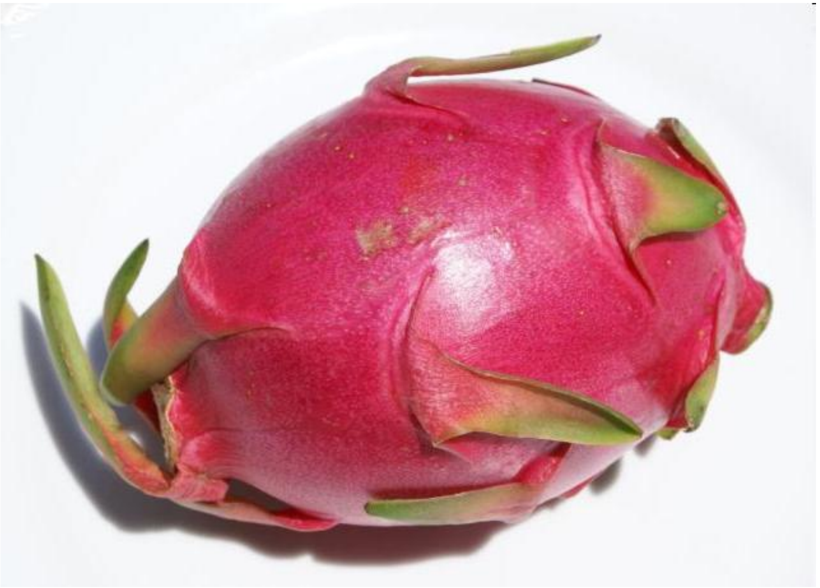
Object and Action Detection from a Single Example

Peyman Milanfar*
EE Department
University of California, Santa Cruz

*Joint work with Hae Jong Seo

AFOSR Program Review, June 4-5, 2009

Take a look at this:

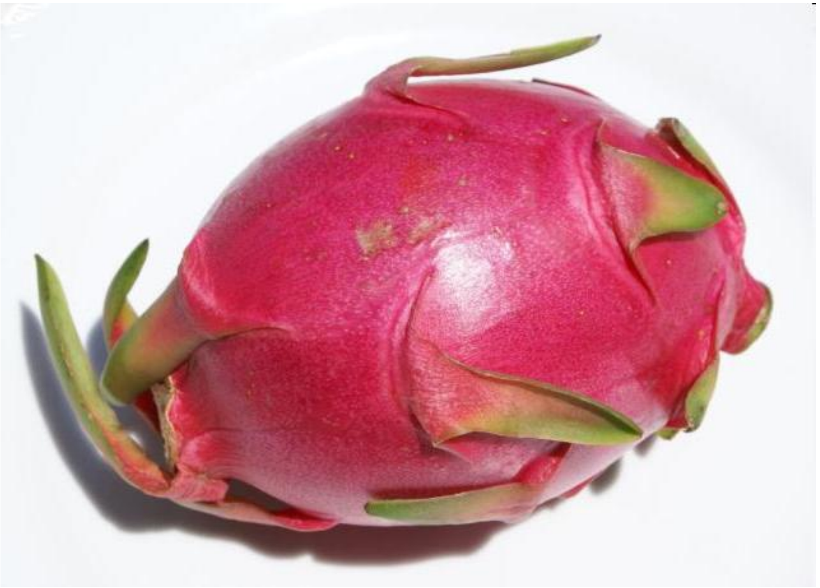


See it here?

How about here?

Or here?

Single Example, No Training!



(Most) people can find the Dragon Fruit from one look.

Even if they've never seen it before.

Outline

I. Motivation

II. Overview

III. Object Detection

IV. Action Detection

V. Conclusion and Future work

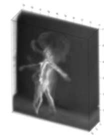
Fundamental Problems in Machine Vision

Develop a unified framework that can robustly detect objects/actions of interest within images/videos <u>without training</u>

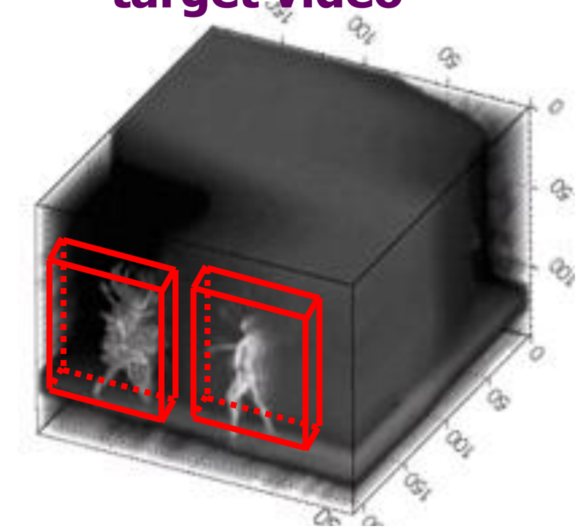
query

target image

query



target video



1) Whether objects (actions) are present or not,

2) How many objects (actions)?

3) Where are they located?

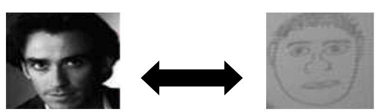
Milanfar et al. EE Dept, UCSC

Challenges in Detection

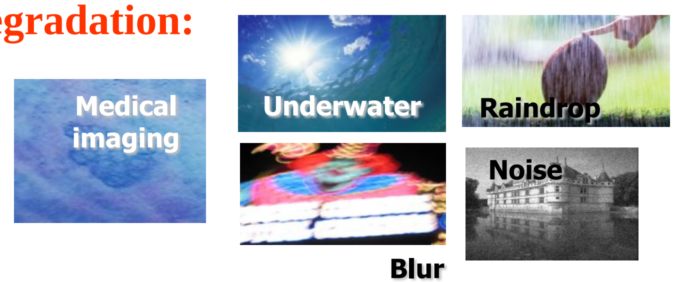
❖**Objects**

Besides,

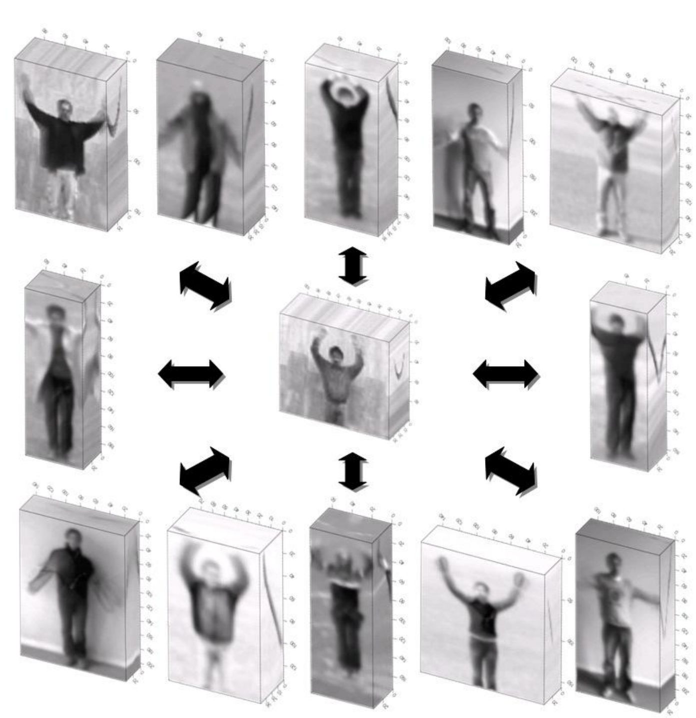
Contexts:



Degradation:



❖**Actions**



1) **different clothes,**
2) **different illumination,**
3) **different background**
4) **action speed**

Milanfar et al. EE Dept, UCSC

Outline

I. Motivation

II. System Overview

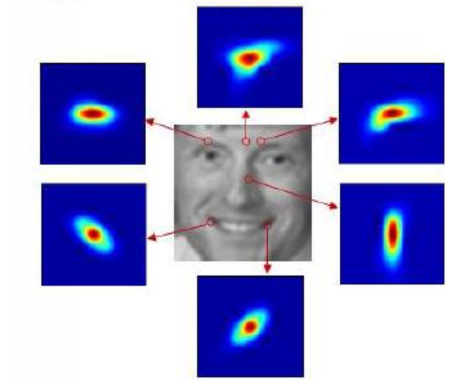
III. Object Detection

IV. Action Detection

V. Conclusion and Future work

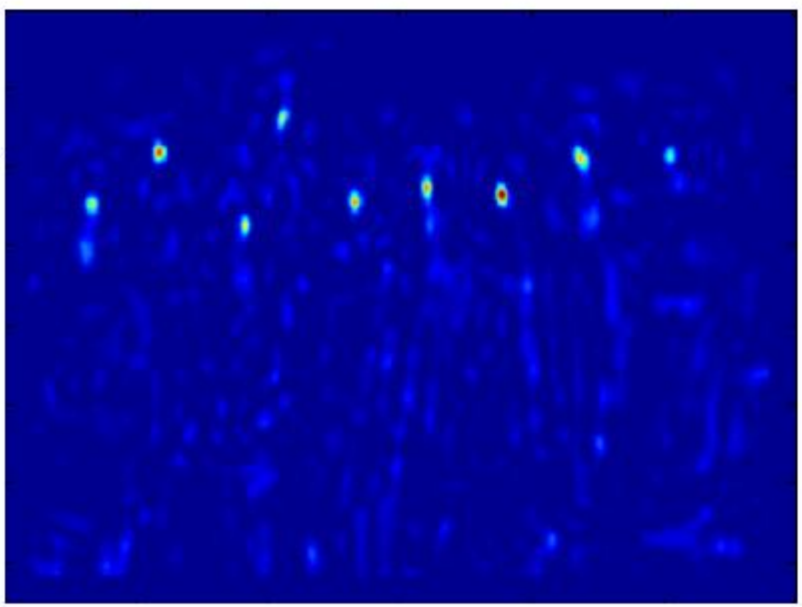
Object Detection using Local Regression Kernels

- Local Steering Kernels as *Descriptors*
- Using a *single* example

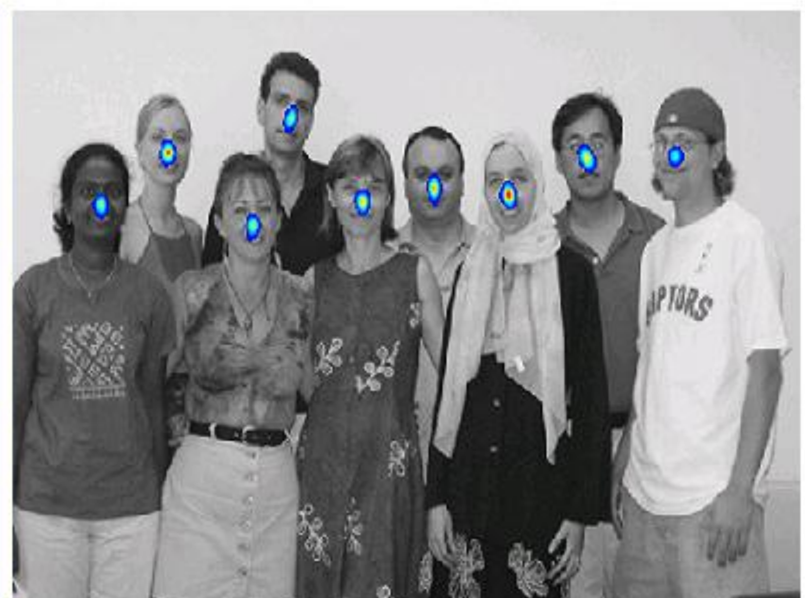


Query

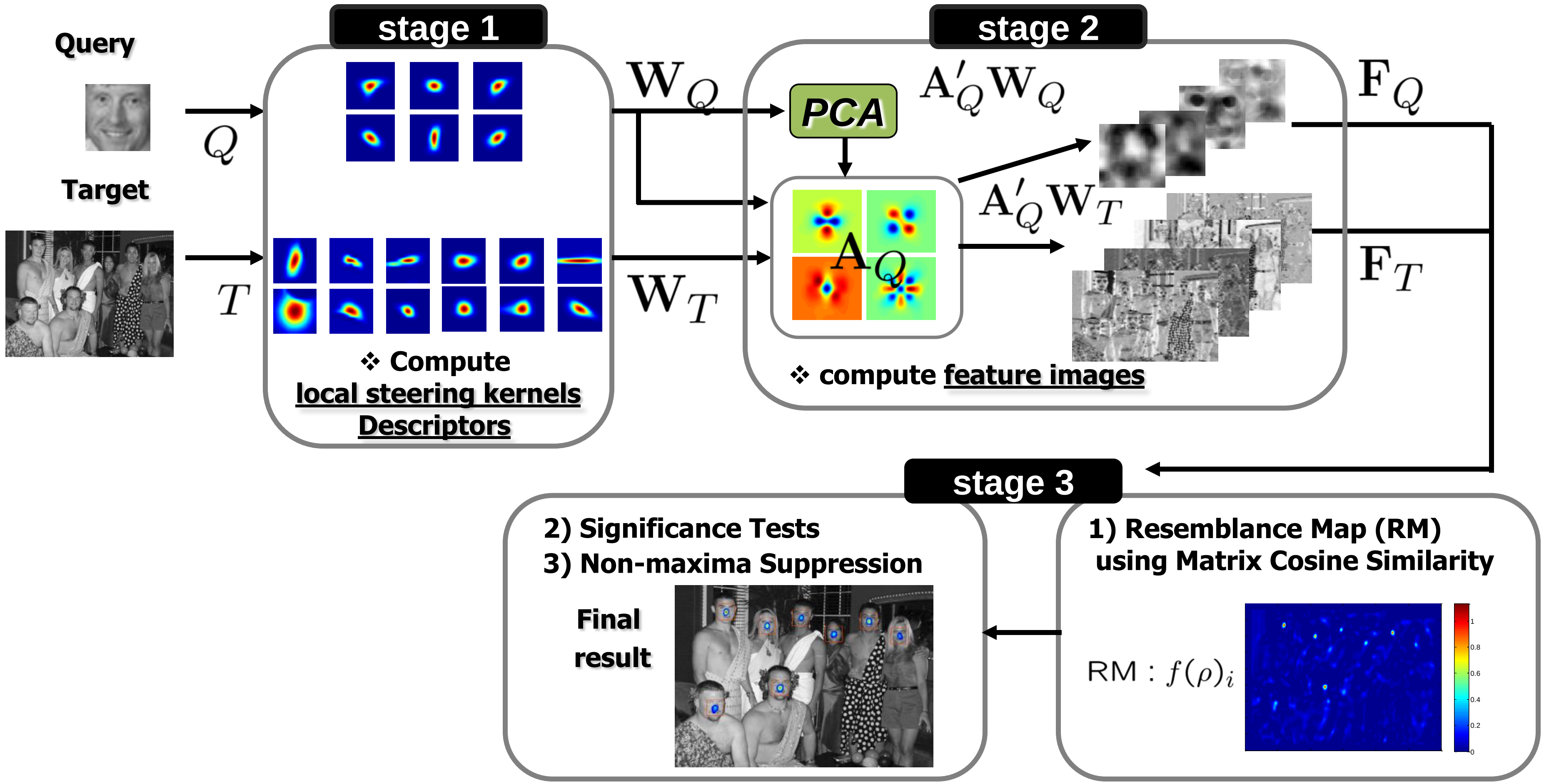
“Resemblance Map”



Detected Similar Objects



Object Detection System Overview



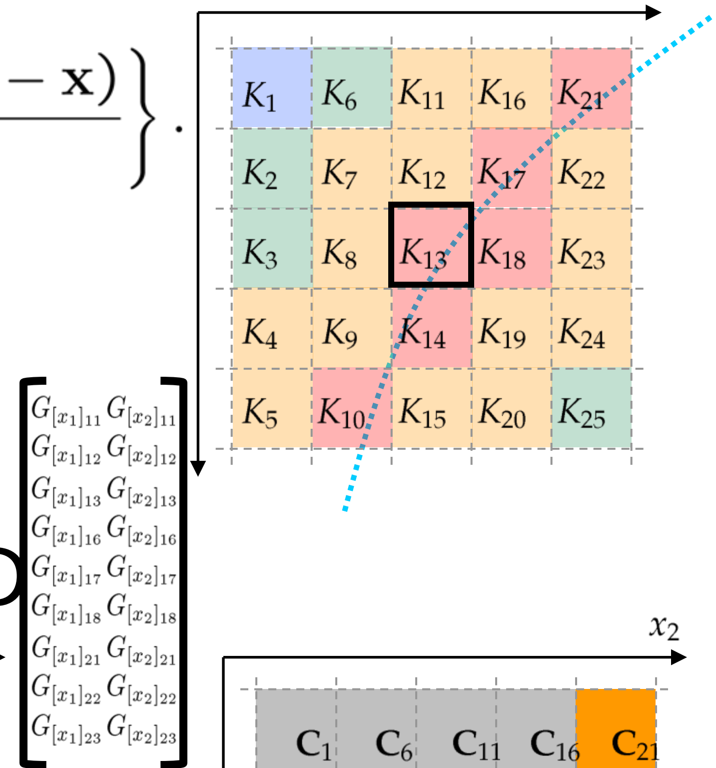
H. Seo and P. Milanfar, "**Training-free, Generic Object Detection using Locally Adaptive Regression Kernels**", Accepted for publication in *IEEE Transactions on Pattern Analysis and Machine Intelligence*

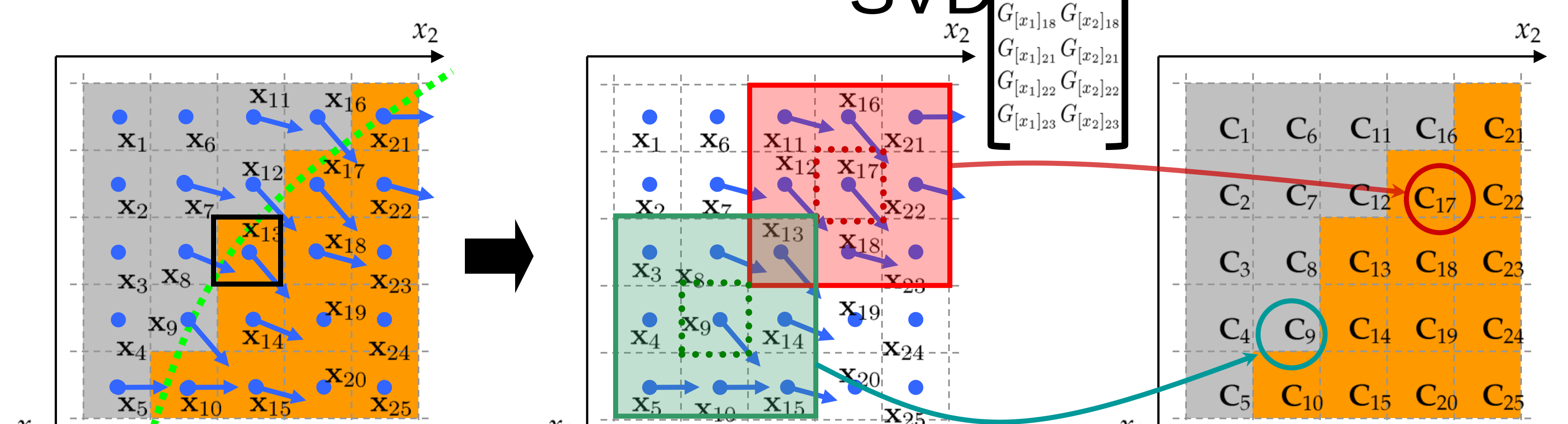
Stage 1: Calculation of Local Descriptors

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

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

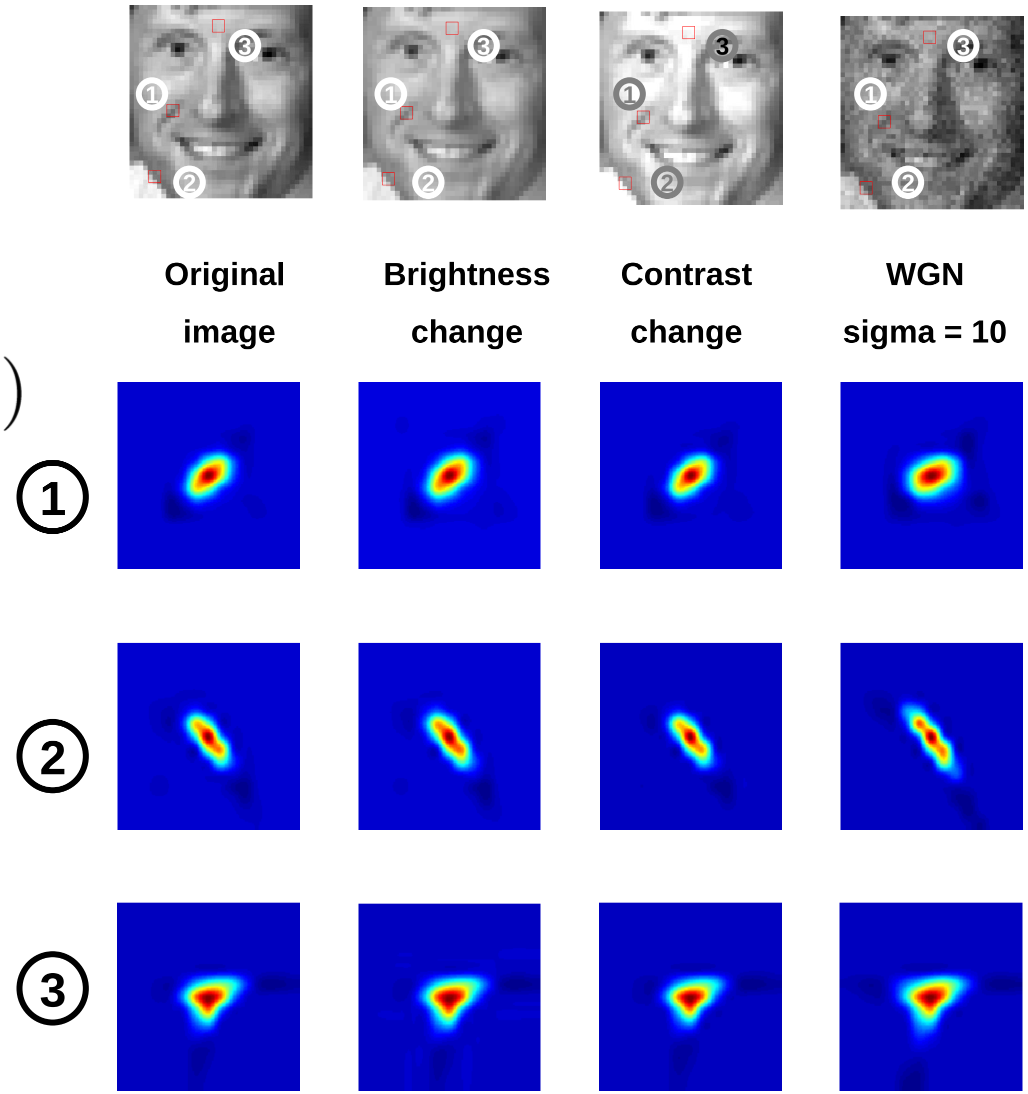
SVD



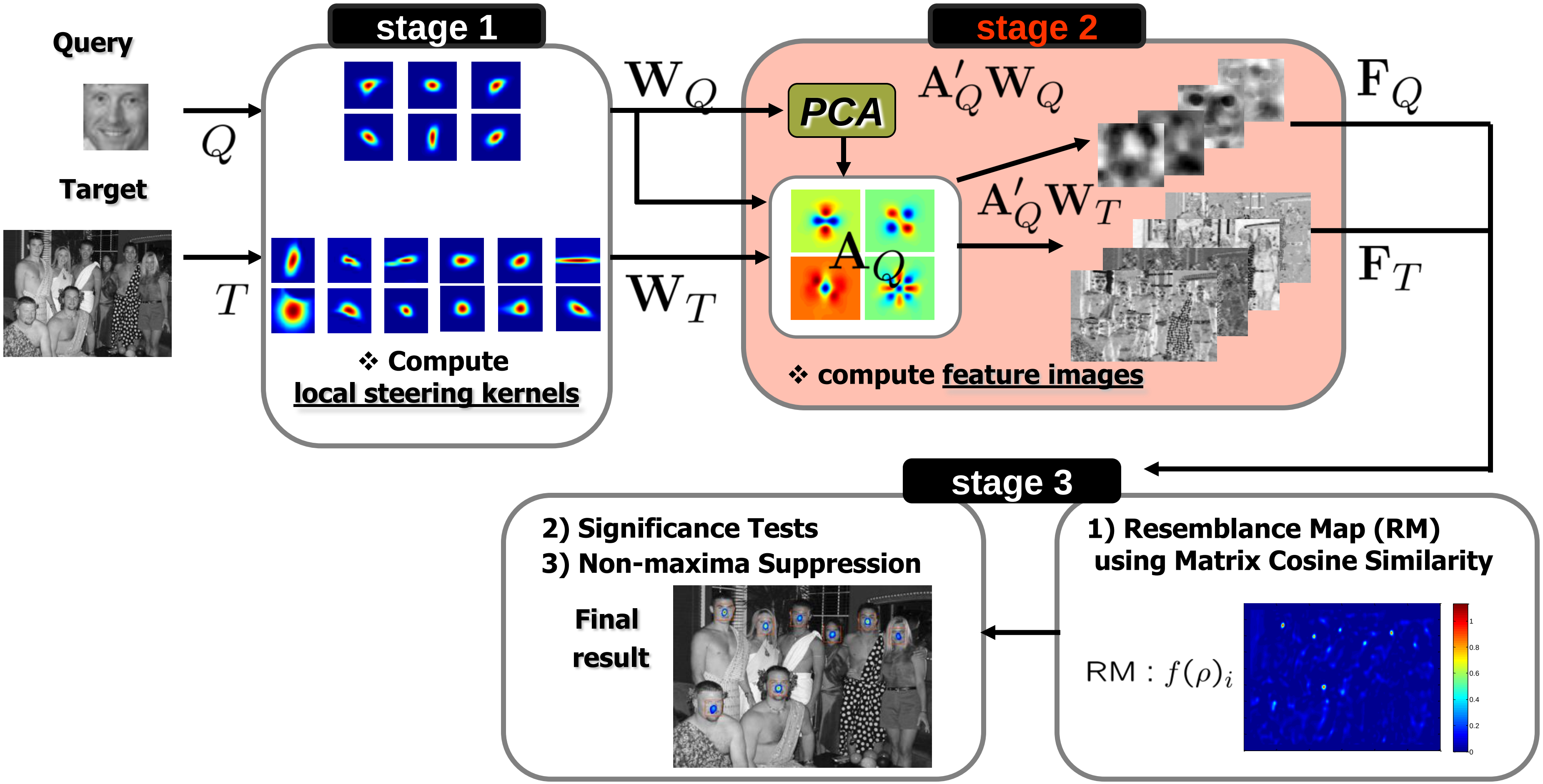


Robustness of LSK Descriptors

$$W_Q(\mathbf{x}_l - \mathbf{x})$$

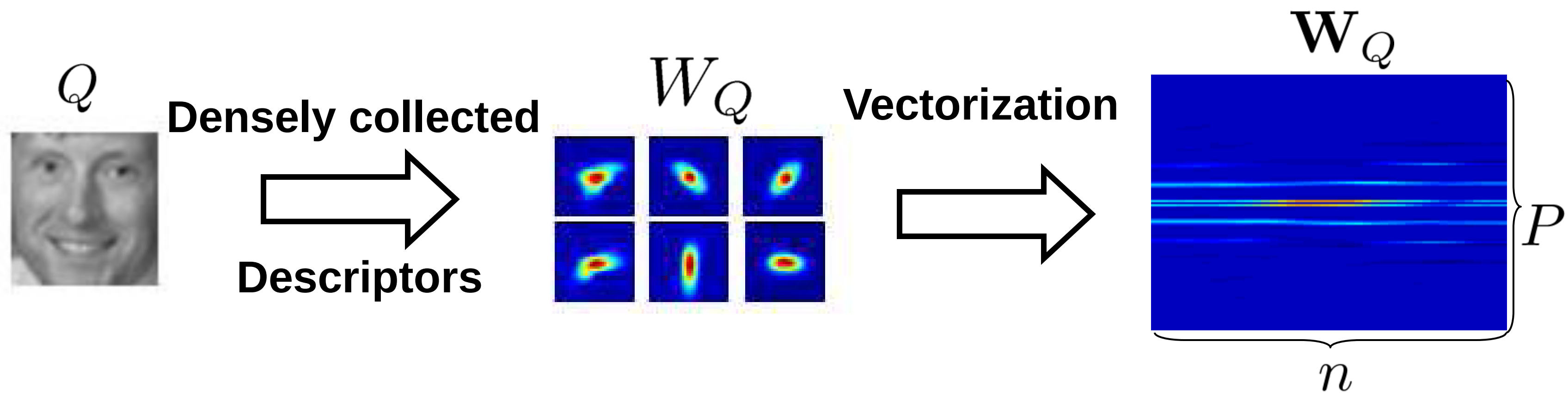


System Overview : Stage 2



Milanfar et al. EE Dept, UCSC

Stage 2: Feature Extraction from Descriptors



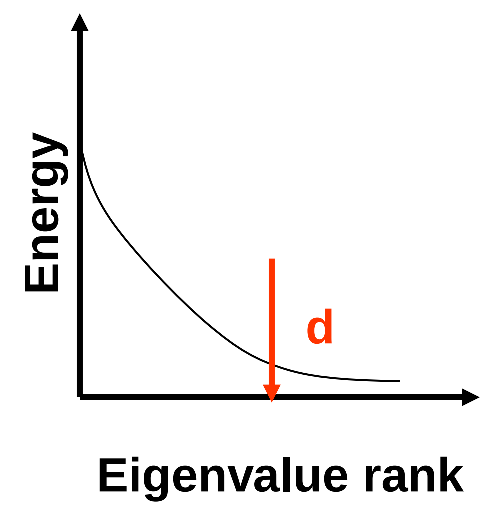
- **Apply PCA to $\mathbf{W}_Q$ for dimensionality reduction**

→ Retain the d largest principal components $\mathbf{A}_Q \in \mathbb{R}^{P \times d}$

→ Project $\mathbf{W}_Q$ and $\mathbf{W}_T$ onto $\mathbf{A}_Q$

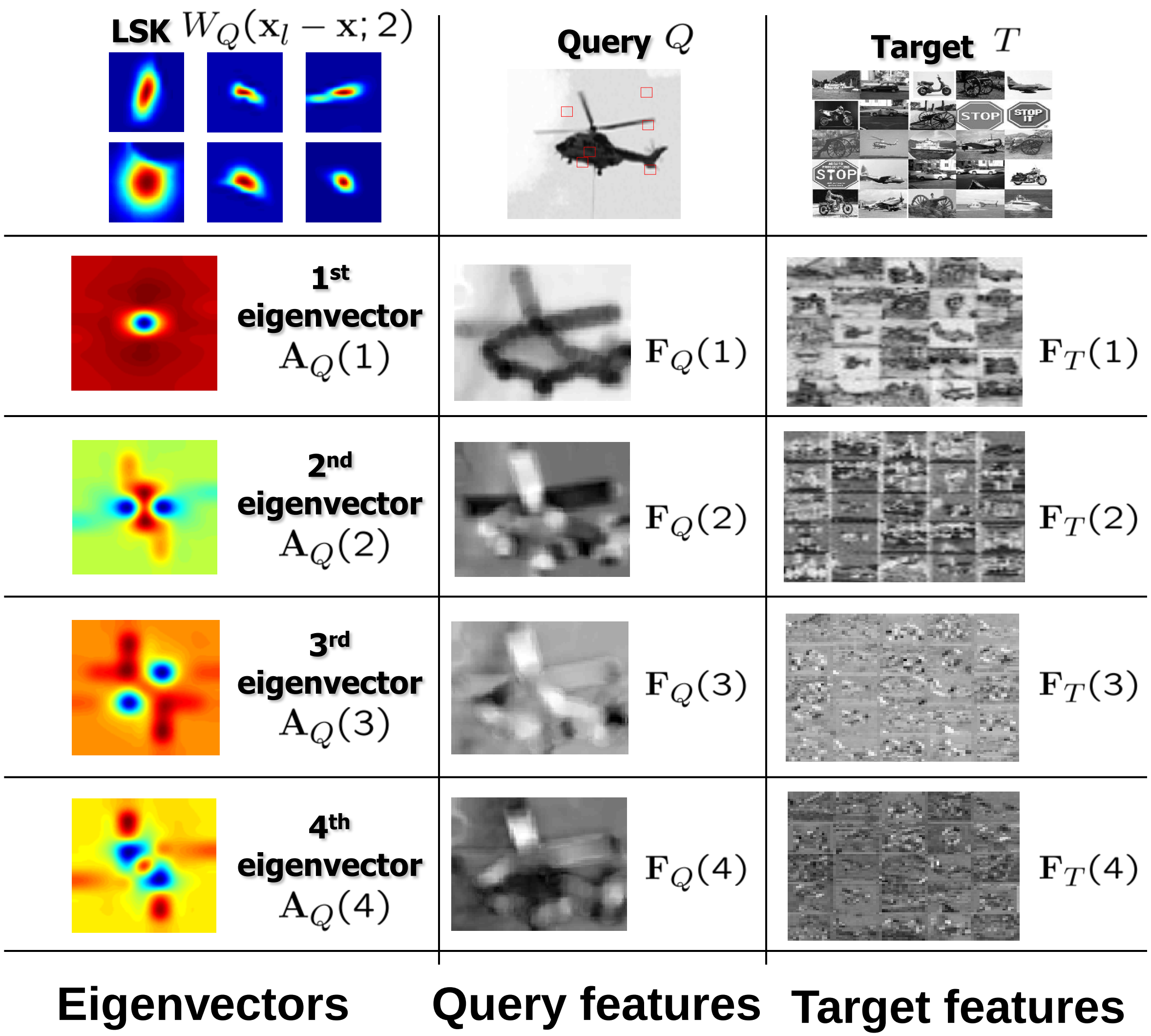
$$\mathbf{F}_Q = [\underline{\mathbf{f}}_Q^1, \cdots, \underline{\mathbf{f}}_Q^n] = \mathbf{A}_Q^{'} \mathbf{W}_Q$$

$$\mathbf{F}_T = [\underline{\mathbf{f}}_T^1, \cdots, \underline{\mathbf{f}}_T^{n_T}] = \mathbf{A}_Q^{'} \mathbf{W}_T$$



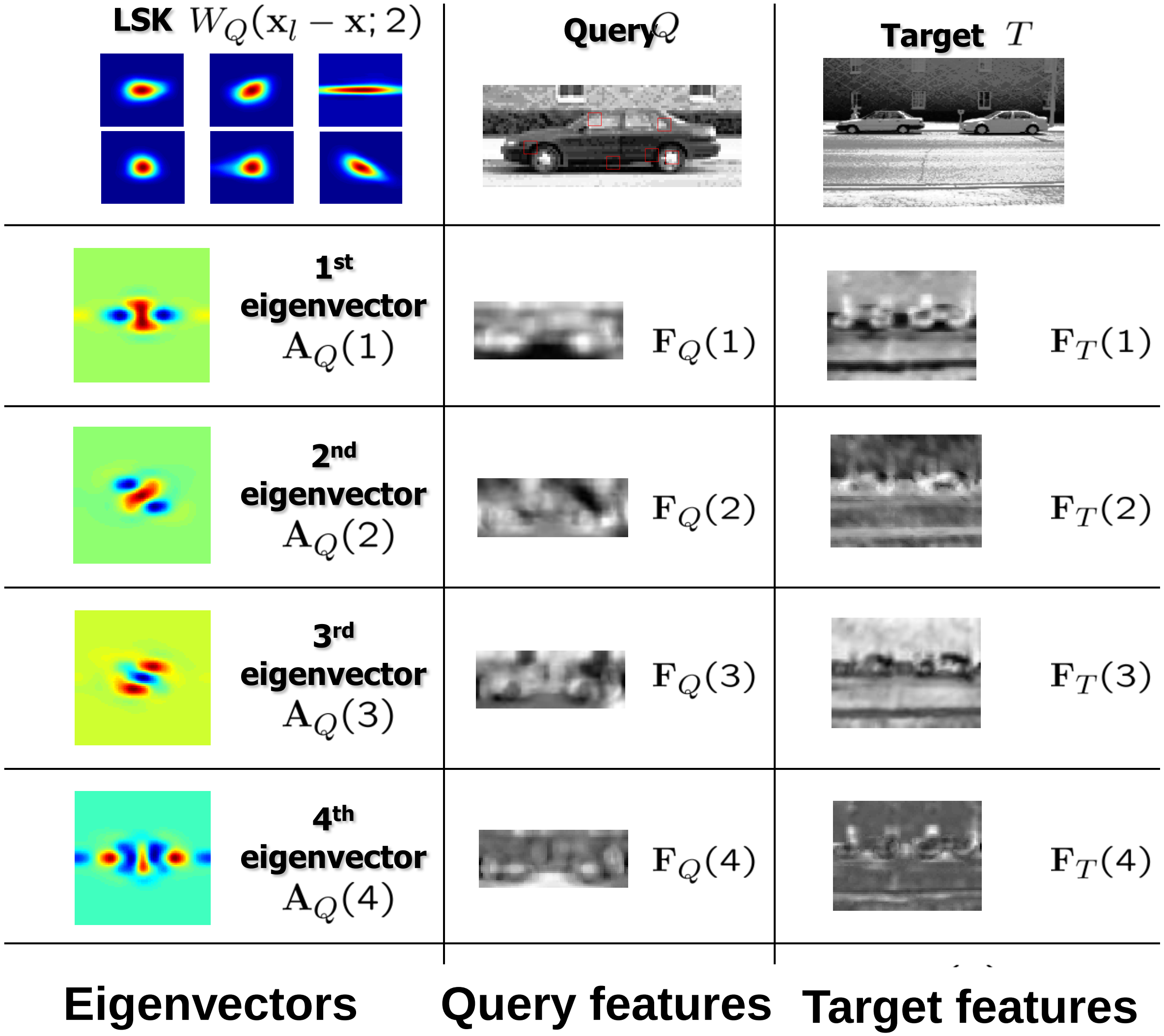
Stage 2: Salient features after PCA

Object: Helicopter

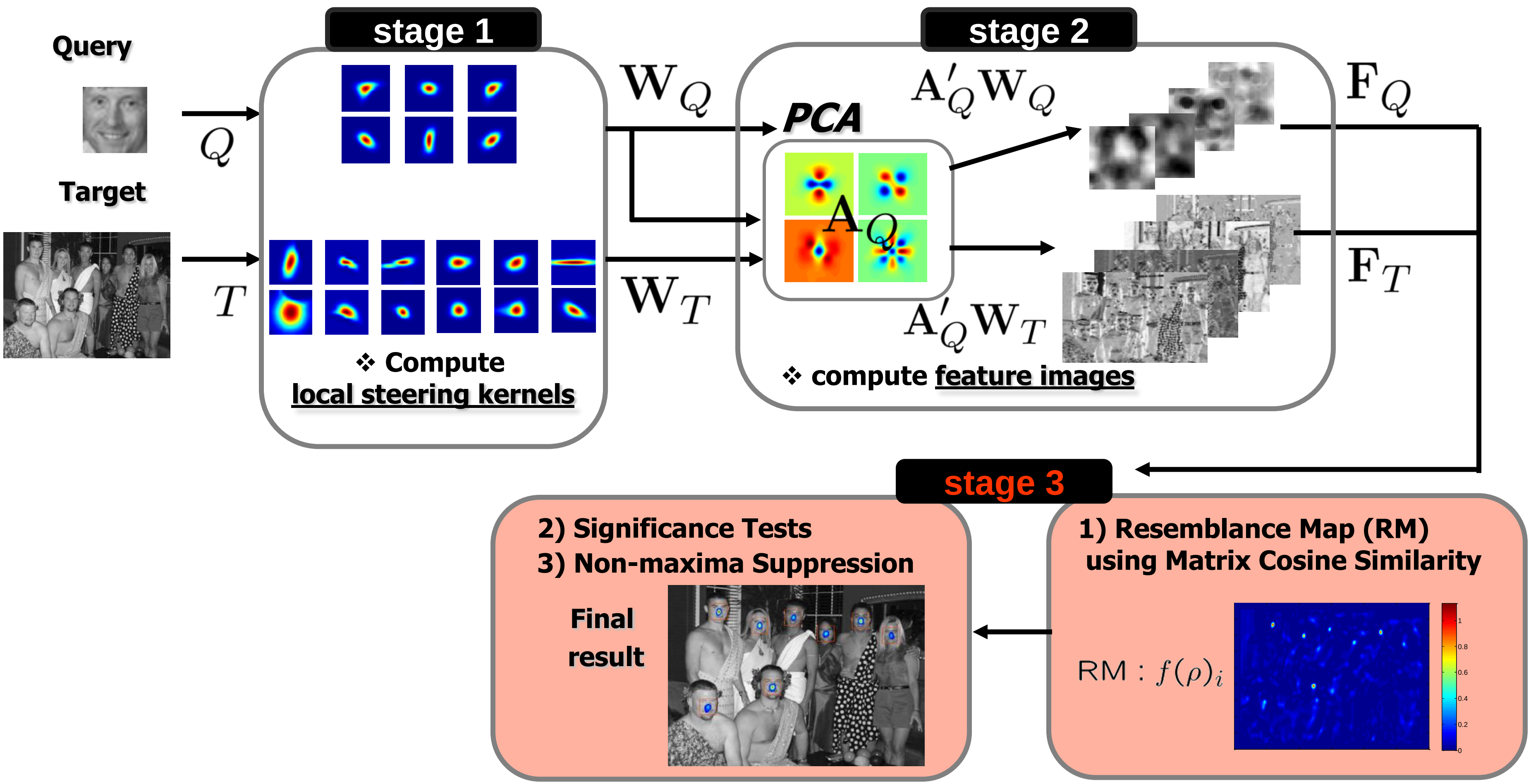


Stage 2: Salient features after PCA

Object: Car

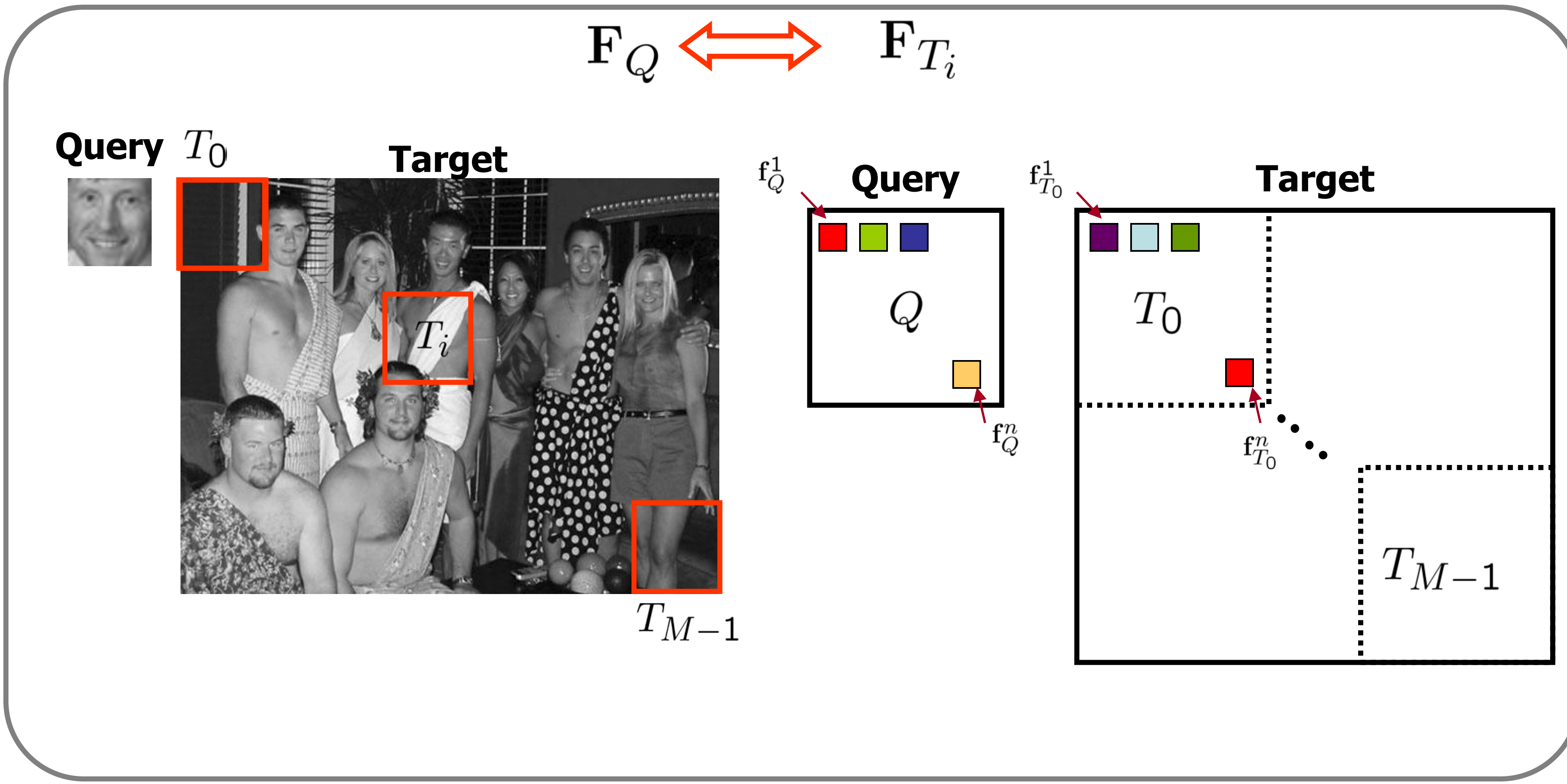


System Overview : Stage 3



Stage 3: Finding similarity between features

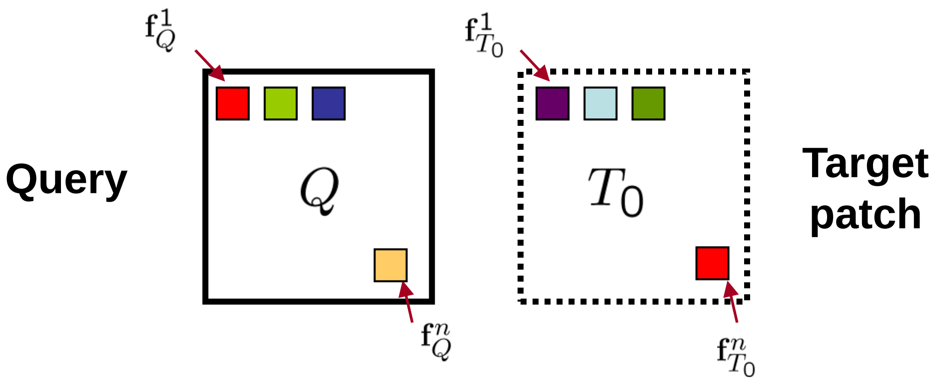
Target image is divided into a set of overlapping patches



Stage 3: Correlation based Metric

The vector cosine similarity

$$\rho(\mathbf{a},\mathbf{b}) = < \frac{\mathbf{a}}{\|\mathbf{a}\|}, \frac{\mathbf{b}}{\|\mathbf{b}\|} > = \frac{\mathbf{a}'\mathbf{b}}{\|\mathbf{a}\|\|\mathbf{b}\|} = \cos\theta \in [-1,1],$$



Inner product between two normalized vectors

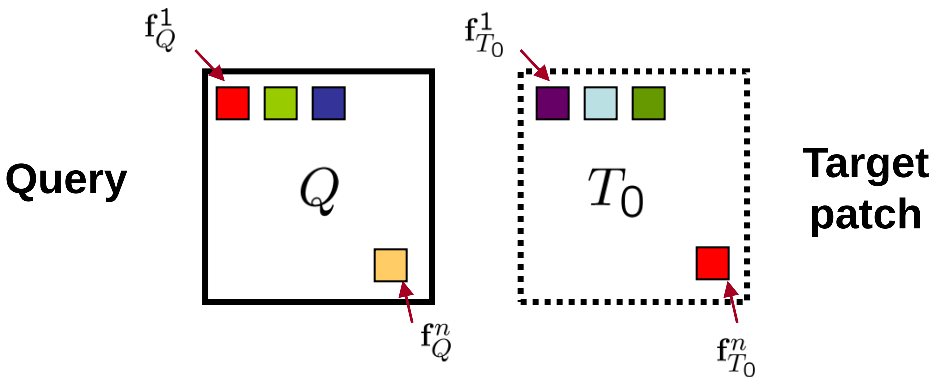
Measures angle while discarding the magnitude

Stage 3: Correlation based Metric

The vector cosine similarity

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

$$\mathbf{f}_Q, \mathbf{f}_{T_i} \in \mathbb{R}^d$$



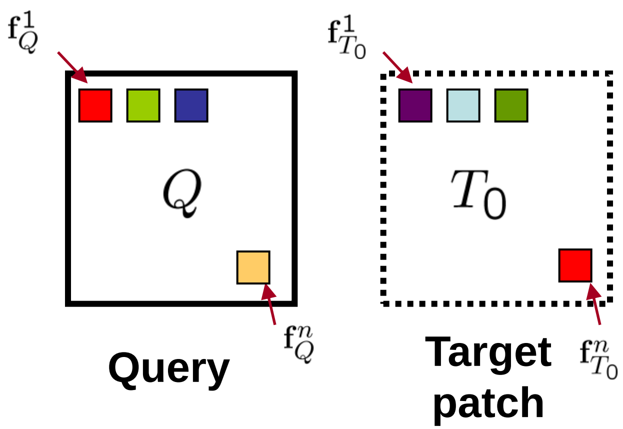
Inner product between two normalized vectors

Measures angle while discarding the magnitude

Stage 3: Matrix Cosine Similarity

What about a set of vectors? **Matrix Cosine Similarity**

→ Frobenius Inner product between normalized matrices

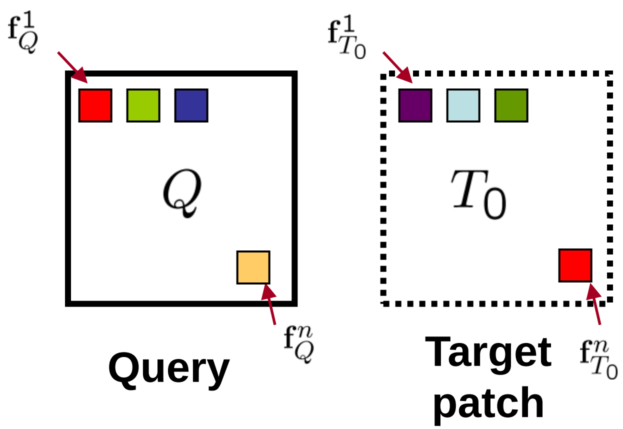


$$
\begin{aligned}
\rho(\mathbf{A},\mathbf{B}) &= <\overline{\mathbf{A}},\overline{\mathbf{B}}>_F = \operatorname{trace}\Big(\frac{\mathbf{A}'\mathbf{B}}{\|\mathbf{A}\|_F\|\mathbf{B}\|_F}\Big) \in [-1,1], \\
&= \sum_{\ell=1}^{n} \frac{\mathbf{a}^{\ell\prime}\mathbf{b}^{\ell}}{\|\mathbf{A}\|_F\|\mathbf{B}\|_F}, \\
&= \sum_{\ell=1}^{n} \rho(\mathbf{a}^{\ell},\mathbf{b}^{\ell})\frac{\|\mathbf{a}^{\ell}\|\|\mathbf{b}^{\ell}\|}{\|\mathbf{A}\|_F\|\mathbf{B}\|_F}.
\end{aligned}
$$

Stage 3: Matrix Cosine Similarity

What about a set of vectors? **Matrix Cosine Similarity**

→ Frobenius Inner product between normalized matrices



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

A weighted sum of the column-wise vector cosine similarities

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

We can prove optimality of this approach in a naïve Bayes sense.

Stage 3: Generate Resemblance Map

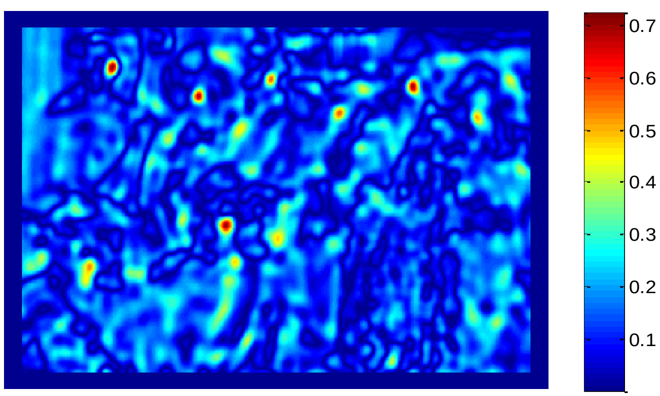
Resemblance Map (RM)

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

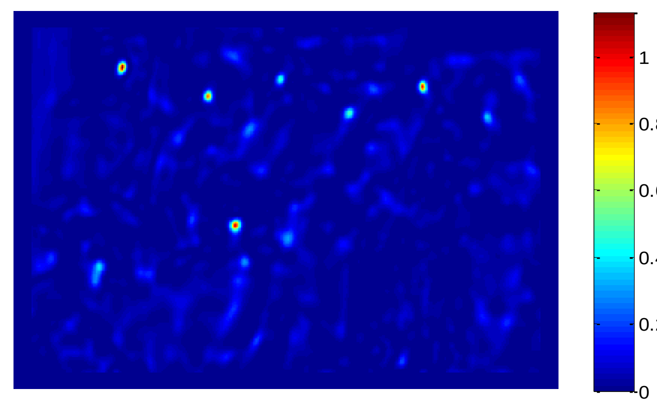
Describes the proportion of variance in common between two features

Lawley-Hotelling Trace statistic

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



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



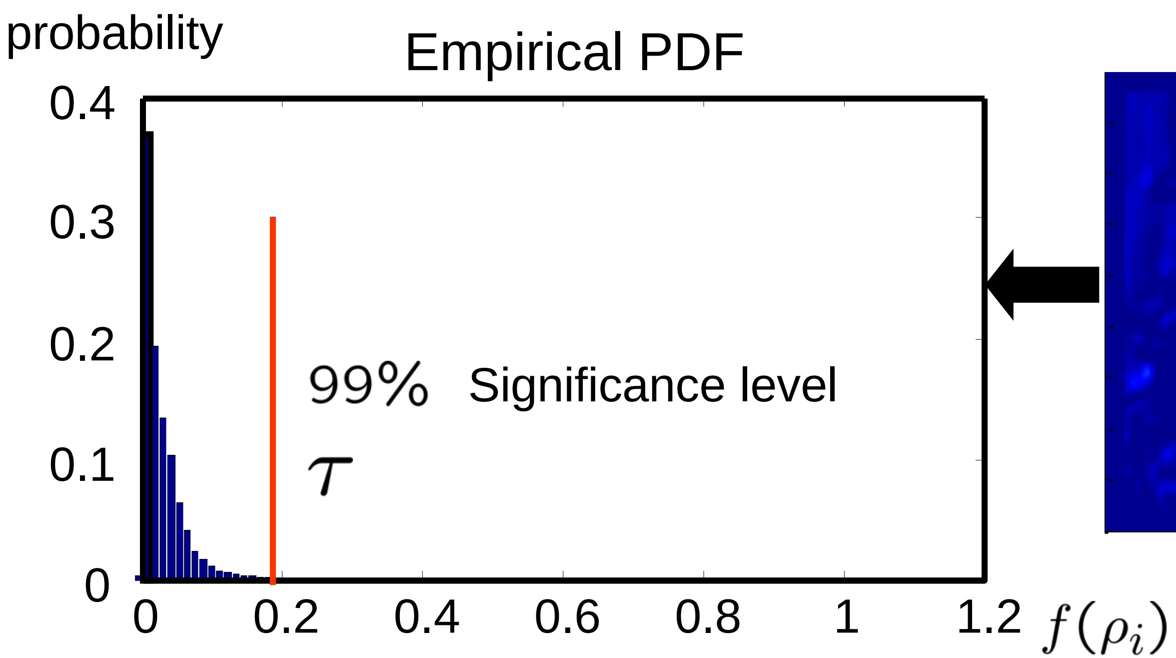
Stage 3: Non-parametric Significance Tests

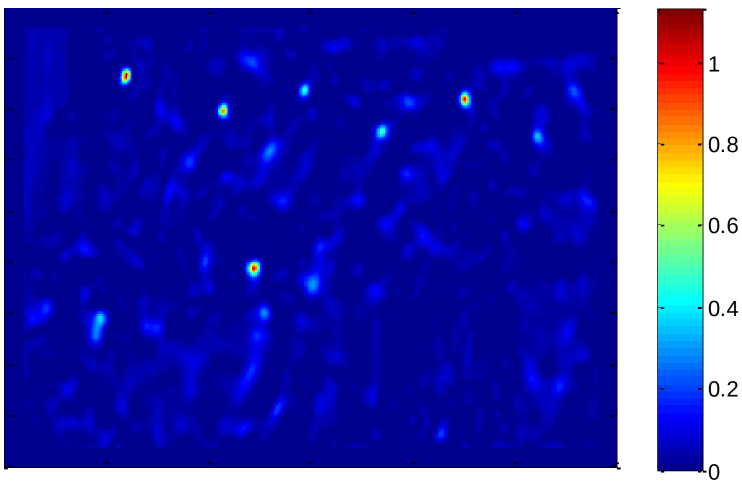
1. Is any sufficiently similar object present?

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

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

2. How many objects of interest are present?





Milanfar et al. EE Dept, UCSC

Experimental Results

Query **Targets**

Dataset from Weizmann Inst.

Experimental Results

query

target

query

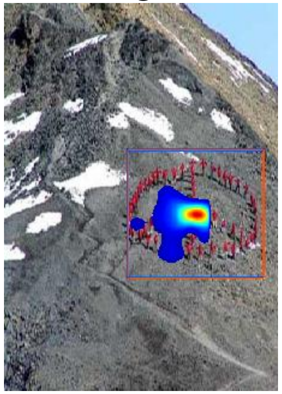
target

Experimental Results

query

target

target

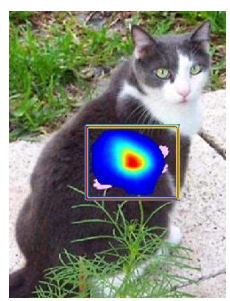


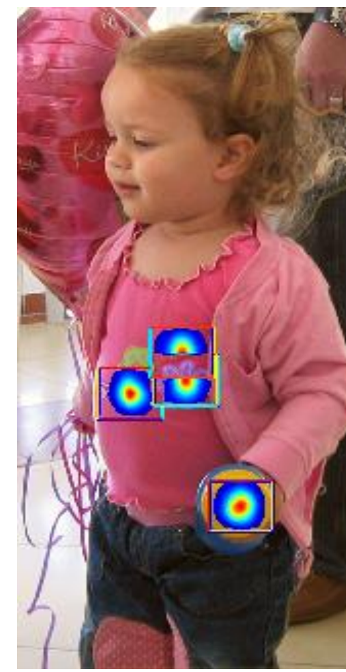
Experimental Results

query

target

target



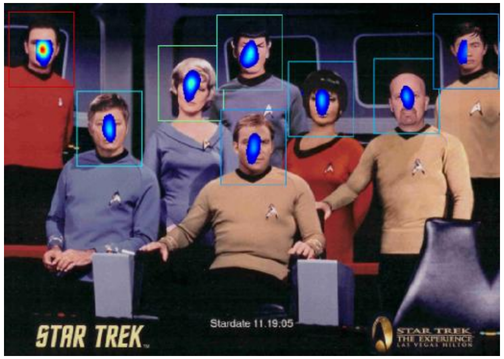


target

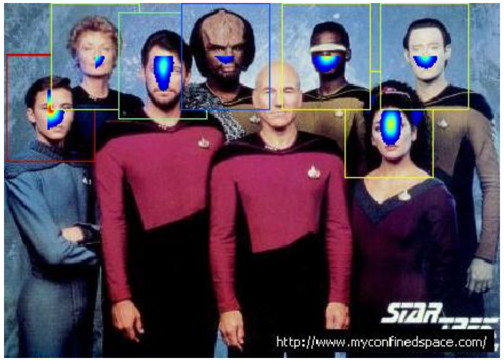
Experimental Results

query

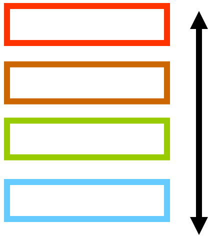
target



target

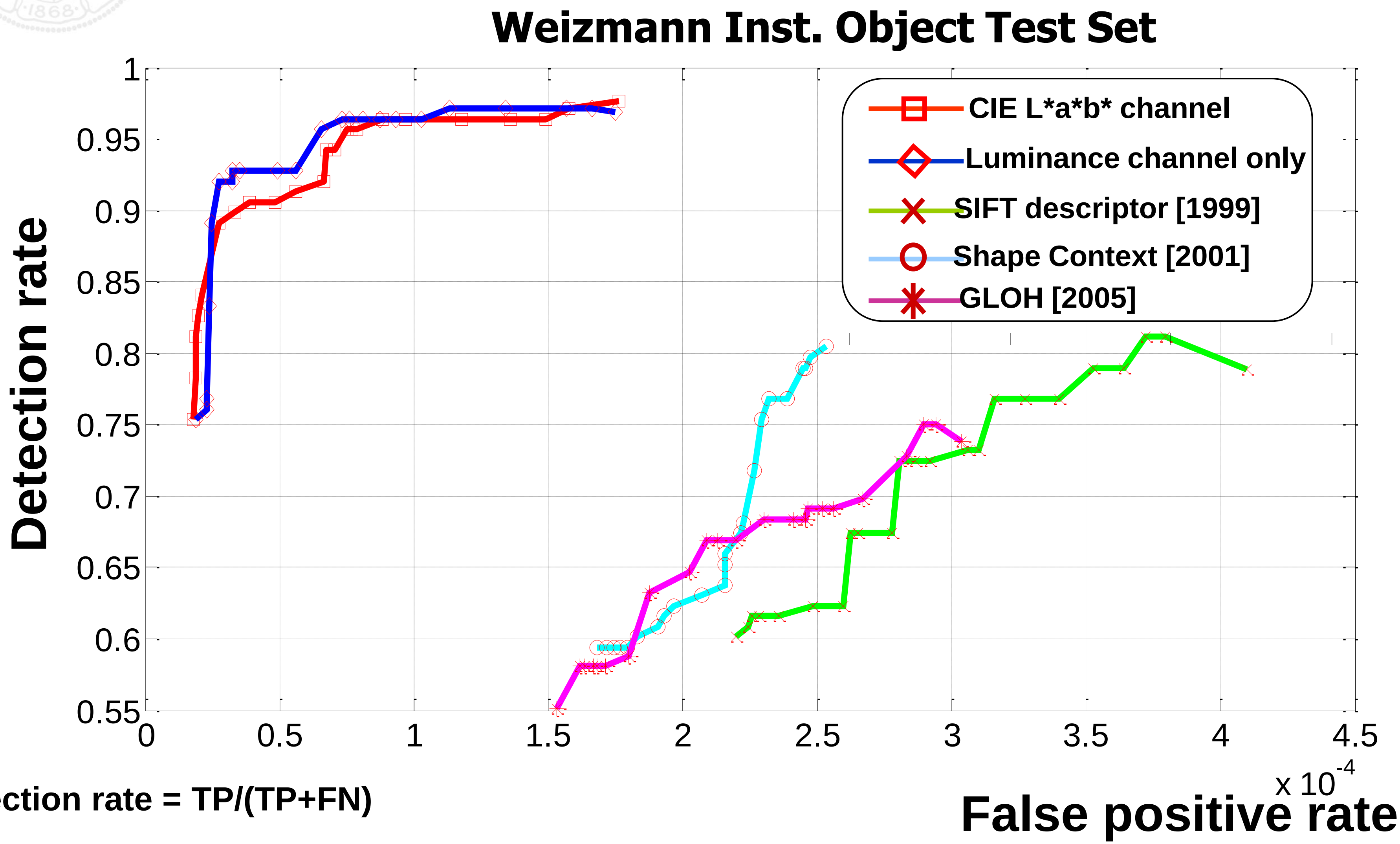


Higher resemblance



Lower resemblance

Experimental Results



Detection rate = TP/(TP+FN)

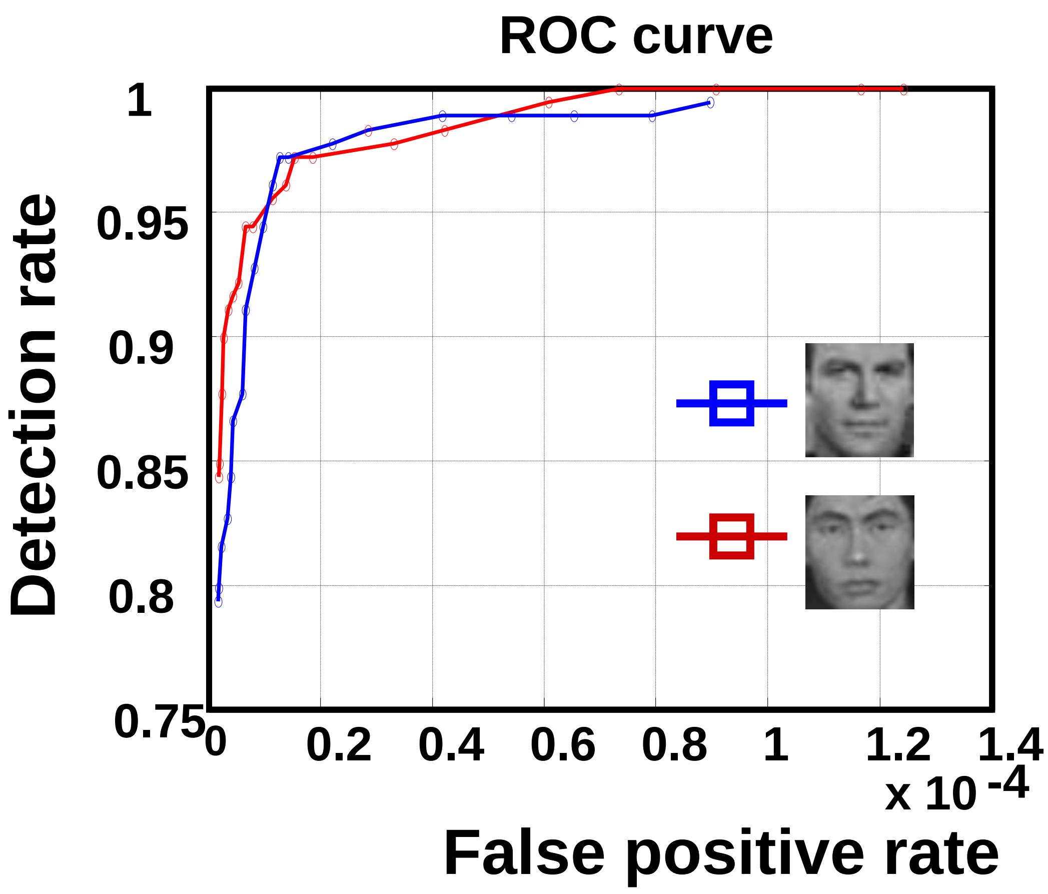
False positive rate = FP/(FP+TN)

Experimental Results

The MIT-CMU Face Test Set

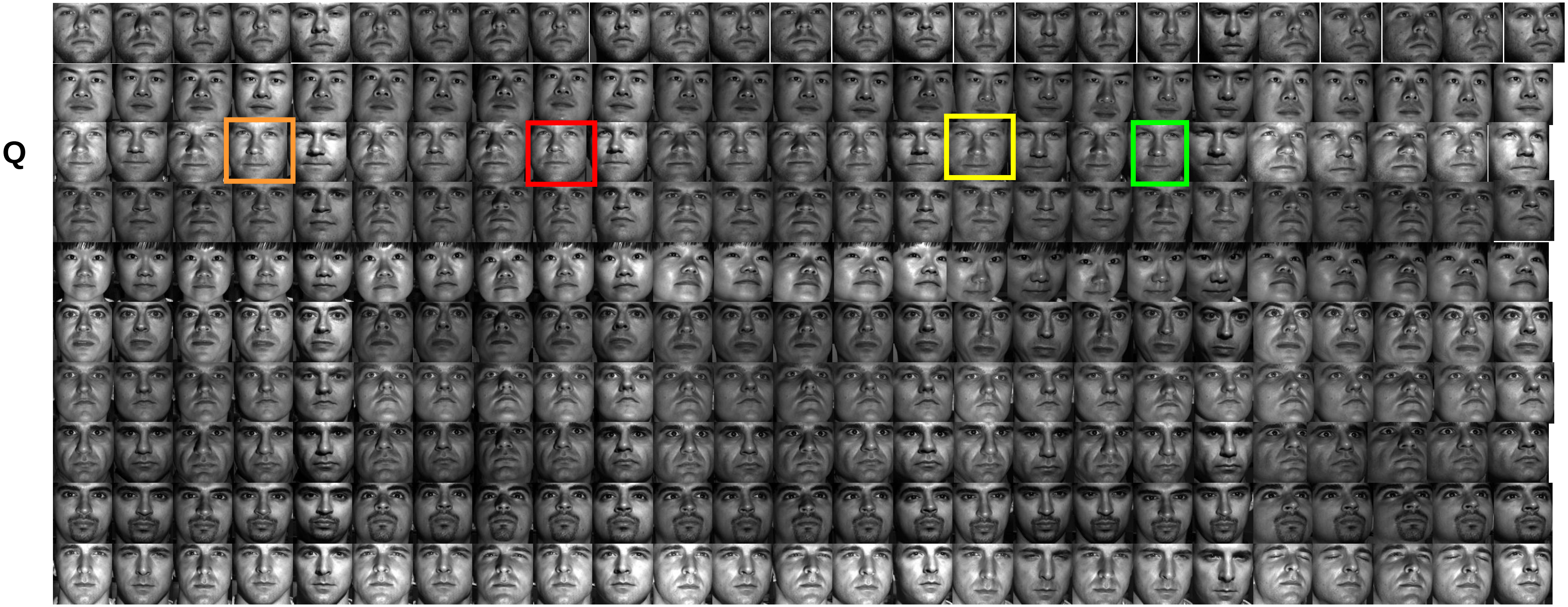
Experimental Results

The MIT-CMU Face Test Set



Gallery Set:10 subjects x 25 different conditions

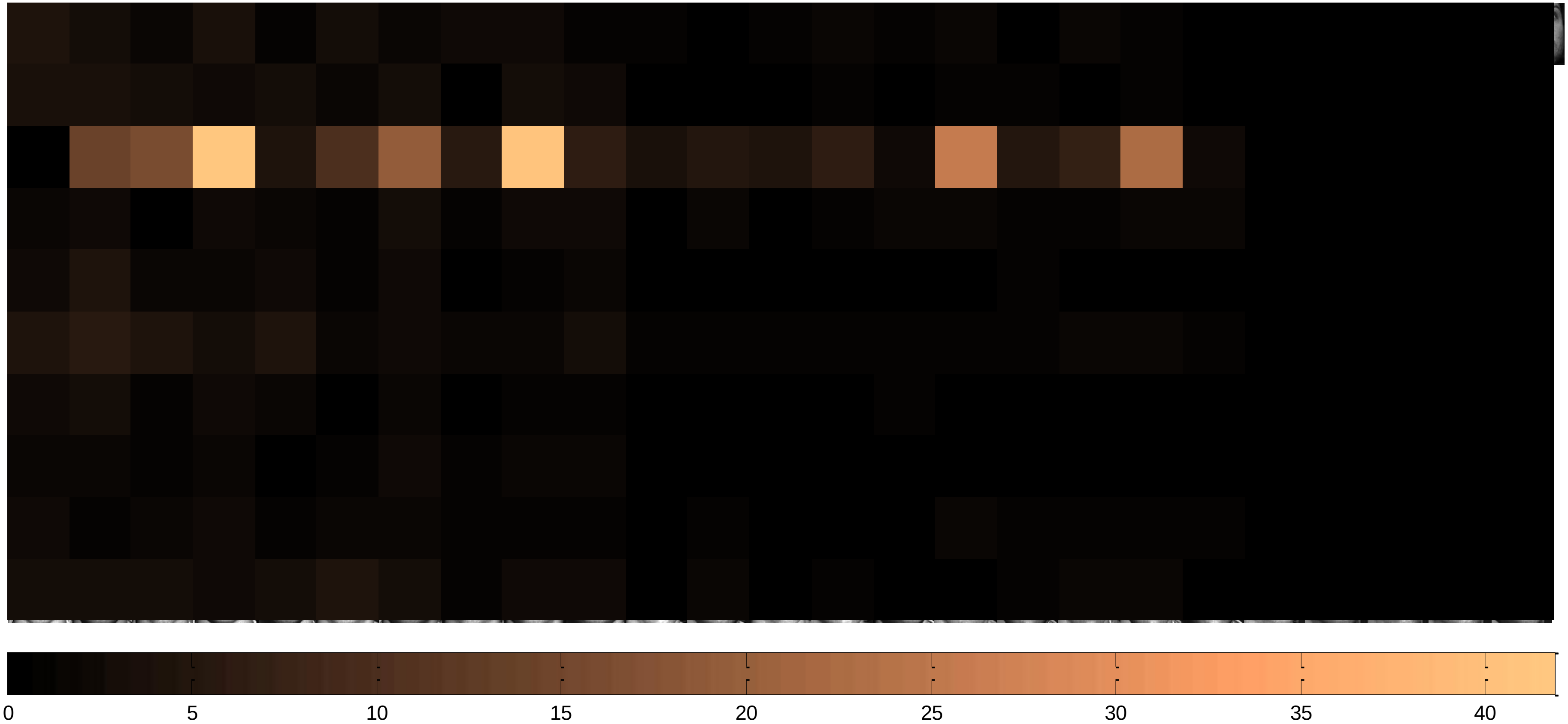
Query



Gallery Set:10 subjects x 25 different conditions

Query

Q

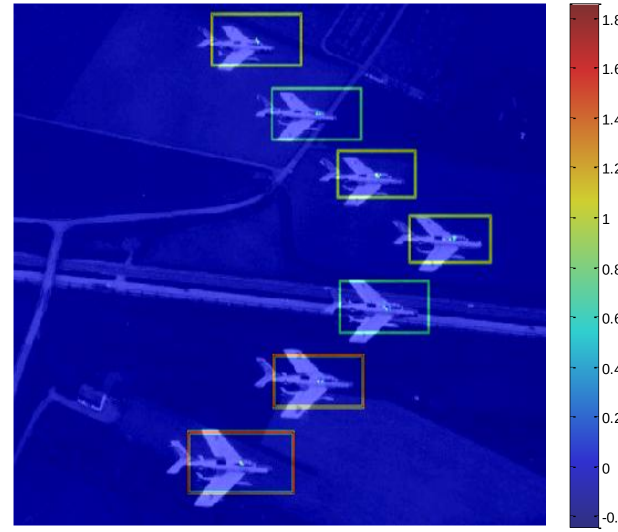


Milanfar et al. EE Dept, UCSC

query

target

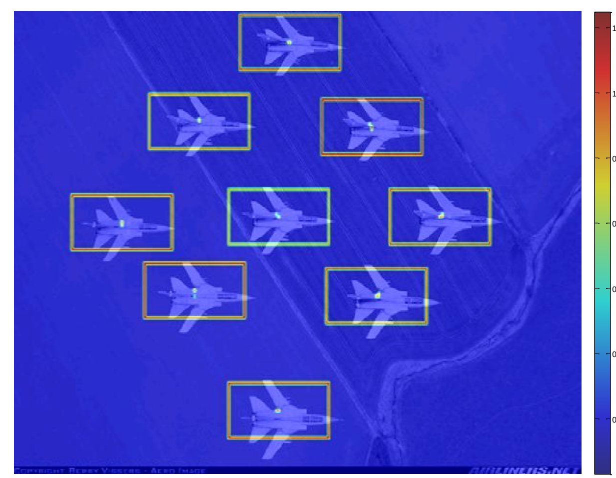
output



query

target

output



Milanfar et al. EE Dept, UCSC

query

target

output

---

query

target

output

Outline

I. Motivation

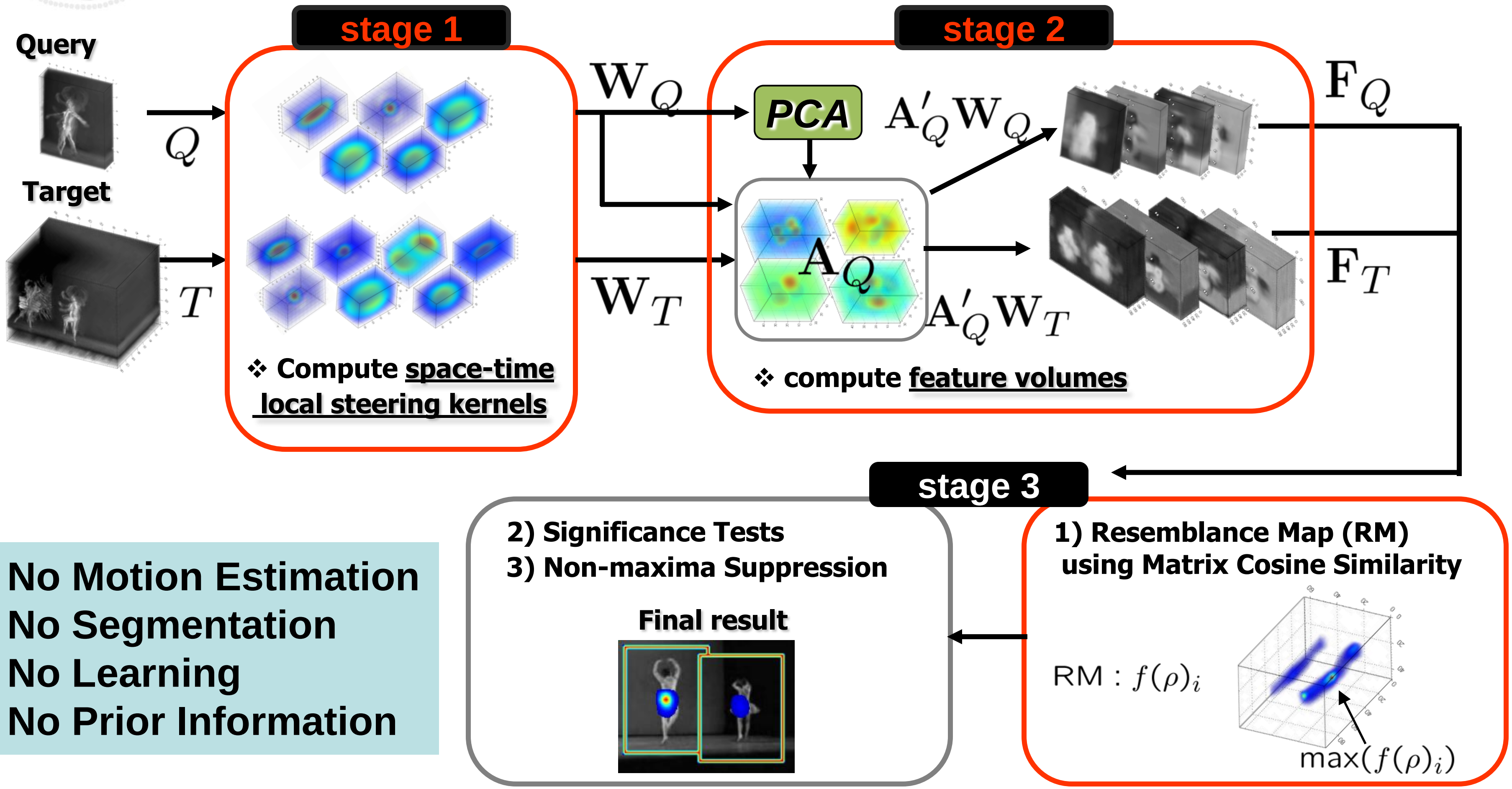
II. System Overview

III. Object Detection

IV. Action Detection

V. Conclusion and Future work

Action Detection System Overview



- No Motion Estimation
- No Segmentation
- No Learning
- No Prior Information

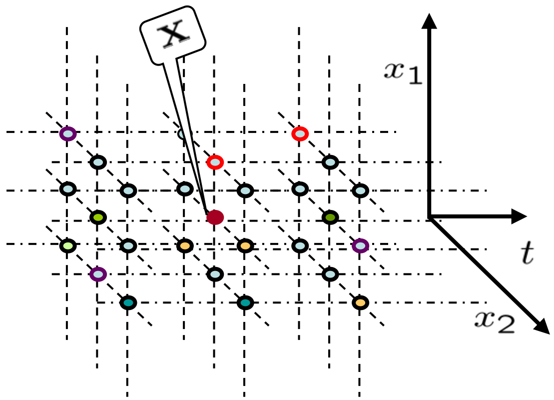
H. Seo and P. Milanfar, "**Generic Action Recognition from a Single Example**", Submitted to *International Journal of Computer Vision* (IJCV), March 2009

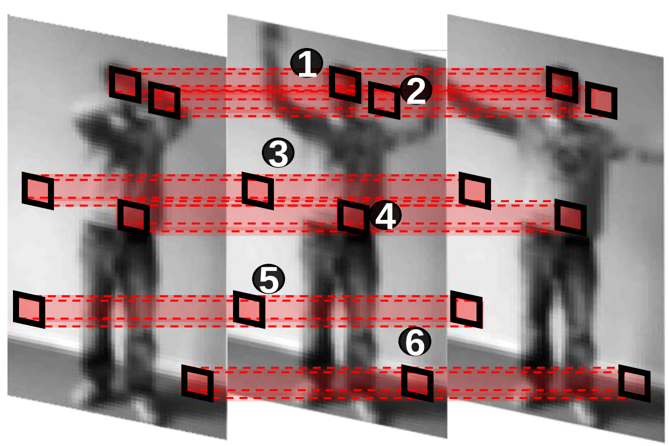
Stage 1: Space –Time Descriptors

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

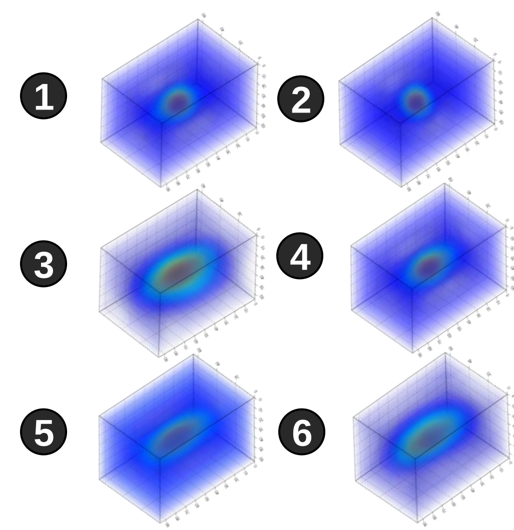
$\mathbf{C}_l$: 3x3 local covariance matrix

$\mathbf{x}$: space-time coordinates $[x_1, x_2, t]$





First frame   Key frame   Last frame



Experimental Results

Shechtman's action test set (Beach walk)

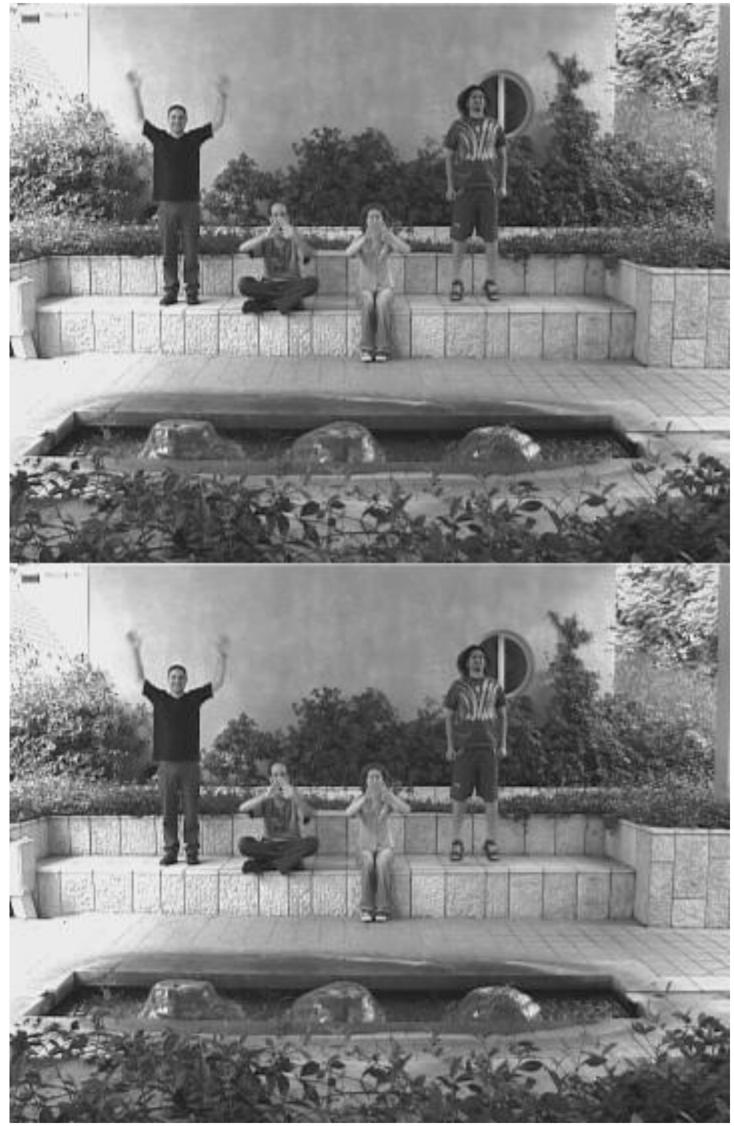
Query

Typical run time for target (50 frames of 144 x 192) and query (13 frames of 90 x 110) : a little over 1 minute

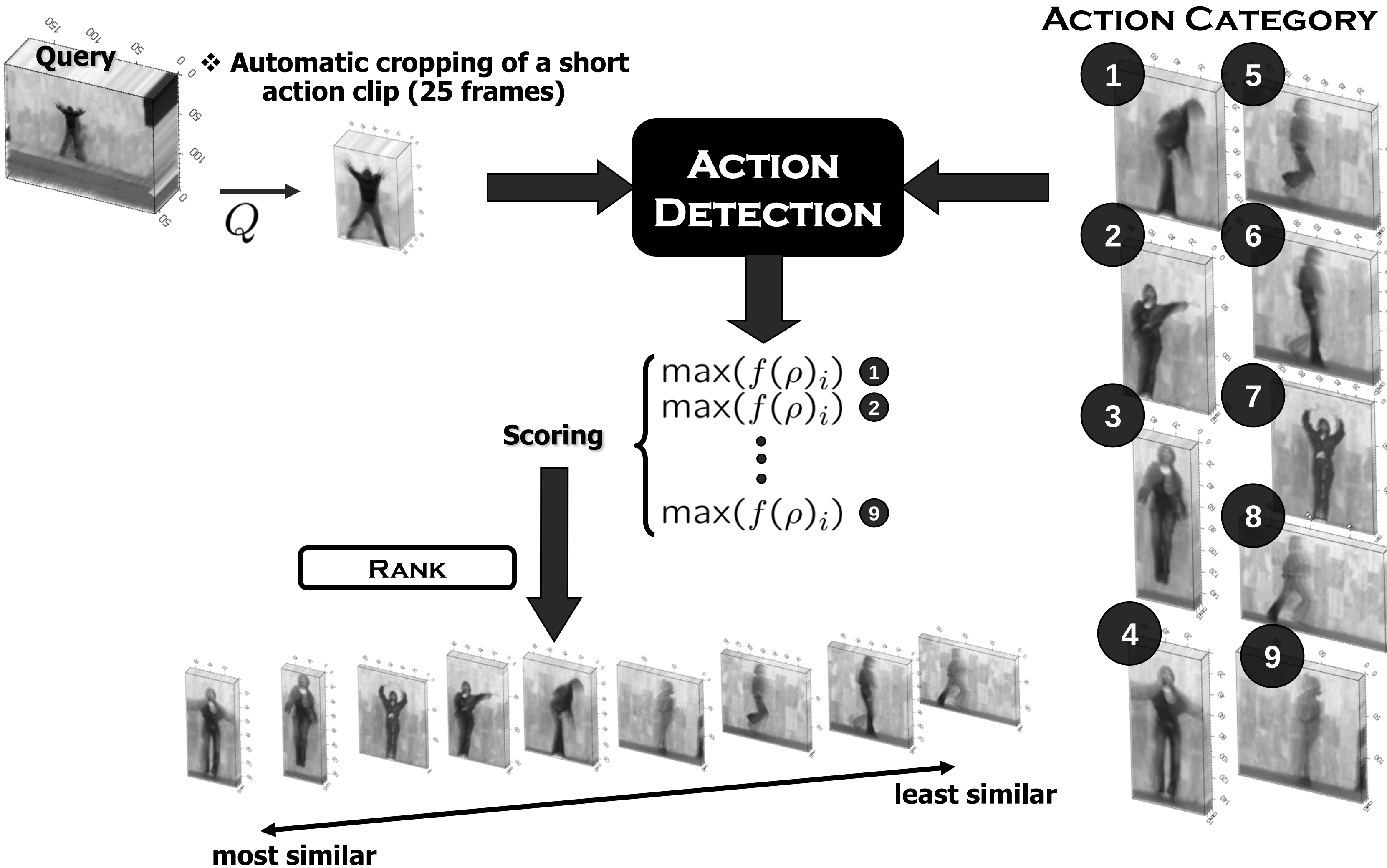
Experimental Results

(Multiple Actions)

Multiple queries
Automatic cropping



Action Recognition



Action Classification Performance

Average confusion matrices

Classification rate: 96 %

	Bend	Jack	Jump	Pjump	Run	Side	Skip	Walk	Wave1	Wave2
Bend	1.0	.00	.00	.00	.00	.00	.00	.00	.00	.00
Jack	.00	1.0	.00	.00	.00	.00	.00	.00	.00	.00
Jump	.00	.00	1.0	.00	.00	.00	.00	.00	.00	.00
Pjump	.00	.00	.00	1.0	.00	.00	.00	.00	.00	.00
Run	.00	.00	.00	.00	1.0	.00	.00	.00	.00	.00
Side	.00	.00	.00	.00	.00	1.0	.00	.00	.00	.00
Skip	.00	.00	.11	.00	.11	.11	.67	.00	.00	.00
Walk	.00	.00	.00	.00	.00	.00	.00	1.0	.00	.00
Wave1	.11	.00	.00	.00	.00	.00	.00	.00	.89	.00
Wave2	.00	.00	.00	.00	.00	.00	.00	.00	.00	1.0

(Weizmann dataset) **90 video sequences**

Classification rate: 95.66 %

	box	hclp	hwav	jog	run	walk
box	.94	.02	.04	.00	.00	.00
hclp	.00	.99	.01	.00	.00	.00
hwav	.00	.00	1.0	.00	.00	.00
jog	.00	.00	.00	.95	.01	.04
run	.00	.00	.00	.05	.92	.03
walk	.00	.00	.00	.05	.01	.94

(KTH dataset) **600 video sequences**

Classification rate = 1 – (# of miss classification) / (total # of sequences)

Evaluation setting: Leave-one-out

Classify each testing video as one of the predefined classes by 3-NN (nearest neighbor)

Action Classification Performance

- **Comparison with state-of-the art methods** (KTH dataset)

Our Approach (1-NN)	89%
Our Approach (2-NN)	93%
Our Approach (3-NN)	95.66%

Our Approach (3-NN)	95.66%
Kim et al. (2008)	95.33%
Ali et al.(2008)	87.7%
Dollar et al. (2005)	81.17%
Ning et al. (2008)	92.31%
Niebles et al. (2008)	81.5%
Wong et al. (2007)	71.16%

Classification rate = 1 – (# of miss classification) / (total # of sequences)

- **We outperform all the state-of-the art methods on KTH dataset.**

Publications

- H. Seo and P. Milanfar, "**Training-free, Generic Object Detection using Locally Adaptive Regression Kernels**", Accepted for publication in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2008
- H. Seo and P. Milanfar, "**Generic Action Recognition from a Single Example**", Submitted to *International Journal of Computer Vision* (IJCV), March 2009
- H. Seo and P. Milanfar, "**Static and Space-time Visual Saliency Detection by Self-Resemblance** ", Submitted to *Journal of Vision* (JoV), May 2009
- H. Seo and P. Milanfar, "**Detection of Human Actions from a Single Example**", Accepted for publication in *International Conference on Computer Vision* (ICCV), March 2009
- H. Seo and P. Milanfar, "**Nonparametric Bottom-Up Saliency Detection by Self-Resemblance**", Accepted for *IEEE Conference on Computer Vision and Pattern Recognition* (CVPR), 1st International Workshop on Visual Scene Understanding (ViSU'09), Miami, June, 2009
- H. Seo and P. Milanfar, **"Using Local Regression Kernels for Statistical Object Detection**", Proceedings of *IEEE International Conference on Image Processing* (ICIP), San Diego, 2008

Conclusions & Future Work

- **Local Steering Kernels are Very Effective Descriptors**
- **Simple Approach: PCA + Matrix Cosine Similarity**
- **Excellent Detection and Recognition is Achieved without Training**
- **Make algorithm scalable for image and (video) retrieval**
- **Increase accuracy by incorporating "context"**
- **Detect /recognize objects of interest in general degraded data without explicit restoration**