approach (2-NN)	Niebles *et al.* ([24])	Ali *et al.* ([28])
90%	90%	95.75%
Our Approach (3-NN)	Jhuang *et al.* ([27])	Batra *et al.* ([29])
96%	98.8%	92%

3.3.2. KTH Action Data Set

In order to further verify the performance of our algorithm, we also conducted experiments on the KTH dataset. The KTH action dataset contains six types of human actions (boxing, hand waving, hand clapping, walking, jogging, and running), performed repeatedly by 25 subjects in 4 different scenarios: outdoors (c_1), outdoors with camera zoom (c_2), outdoors with different clothes (c_3), and indoors (c_4). This dataset seems more challenging than the Weizmann dataset because there are large variations in human body shape, view angles, scales, and appearance. The "leave-one-out" cross validation

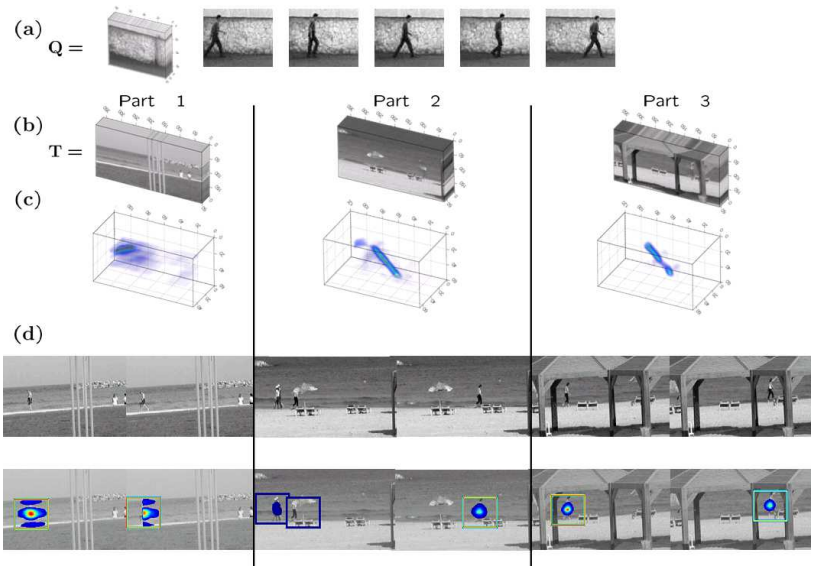

Fig. 9. Results searching for walking person on the beach (a) query video (a short walk clip) (b) target video (c) Resemblance maps (RM) (d) a few frames from T (e) frames with resemblance map on top of it.

is again used to measure the performance. More specifically, for each run the videos of 24 subjects are designated as labeled video sets and the videos of the remaining subject is used for testing. We were able to achieve a recognition rate of 95.66% on these six actions. The recognition rate comparison with competing methods is provided in Table 2 as well. It is worth noting that our method outperforms all the other state-of-the-art methods and is fully automatic.

Table 2. Comparison of average recognition rate on the KTH dataset

Our Approach (1-NN)	Kim *et al.* ([30])	Ning *et al.* ([12])
89%	95.33%	92.31% (3-NN)
Our Approach (2-NN)	Ali *et al.* ([28])	Niebles *et al.* ([24])
93%	87.7%	81.5%
Our Approach (3-NN)	Dollar *et al.* ([31])	Wong *et al.* ([32])
95.66%	81.17%	84%

3.4. Discussion

Our system is designed with recognition accuracy as a high priority. A typical run of the object detection takes about 25 second on a target image T of size 550×800 using a query Q of size 60×60 in the Intel Pentium CPU 2.66 Ghz machine. A typical run of the action detection system takes a little over 1 minute on a target video T (50 frames of 144×192 pixels) using a query Q (13 frames of 90×110). Most of the run-time is taken up by the computation of MCS (about 9 seconds, and 16.5 seconds for the computation of 3-D LSKs from Q and T respectively, which needs to be computed only once.) There are many factors that affect the precise timing of the calculations, such as query size, complexity of the video, and 3-D LSK size. Our system runs in Matlab but could be easily implemented using multi-threads or parallel programming as well as General Purpose GPU for which we expect a significant gain in speed. Even though our method is stable in the presence of moderate amount of camera motion, our system can benefit from camera stabilization methods as done in [33] and [34] in case of large camera movements.

4. CONCLUSION

In this paper, we have described a novel training-free non-parametric object and action detection, and recognition algorithm by employing *local steering kernels* (LSKs) which robustly capture underlying space-time data structure. The proposed method can automatically detect in the target the presence, the number, as well as location of objects (actions) similar to the given query. Challenging sets of real-world experiments demonstrated that the proposed approach achieves a high accuracy and improves upon other state-of-the-art methods. The proposed method does not require any prior knowledge (learning) about actions being sought; and does not require any segmentation or pre-processing step of the target.

Acknowledgment

This work was supported in part by AFOSR Grant FA 9550-07-01-0365

5. REFERENCES

[1] R. Fergus, L. Fei-Fei, P. Perona, and A. Zisserman, "Learning object categories from google's image search," *Proc. of the 10th Inter. Conf. on Computer Vision, ICCV*, vol. 2, pp. 1816–1823, October 2005.

[2] P. Viola and M. Jones, "Robust real-time object detection," *International Journal of Computer Vision*, vol. 57, no. 2, pp. 137–154, 2004.

[3] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2169–2178, 2006.

[4] 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.," *IEEE. CVPR Workshop on Generative-Model Based Vision*, , no. RR-2724, 2004.

[5] E. Shechtman and M. Irani, "Matching local self-similarities across images and videos," *In Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1–8, June 2007.

[6] 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 Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, pp. 2045–2056, November 2007.

[7] C. Yeo, P. Ahammad, K. Ramchandran, and S. S. Satry, "High-speed action recognition and localization in compressed domain videos," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 18, pp. 1006–1015, Aug 2008.

[8] T.K Kim, S.F. Wong, and R. Cipolla, "Tensor canonical correlation analysis for action classification," *In Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1–8, 2007.

[9] O. Boiman, E. Shechtman, and M. Irani, "In defense of nearest-neighbor based image classification," *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1–8, Anchorage, June 2008.

[10] H. Zhang, A.C. Berg, M. Maire, and J. Malik, "SVM-KNN: Discriminative nearest neighbor classification for visual category recognition," *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. 2, pp. 2126–2136, 2006.

[11] K. Grauman and T. Darrell, "The pyramid match kernel: Efficient learning with sets of features," *Journal of Machine Learning Research*, vol. 8, pp. 725–760, 2007.

[12] H. Ning, T.X. Han, D.B. Walther, M. Liu, and T.S. Huang, "Hierarchical space-time model enabling efficient search for human actions," *IEEE Transactions on Circuits and Systems for Video Technology, in press*, 2008.

[13] F. Devernay, "A non-maxima suppression method for edge detection with sub-pixel accuracy," *Technical report, INRIA*, , no. RR-2724, 1995.

[14] H. J. Seo and P. Milanfar, "Training-free, generic object detection using locally adaptive regression kernels," *Submitted to IEEE Transactions on Pattern Analysis and Machine Intelligence*, October 2008.

[15] H. J. Seo and P. Milanfar, "Generic human action recognition from a single example," *Submitted to IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2009.

[16] H. Takeda, S. Farsiu, and P. Milanfar, "Kernel regression for image processing and reconstruction," *IEEE Transactions on Image Processing*, vol. 16, no. 2, pp. 349–366, February 2007.

[17] H. Takeda, P. Milanfar, Matan Protter, and Michael Elad, "Super-resolution without explicit subpixel motion estimation," *EEE Transactions on Image Processing*, vol. 18, no. 9, 2009.

[18] M.M. Tatsuoka, *Multivariate Analysis*, Macmillan, 1988.

[19] L. Gorelick, M. Blank, E. Shechtman, M. Irani, and R. Basri, "Actions as space-time shapes," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, pp. 2247–2253, December 2007.

[20] C. Schuldt, I. Laptev, and B. Caputo, "Recognizing human actions: A local svm approach," *IEEE Conference on Pattern Recognition (ICPR)*, vol. 3, pp. 32–36, June 2004.

[21] D.G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 20, pp. 91–110, 2004.

[22] K. Mikolajczyk and C. Schmid, "A performance evaluation of local descriptors," *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, vol. 27, no. 10, pp. 1615–1630, 2005.

[23] S. Belongie, J. Malik, and J. Puzicha, "Shape matching and object recognition using shape contexts," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 4, pp. 509–522, April 2002.

[24] J.C. Niebles, H. Wang, and L. Fei Fei, "Unsupervised learning of human action categores usig spatial-temporal words," *International Journal of Computer Vision*, vol. 79, pp. 299–318, March 2008.

[25] I.N. Junejo, E. Dexter, I. Laptev, and P. Prez, "Cross-view action recognition from temporal self-similarities.," *In Proc. European Conference Computer Vision (ECCV'08)*, vol. 2, pp. 293–306, October 2008.

[26] J. Liu, S. Ali, and M. Shah, "Recognizing human actions using multiple features," *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1–8, June 2008.

[27] H. Jhuang, T.Serre, L.Wolf, and T.Poggio, "A biologically inspired system for action recogntion," *IEEE International Conference on Computer Vision(ICCV)*, pp. 1–8, October 2007.

[28] S. Ali and M. Shah, "Human action recognition in videos using kinematic features and multiple instance learning," *Accepted for publication in IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 2008.

[29] D. Batra, T. Chen, and R. Sukthankar, "Space-time shapelets for action recogntion," *IEEE Workshop on Motion and video Computing (WMVC)*, pp. 1–6, January 2008.

[30] T.K. Kim, S.F. Wong, and R. Cipolla, "Tensor canonical correlation analysis for action classification," *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1–3, 2007.

[31] P. Dollar, V. Rabaud, G. Cottrell, and S. Belongie, "Behavior recognition via sparse spatio-temporal features," *In proceeding of Visual Surveillance and Performance Evaluation of Tracking and Surveillance (VS-PETS)*, pp. 65–72, October 2005.

[32] S.-F. Wong, T.-K. Kim, and R. Cipolla, "Learning motion categories using both semantic and structural informtion," *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1–6, June 2007.

[33] G. Medioni, I. Cohen, F. Bremond, S. Hongeng, and R. Nevatia, "Event detection and analysis from video streams," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, pp. 873–890, August 2001.

[34] P.B. Thomas Veit, F. Cao, and P. Bouthemy, "Probabilistic parameter-free motion detection," *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 715–721, June 2004.