TRAINING-FREE, ONE-SHOT DETECTION

Siddhant Katyan - 2018801005 [**]

November 2018

1 Introduction

One shot, generic object detection involves searching for a single query object in a larger target image. Relevant approaches have benefited from features that typically model the local similarity patterns. 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.

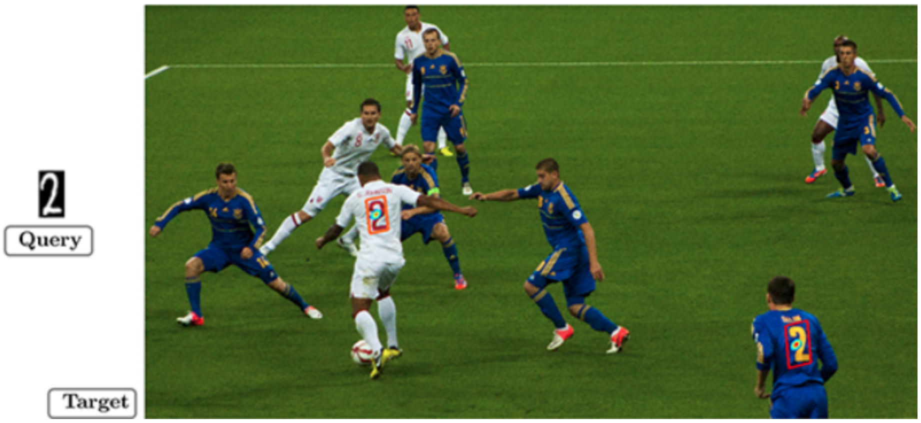


Figure 1: Overview of our one shot detection scheme.

[*] Enrolled in CSE/ECE 478, Monsoon 2018 at IIIT Hyderabad

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.

2 Methodology

The pipeline of our application is shown in Figure 2. We describe each of the stages of our pipeline in detail in this section.

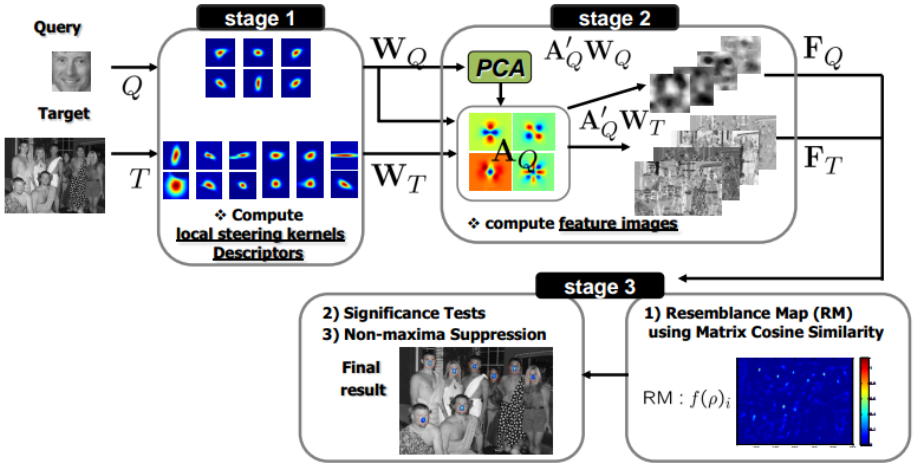
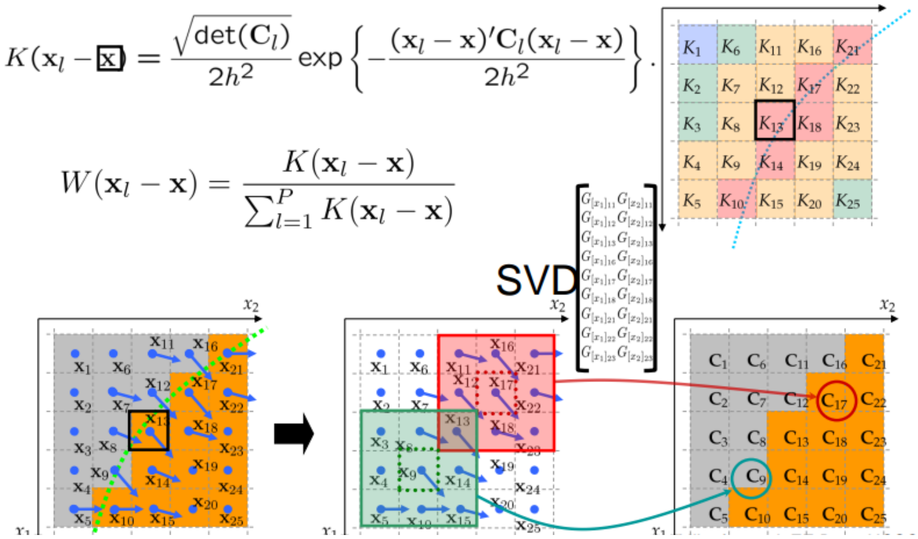


Figure 2: Pipeline of One-Shot Detection (there are broadly three stages).

2.1 Stage 1: Calculation of Local Descriptors

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 distortion. The origin and motivation behind the use of these local kernels is the earlier work on adaptive kernel regression for image processing and reconstruction [5]. 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 as shown in Figure 3. 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.



$$K(\mathbf{x}_l - \boxed{\mathbf{x}}) = \frac{\sqrt{\det(\mathbf{C}_l)}}{2h^2}\exp\left\{-\frac{(\mathbf{x}_l-\mathbf{x})'\mathbf{C}_l(\mathbf{x}_l-\mathbf{x})}{2h^2}\right\}.$$

$$W(\mathbf{x}_l - \mathbf{x}) = \frac{K(\mathbf{x}_l - \mathbf{x})}{\sum_{l=1}^{P} K(\mathbf{x}_l - \mathbf{x})}$$

Figure 3: Calculation of Local Descriptors

2.2 Stage 2: Feature Extraction from Descriptors

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. The feature collections from Q and T_i form feature matrices F_Q and F_T .We compare the feature matrices F_T and F_Q from the ith patch of T and Q to look for matches. So Instead of using selective feature techniques such as SIFT, which 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.

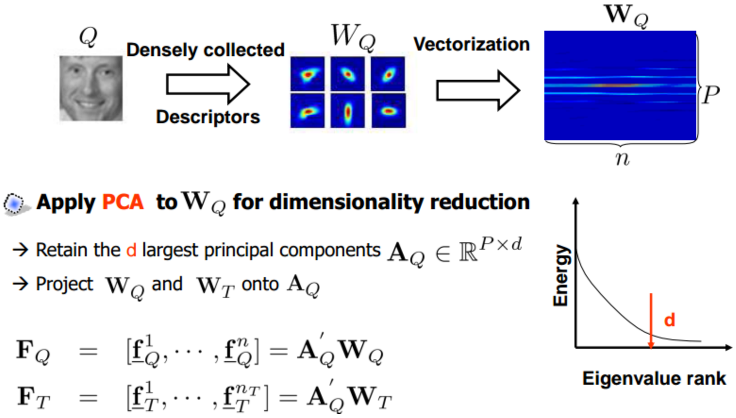

Q
Densely collected
Descriptors
W_Q
Vectorization
$\mathbf{W}_Q$
P
n
Apply PCA to $\mathbf{W}_Q$ for dimensionality reduction
→ Retain the d largest principal components $\mathbf{A}_Q \in \mathbb{R}^{P\times d}$
→ Project $\mathbf{W}_Q$ and $\mathbf{W}_T$ onto $\mathbf{A}_Q$
$\mathbf{F}_Q = [\mathbf{f}_Q^1, \cdots, \mathbf{f}_Q^n] = \mathbf{A}_Q^{'}\mathbf{W}_Q$
$\mathbf{F}_T = [\mathbf{f}_T^1, \cdots, \mathbf{f}_T^{n_T}] = \mathbf{A}_Q^{'}\mathbf{W}_T$
Energy
d
Eigenvalue rank

Figure 4: Learning relevant features of the data.

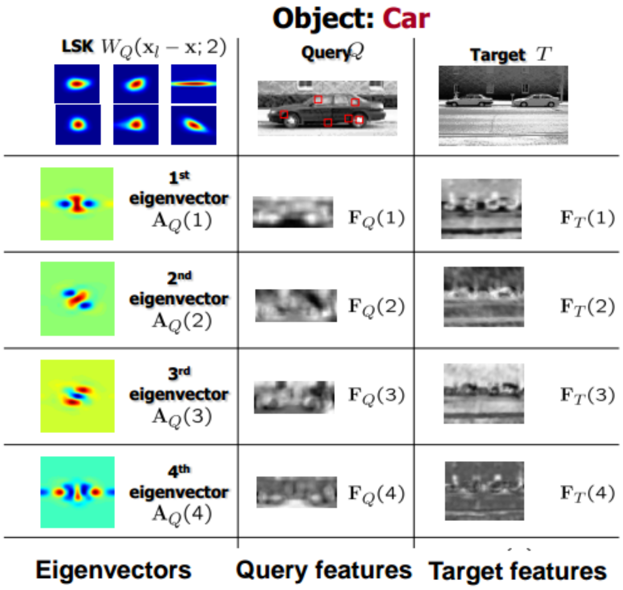


Figure 5: Salient Features after PCA.

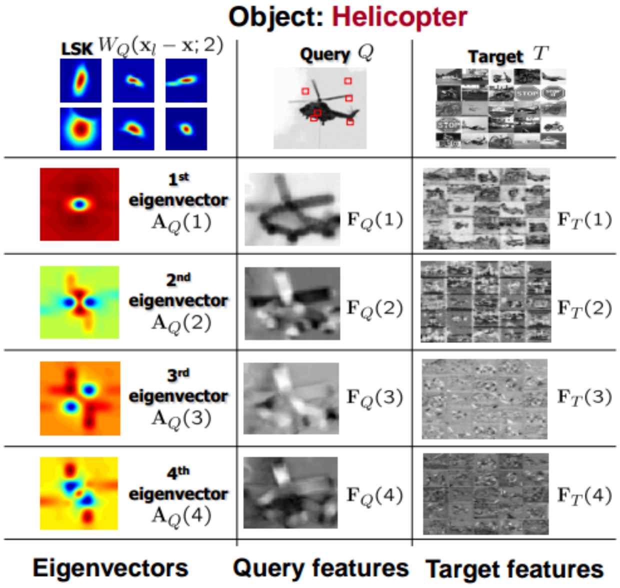


Figure 6: Salient Features after PCA.

2.3 Stage 3: Finding similarity between features.

Instead of the conventional Euclidean Distance, we employ and justify use of “Matrix Cosine Similarity” as a similarity measure which generalizes the cosine similarity between two vectors 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.

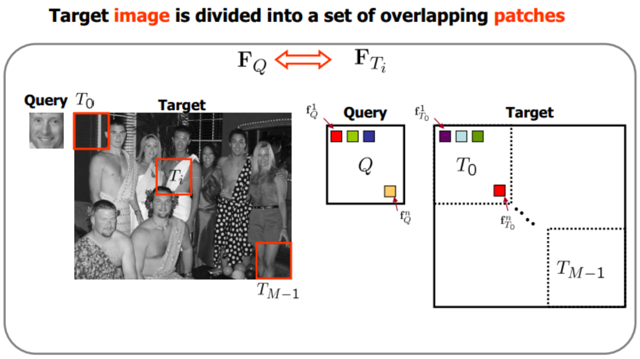
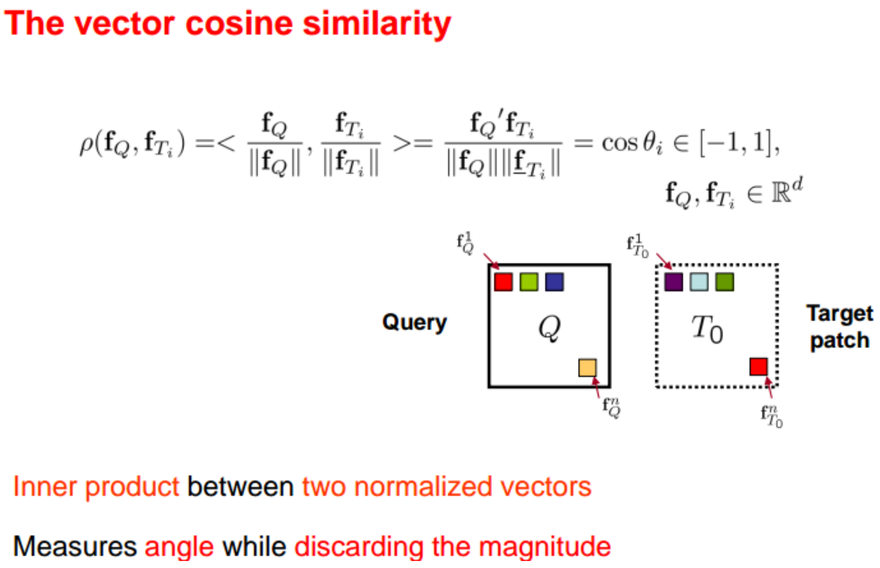


Figure 7: Dividing target images into patches.

2.4 Stage 3: Coorelation Based Metric

In general, “correlation” indicates the strength and direction of a linear relationship between two random variables. But the idea of correlation is quite malleable. 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. 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:



The vector cosine similarity

$$\rho(\mathbf{f}_Q, \mathbf{f}_{T_i}) =< \frac{\mathbf{f}_Q}{\|\mathbf{f}_Q\|}, \frac{\mathbf{f}_{T_i}}{\|\mathbf{f}_{T_i}\|} >= \frac{\mathbf{f}_Q{}'\mathbf{f}_{T_i}}{\|\mathbf{f}_Q\|\|\mathbf{f}_{T_i}\|} = \cos\theta_i \in [-1, 1], \quad \mathbf{f}_Q, \mathbf{f}_{T_i} \in \mathbb{R}^d$$

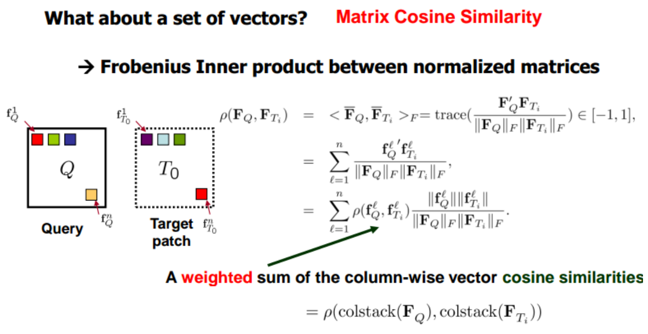
Inner product between two normalized vectors

Measures angle while discarding the magnitude

Figure 8: Vector Cosine Similarity

2.5 Stage 3: Matrix Cosine Similarity

If we deal with the features F_Q, F_T 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:



What about a set of vectors? Matrix Cosine Similarity

→ Frobenius Inner product between normalized matrices

$$\begin{aligned}\rho(\mathbf{F}_Q, \mathbf{F}_{T_i}) &= <\overline{\mathbf{F}}_Q, \overline{\mathbf{F}}_{T_i}>_F = \mathrm{trace}(\frac{\mathbf{F}_Q'\mathbf{F}_{T_i}}{\|\mathbf{F}_Q\|_F\|\mathbf{F}_{T_i}\|_F}) \in [-1, 1], \\ &= \sum_{\ell=1}^{n} \frac{\mathbf{f}_Q^\ell{}'\mathbf{f}_{T_i}^\ell}{\|\mathbf{F}_Q\|_F\|\mathbf{F}_{T_i}\|_F}, \\ &= \sum_{\ell=1}^{n} \rho(\mathbf{f}_Q^\ell, \mathbf{f}_{T_i}^\ell)\frac{\|\mathbf{f}_Q^\ell\|\|\mathbf{f}_{T_i}^\ell\|}{\|\mathbf{F}_Q\|_F\|\mathbf{F}_{T_i}\|_F}.\end{aligned}$$

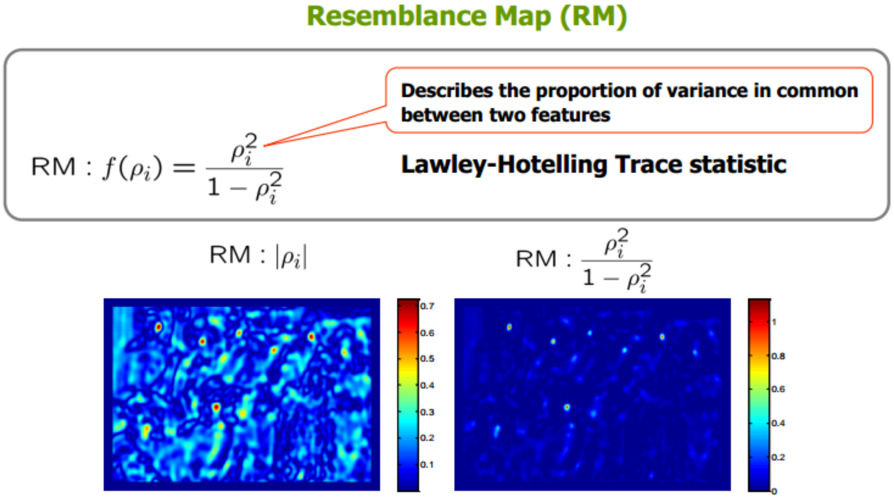
A weighted sum of the column-wise vector cosine similarities

$$= \rho(\mathrm{colstack}(\mathbf{F}_Q), \mathrm{colstack}(\mathbf{F}_{T_i}))$$

Figure 9: Matrix Cosine Similarity

2.6 Stage 3: Generating Resemblance Map

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," ρ_i^2 describes the proportion of variance in common between the two feature sets as opposed to ρ_i, which indicates a linear relationship between two matrices F_Q, F_T. 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-ρ_i^2). More specifically, RM is computed using the mapping function f as follows:



Resemblance Map (RM)

$$\text{RM} : f(\rho_i) = \frac{\rho_i^2}{1-\rho_i^2}$$

Describes the proportion of variance in common between two features

Lawley-Hotelling Trace statistic

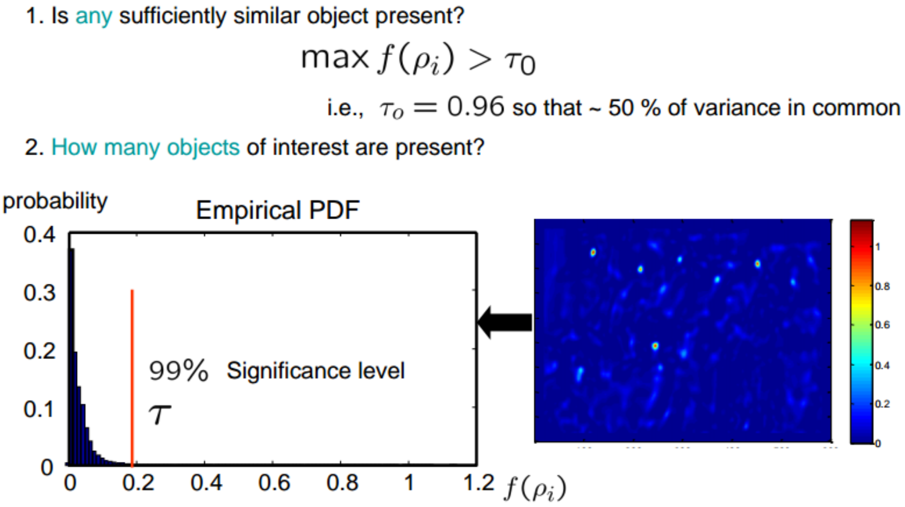
$\text{RM} : |\rho_i|$

$\text{RM} : \frac{\rho_i^2}{1-\rho_i^2}$

Figure 10: Generating Resemblance Map

2.7 Stage 3: Non-parametric Significance Tests

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 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_0 and a threshold τ to detect the possibly multiple objects present in the target image.



1. Is any sufficiently similar object present?

$$\max f(\rho_i) > \tau_0$$

i.e., $\tau_o = 0.96$ so that ~ 50 % of variance in common

2. How many objects of interest are present?

Figure 11: Significance Tests

After the two significance tests with τ_0 and τ are performed, we employ nonmaxima suppression 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 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.

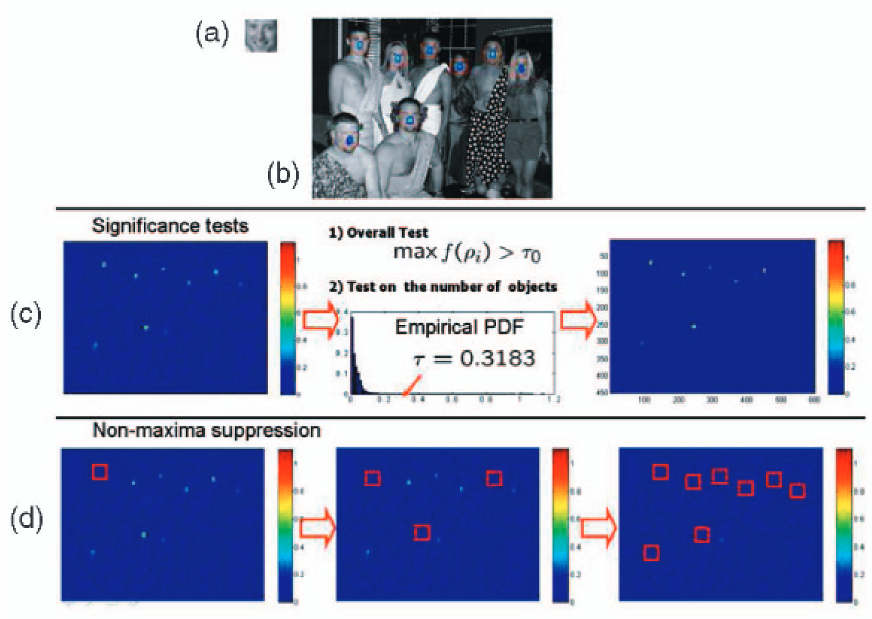


Figure 12: Non Maxima Suppression

3 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, MIT-CMU face data set and Shechtman's general object data set. The proposed algorithm provides a series of bounding boxes around objects of interest. 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.

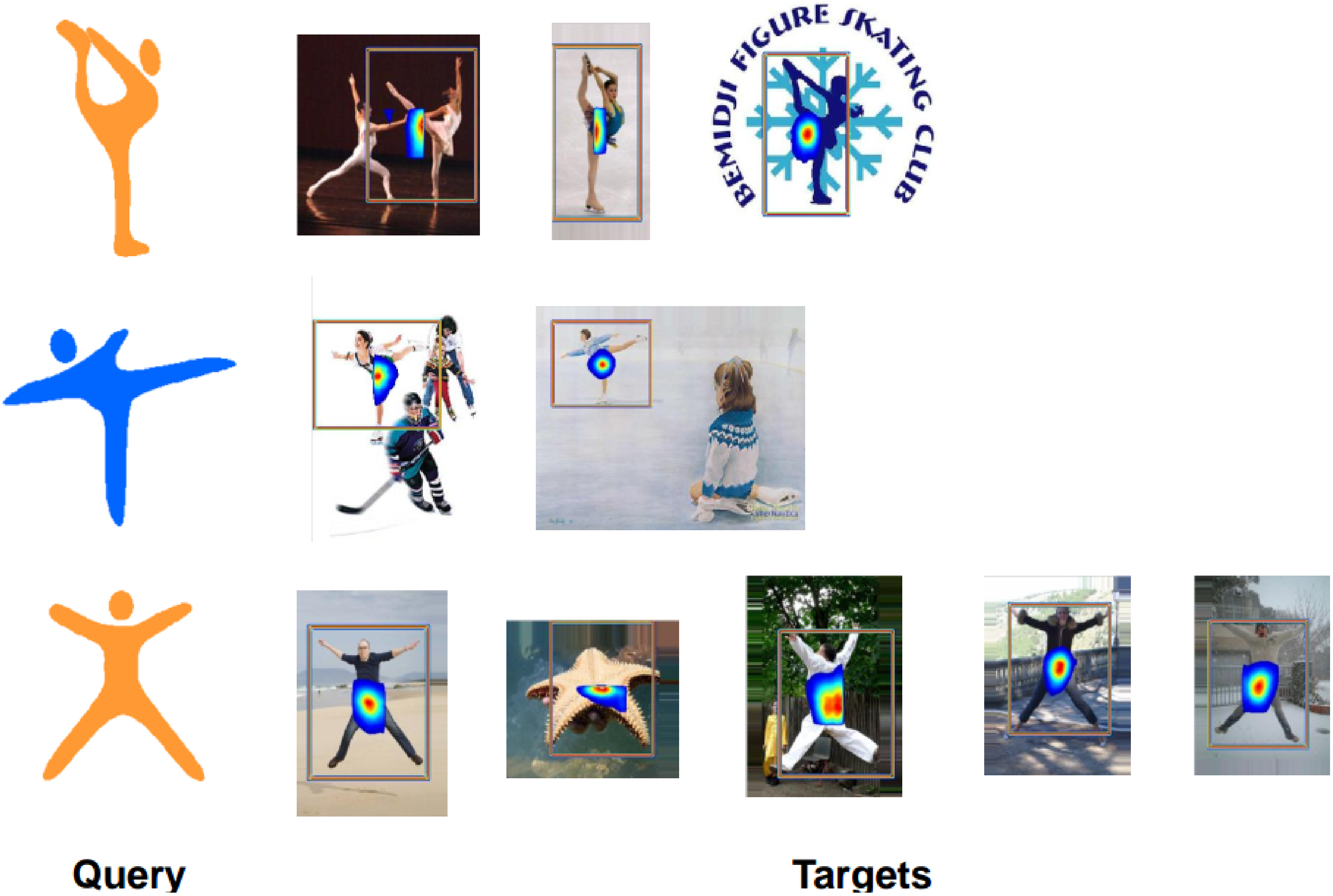


Figure 13: Hand-drawn sketch query (human poses). Right: Targets and examples of correction detections/localizations in Shechtman and Irani's object test set.

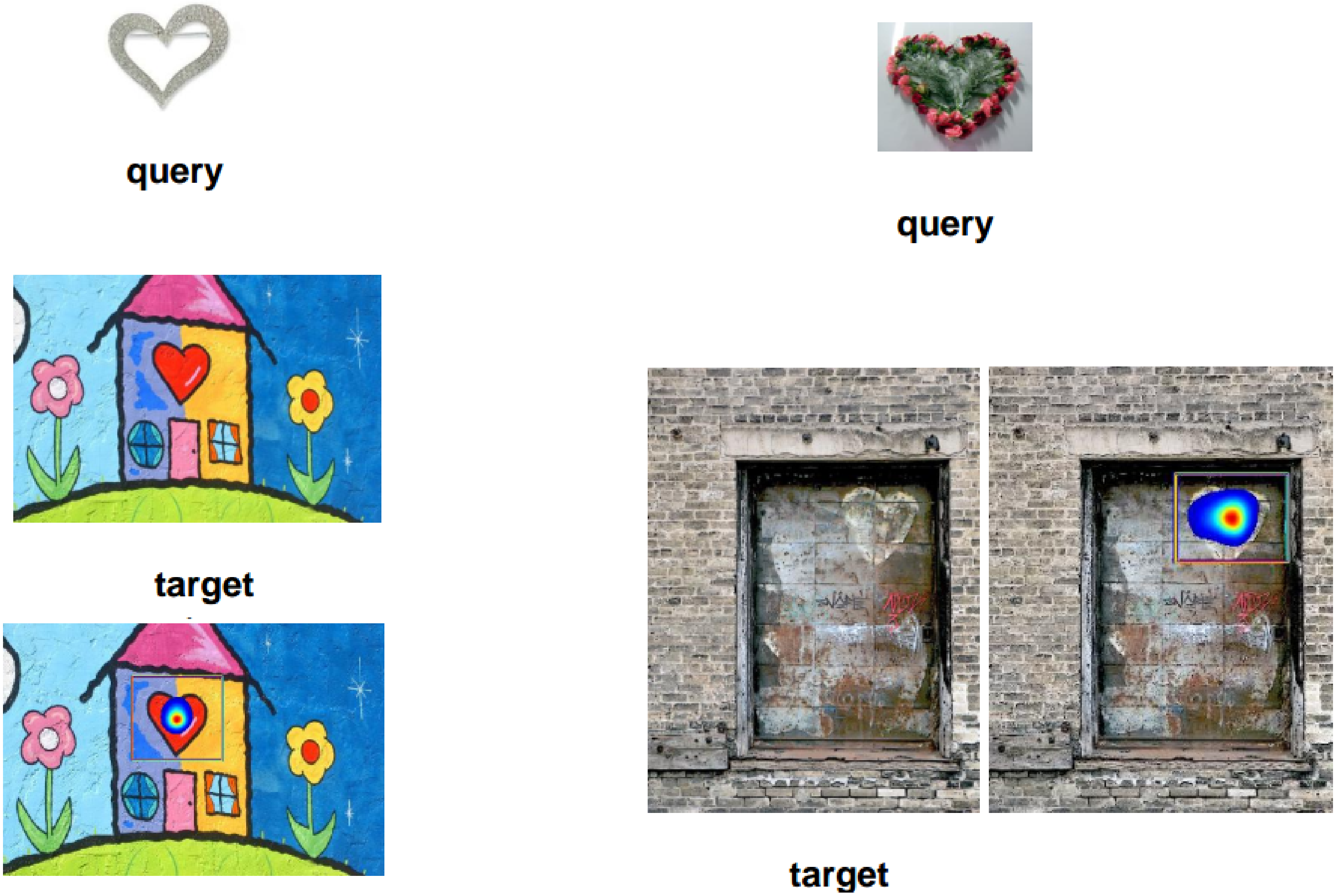


Figure 14: Query: Hearts. Targets and examples of correction detections/localizations in Shechtman and Irani's object test set are shown.

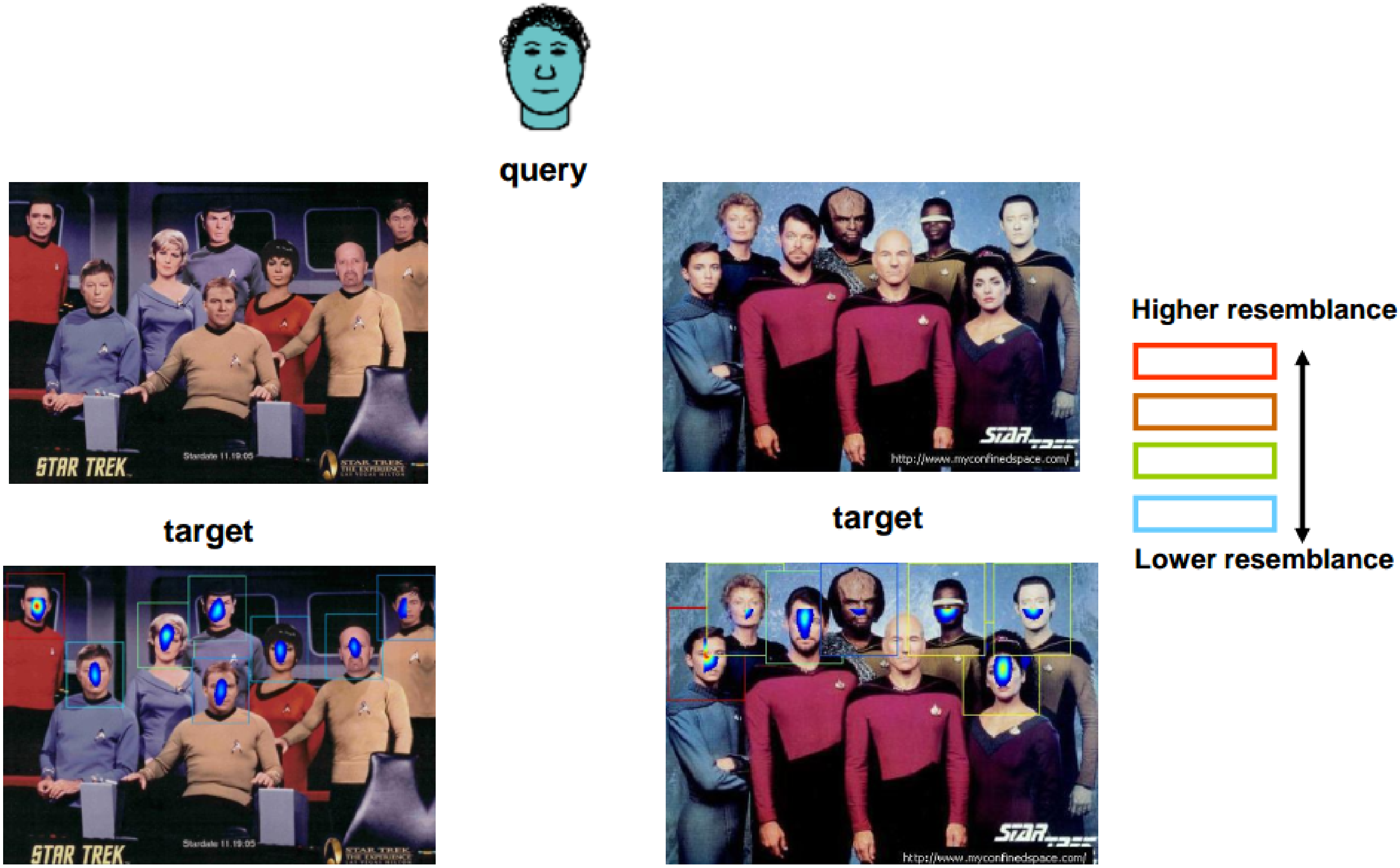


Figure 15: Query: Hand-drawn face

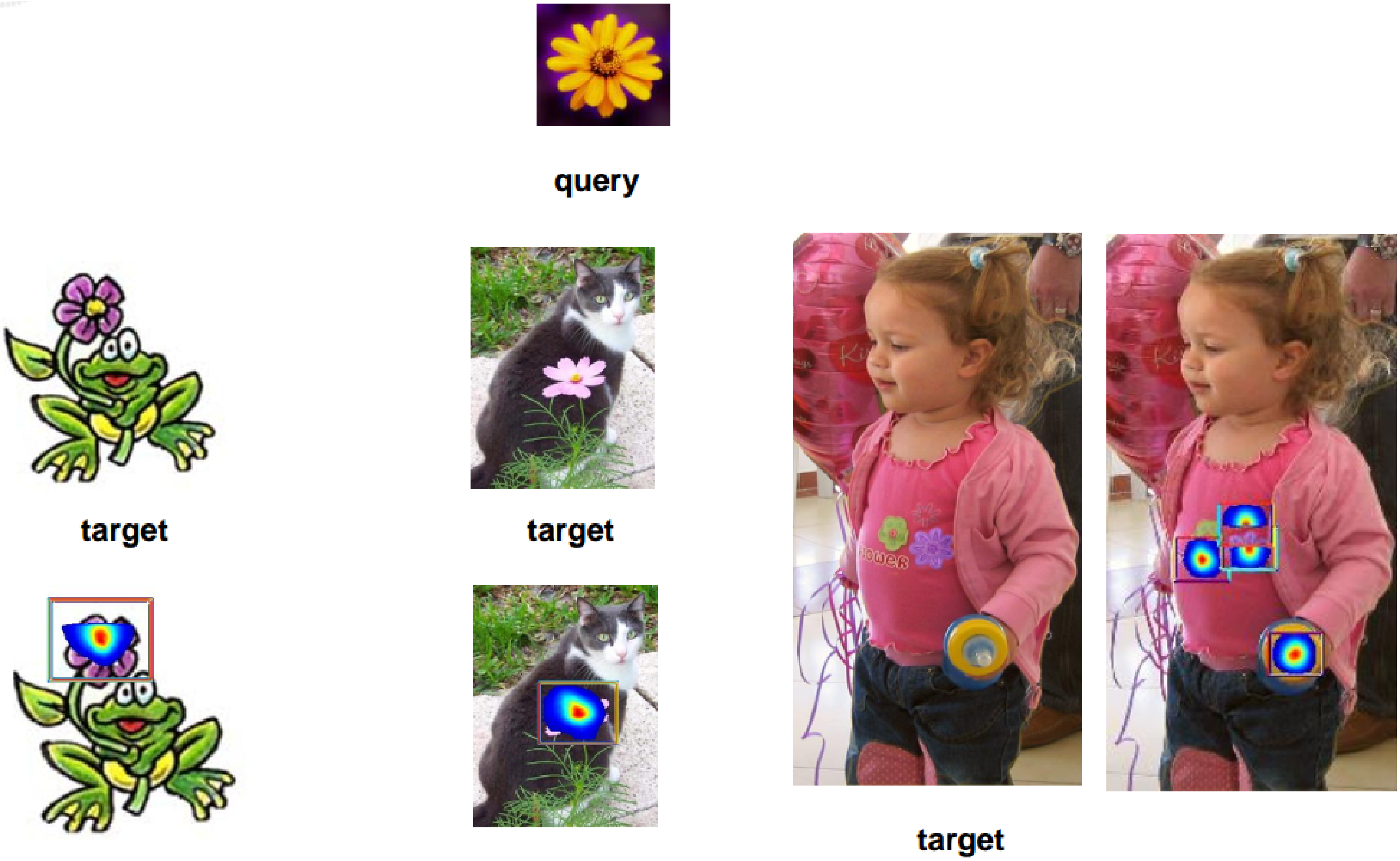


Figure 16: Query: Flower. Some false positives appeared in a girl's T-shirt and candle.

3.1 Limitations of the algorithm

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. For the proposed method to be feasible for scalable image retrieval, we may adopt the idea of encoding the features as proposed in Tree Histogram Coding.

References

[1] J. Ponce, M. Hebert, C. Schmid, and A. Zisserman, *"Toward Category-Level Object Recognition". Springer, 2007.*

[2] 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*

[3] 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.*

[4] Peyman Milanfar *"Face Verification Using the LARK Representations",IEEE Transactions on Information Forensics and Security.*

[5] Jianjun QianJian Yang *"A Novel Feature Extraction Method for Face Recognition under Different Lighting Conditions",Chinese Conference on Biometric Recognition.*