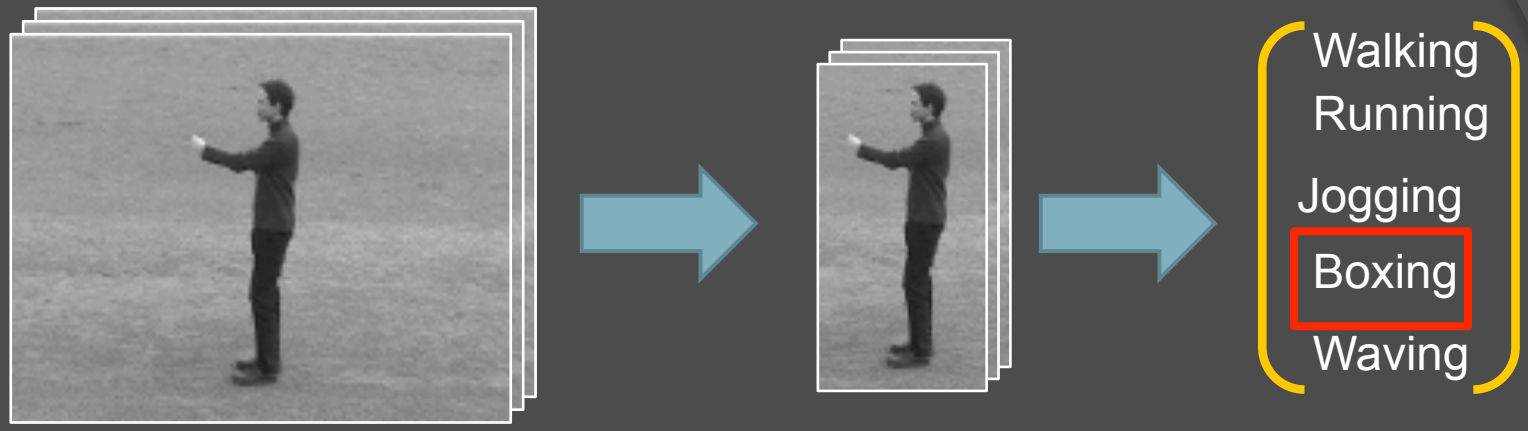
Learning Transferable Distance Functions For Human Action Recognition and Detection

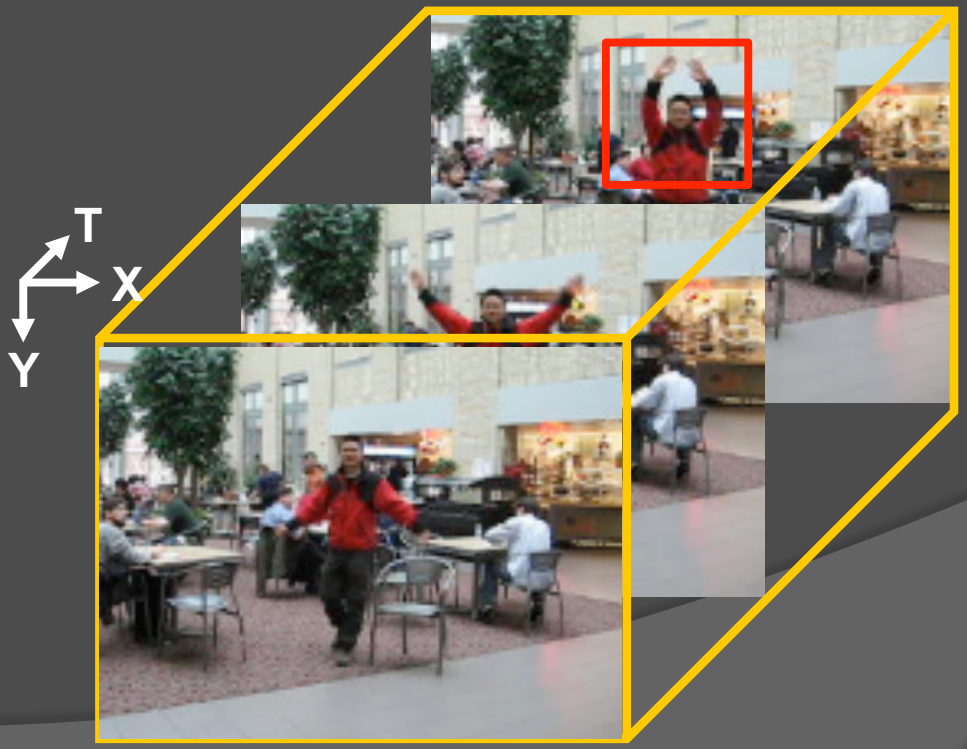
Weilong Yang
Simon Fraser University

SFU
THINKING OF THE WORLD

SFU Vision and Media Lab

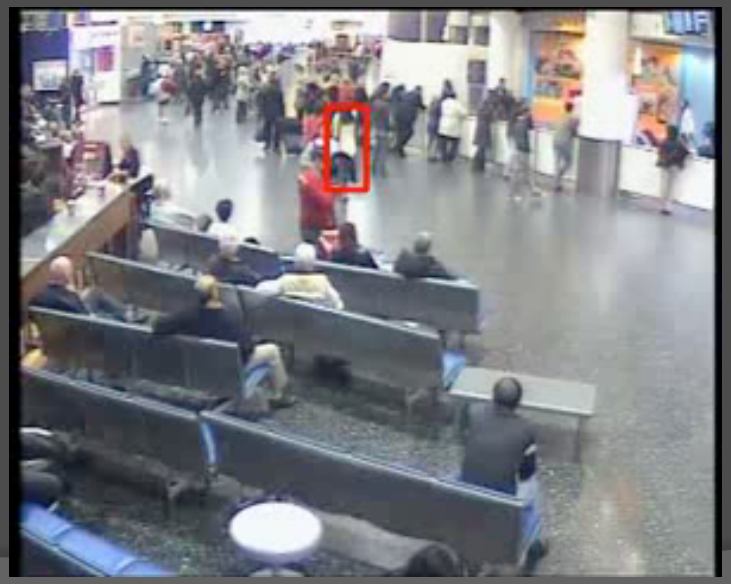
Action Recognition and Detection





Applications

- Action related video search
  - Sports and Dancing video search
- Event Detection
  - Automatic abnormality detection in surveillance videos

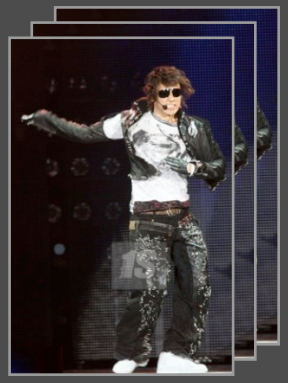


Motivation

- On KTH & Weizmann action datasets, almost 100% accurancy is achieved. [Jhuang et al. ICCV07, Fathi & Mori CVPR08]
- Most of methods rely on a large amout of training set.
  - Half-half split or Leave-one-out cross validation
- It is unrealistic to collect this many training samples for some action.

Query Action

Template A



Template B

Template C

Template D

Label: (Throwing) (R-dancing) (M-dancing) (Kicking)

Template Set

Related Works

- One shot learning of object categories [Fei-Fei et al. ICCV03]
- Visual Object Identification [Ferencz et al. IJCV07]

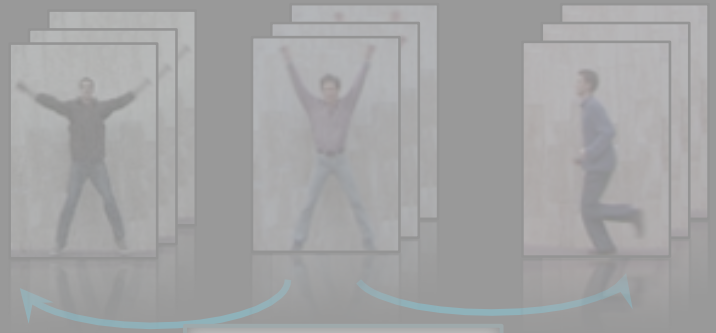
Inter-class Transfer

Transfer Learning:

The ability of a system to recognize and apply knowledge and skills learned in previous tasks to novel tasks .[Pan & Yang, TKDE 2009]

$$w_i = \langle \mathbf{P} \cdot \mathbf{f}_i \rangle$$

$$\langle \mathbf{P}^T \mathbf{F}_q \cdot (\mathbf{d}_{qt_j} - \mathbf{d}_{qt_i}) \rangle > 0$$



Learning

$\mathbf{P} \times$ Hyper-Features Fq

Query

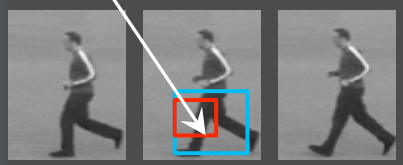
$$D_{qt} = \langle \mathbf{w}_q \cdot \mathbf{d}_{qt} \rangle$$

Templates

A B C

Action Recognition

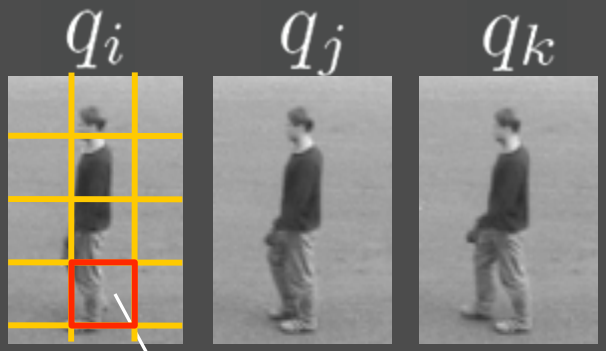
$$D_{qt} = \sum_{s=1}^{M \times S} d_{qt,s}$$



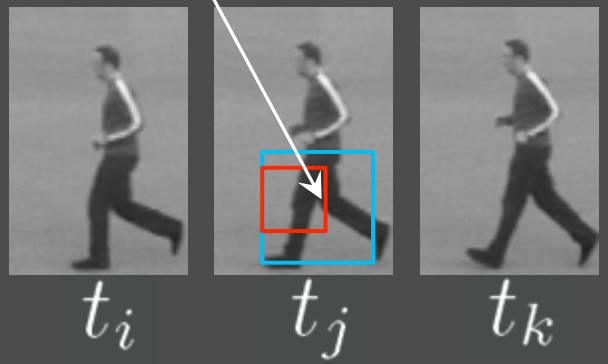
Distance

Action Detection

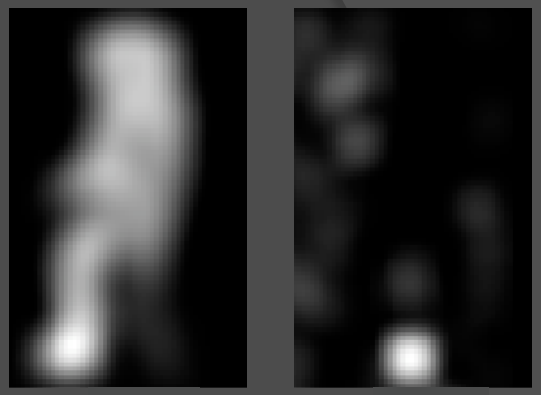
Patch based Action comparison

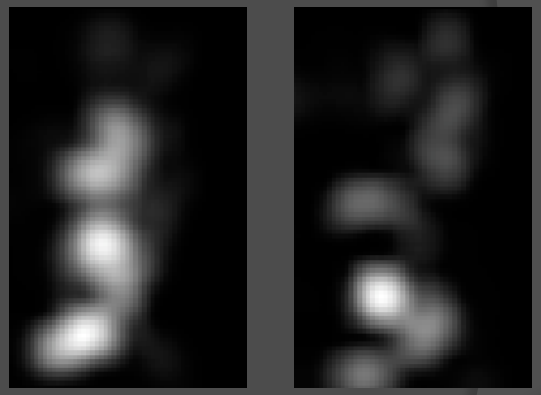


Query



Template



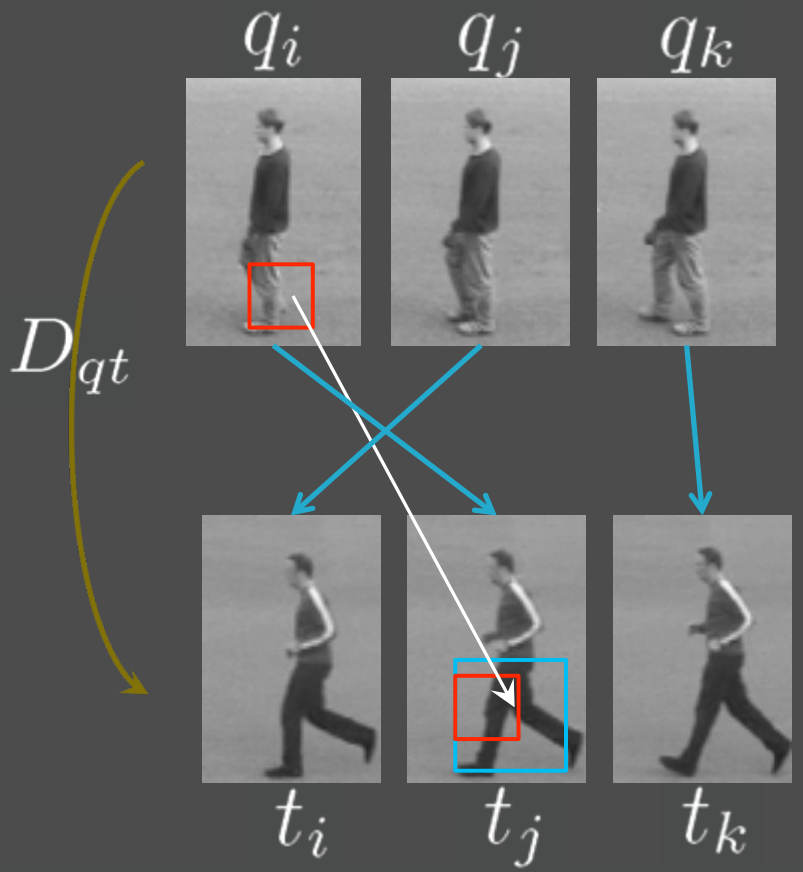


$$D_{qt} = \sum_{i=1}^{M} \min_{j \in [1,N]} \left\{ \sum_{s=1}^{S} \min_{r \in R_s} d(q_{is}, t_{jr}) \right\}$$

Frame-to-Frame Distance

Motion Descriptor
[Efros et al. ICCV 03]

Patch based Action comparison



$$D_{qt} = \sum_{s=1}^{M \times S} d_{qt,s}$$

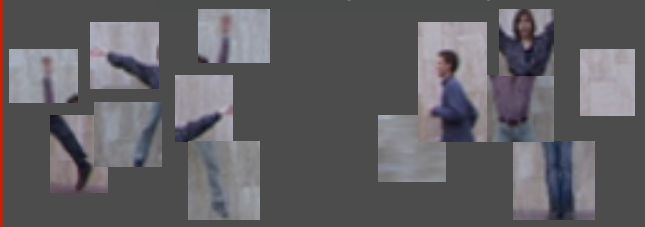
Elementary Patch-to-Patch Distance

$$D_{qt} = \sum_{i=1}^{M} \min_{j \in [1,N]} \left\{ \sum_{s=1}^{S} \min_{r \in R_s} d(q_{is}, t_{jr}) \right\}$$

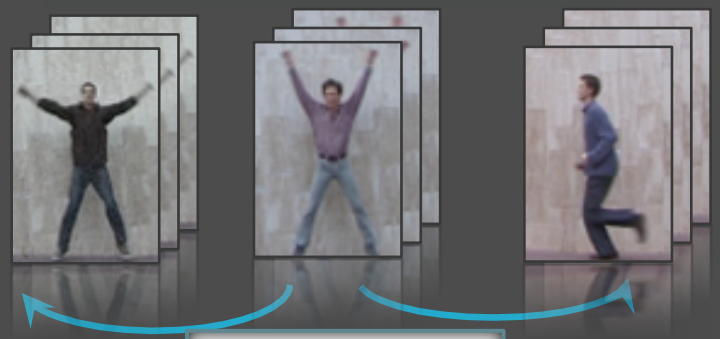
Frame Correspondence

Frame-to-Frame Distance

$$w_i = \langle \mathbf{P} \cdot \mathbf{f}_i \rangle$$



$$\langle \mathbf{P}^T \mathbf{F}_q \cdot (\mathbf{d}_{qt_j} - \mathbf{d}_{qt_i}) \rangle > 0$$



Learning

$\mathbf{P} \times$ Hyper-Features Fq

Query

$$D_{qt} = \langle \mathbf{w}_q \cdot \mathbf{d}_{qt} \rangle$$

Templates

A B C

Action Recognition

$$D_{qt} = \sum_{s=1}^{M \times S} d_{qt,s}$$

Distance

Action Detection

Local Distance Function

$$D_{qt} = \sum_{s=1} d_{qt,s}$$

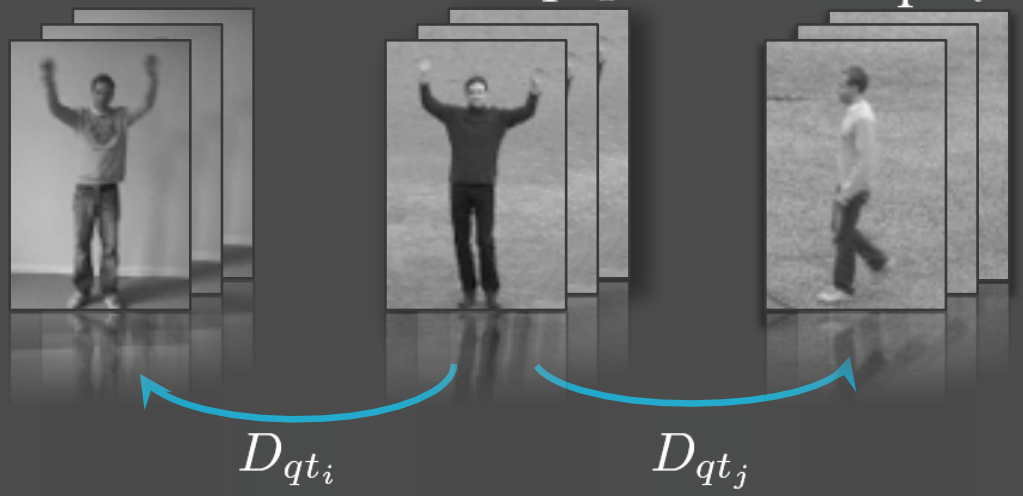
$$D_{qt} = \sum_{s=1} w_{q,s} d_{qt,s} = \langle \mathbf{w}_q \cdot \mathbf{d}_{qt} \rangle$$

[Frome et al. NIPS06]

Local Distance Function

$$D_{qt} = \sum_{s=1} w_{q,s} d_{qt,s} = \langle \mathbf{w}_q \cdot \mathbf{d}_{qt} \rangle$$

Triplet



$$D_{qt_i} < D_{qt_j}$$

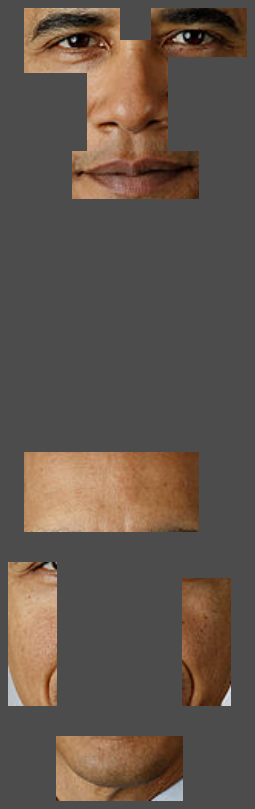
$$\langle \mathbf{w}_q \cdot (\mathbf{d}_{qt_j} - \mathbf{d}_{qt_i}) \rangle > 0$$

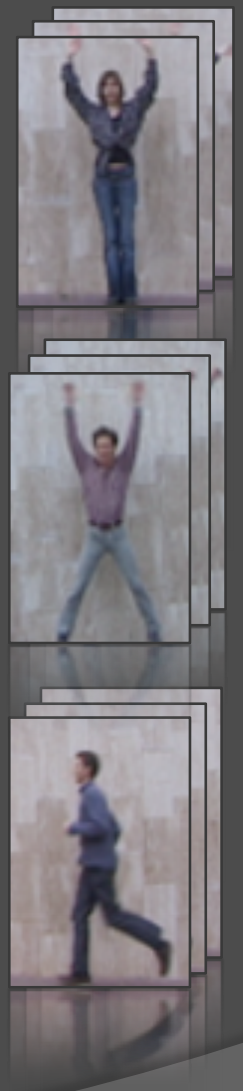
Large Training set required

[Frome et al. NIPS06]

Transferable Distance Function

Transferable Distance Function





Transferable Distance Function

$$w_i = \langle \mathbf{P} \cdot \mathbf{f}_i \rangle$$

Hyper-Feature

Transferable

$$\langle \mathbf{w}_q \cdot (\mathbf{d}_{qt_j} - \mathbf{d}_{qt_i}) \rangle > 0$$

$$\langle \mathbf{P}^T \mathbf{F}_q \cdot (\mathbf{d}_{qt_j} - \mathbf{d}_{qt_i}) \rangle > 0$$

Max-Margin Formulation

$$\min_{\mathbf{P},\xi} \quad \frac{1}{2}\|\mathbf{P}\|^2 + C\sum_{ijk}\xi_{ijk}$$

$$s.t.: \quad \forall i,j,k: \quad \langle \mathbf{P}^T\mathbf{F}_i \cdot \mathbf{x}_{ijk}\rangle \geq 1 - \xi_{ijk}$$

$$\forall m: \quad \langle \mathbf{P} \cdot \mathbf{f}_m\rangle \geq 0$$

$$\forall i,j,k: \quad \xi_{ijk} \geq 0$$

- Triplet $\mathbf{x}_{ijk} = \mathbf{d}_{ij} - \mathbf{d}_{ik}$
- It is convex and similar to the primal problem of SVM

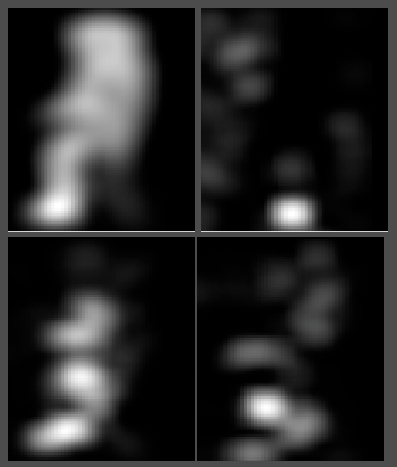
Hyper-Features

- Codebook representation
  - Descriptor for each patch $h = \{g, x, y\}$
    - HOG + Positions
  - Obtaining codebook with the size of $|V|$
    - K-means clustering
  - Hyper-feature for each patch
    - A $|V|$ dimensional vector $\mathbf{f}$

$$\mathbf{f}_m(v_i) = \frac{K_\sigma(D(v_i, h_m))}{\sum_{j=1}^{|V|} K_\sigma(D(v_j, h_m))}$$

Summary of Features

Patch Matching



Motion Cue

Patch Weighting

Hyper-Feature

Shape Cue & Positions

$$w_i = \langle \mathbf{P} \cdot \mathbf{f}_i \rangle$$

$$\langle \mathbf{P}^T \mathbf{F}_q \cdot (\mathbf{d}_{qt_j} - \mathbf{d}_{qt_i}) \rangle > 0$$

Learning

$\mathbf{P} \times$ Hyper-Features Fq

Query

$$D_{qt} = \langle \mathbf{w}_q \cdot \mathbf{d}_{qt} \rangle$$

Templates

A B C

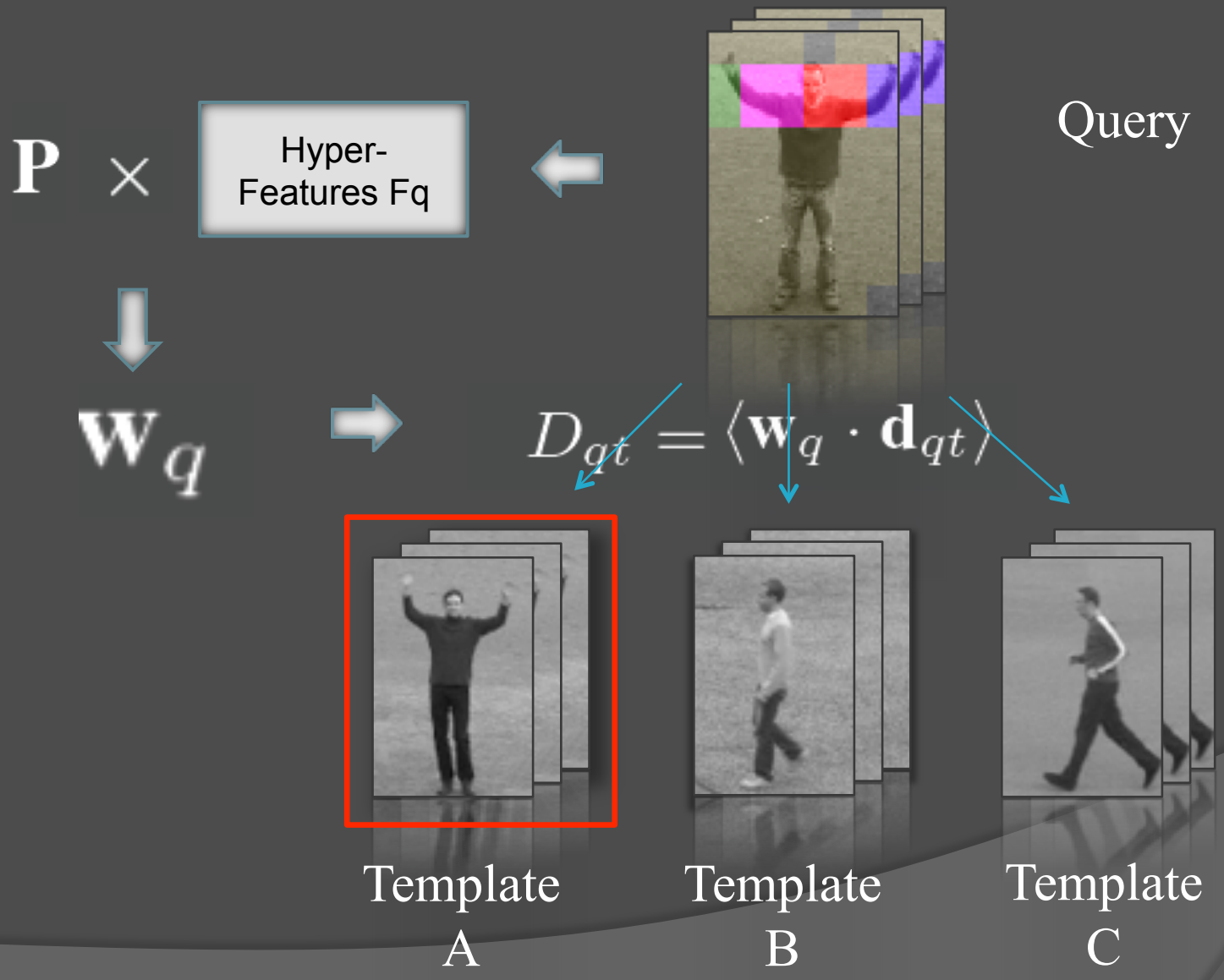
Action Recognition

$$D_{qt} = \sum_{s=1}^{M \times S} d_{qt,s}$$

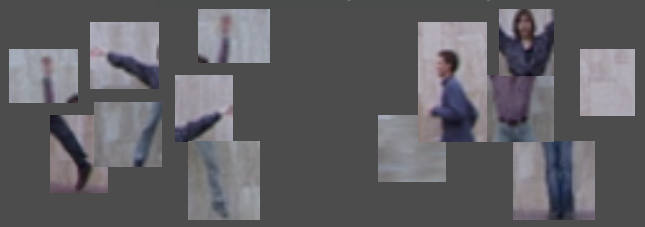
Distance

Action Detection

Recognizing an Action



$w_i = \langle \mathbf{P} \cdot \mathbf{f}_i \rangle$

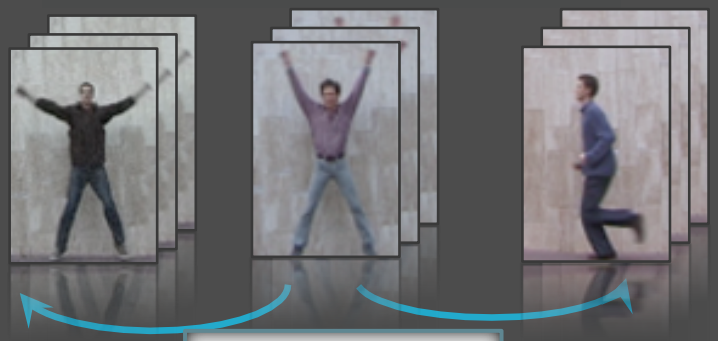


$\mathbf{P} \times$ Hyper-Features Fq

Query

$D_{qt} = \langle \mathbf{w}_q \cdot \mathbf{d}_{qt} \rangle$

$\langle \mathbf{P}^T \mathbf{F}_q \cdot (\mathbf{d}_{qt_j} - \mathbf{d}_{qt_i}) \rangle > 0$



Templates

A

B

C

Learning

Action Recognition

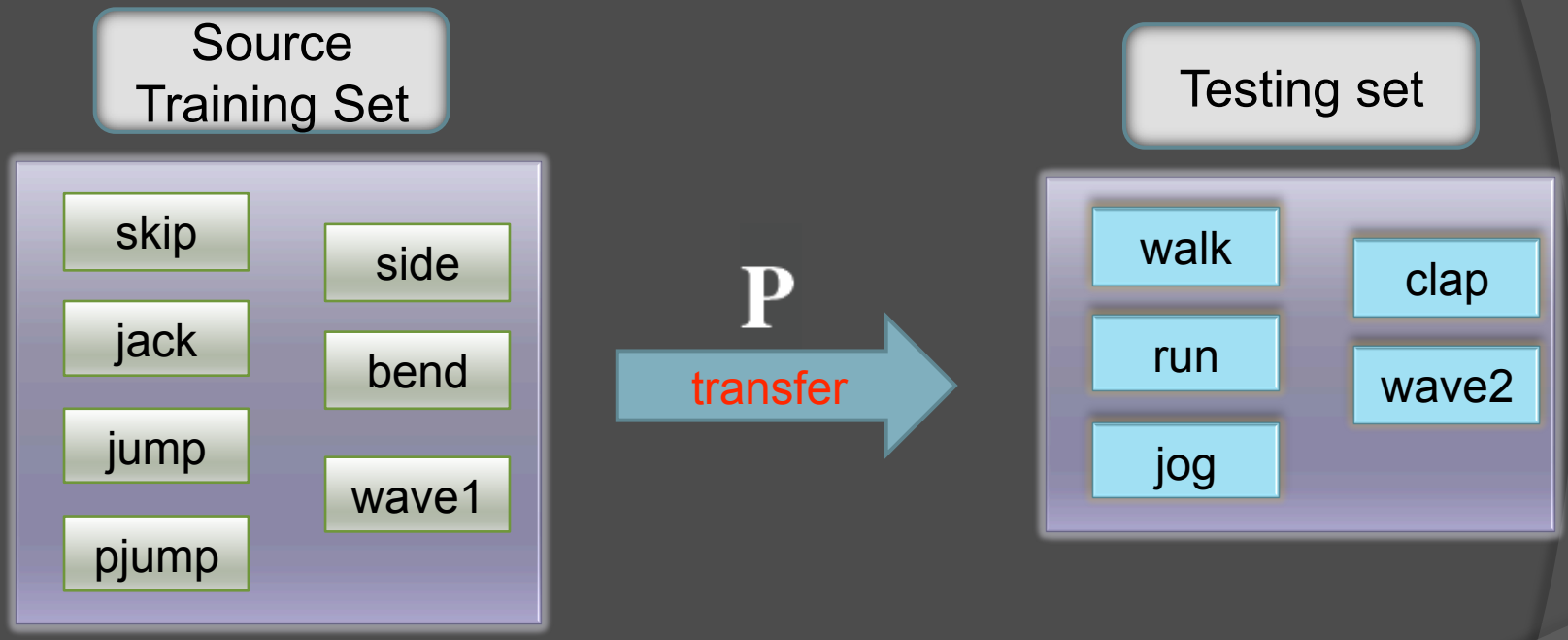
$D_{qt} = \sum_{s=1}^{M \times S} d_{qt,s}$

Distance

Action Detection

Experiments on Action Recognition

- Train the transferable distance function on Weizmann, and test on KTH.

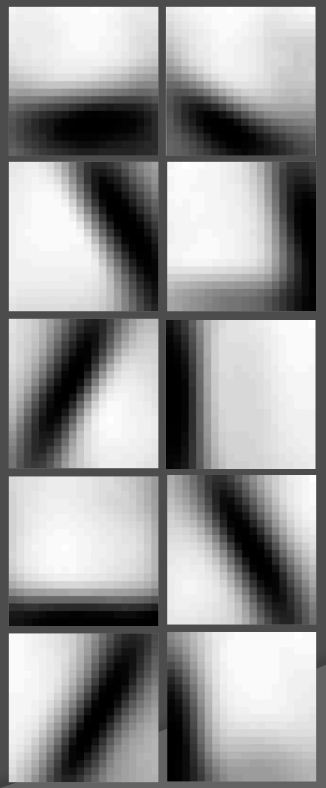


- The source training set does not contain the actions of the template set
- Each Action in the testing set has only one clip as template

Visualization

$$w_i = \langle \mathbf{P} \cdot \mathbf{f}_i \rangle$$

Learnt Weights on Testing Actions



Codeword Ranking

Five Rounds of Experiments

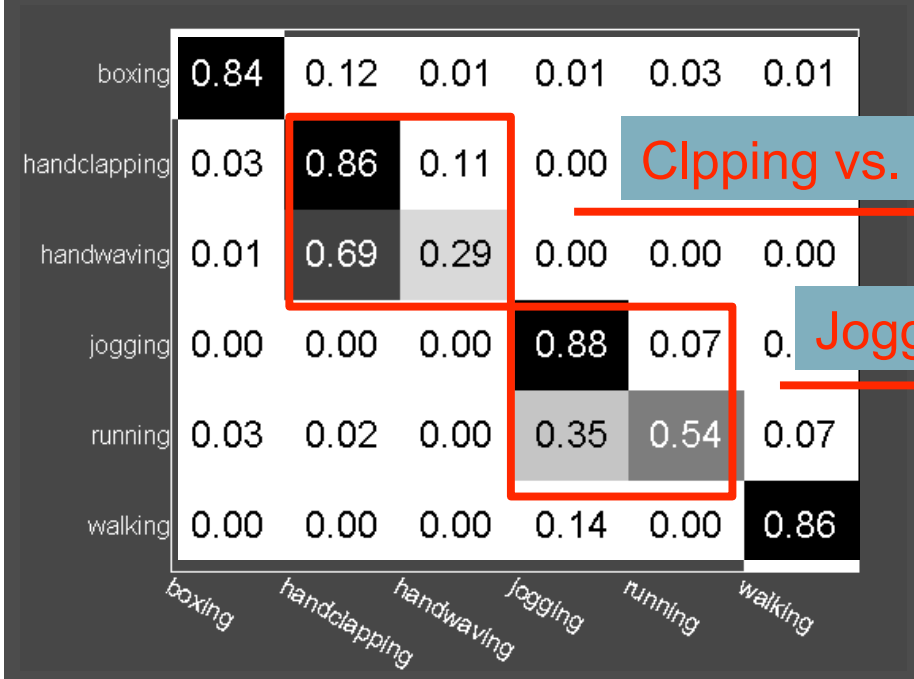
- For each round, we randomly select one actor, then choose one clip per action from this actor as the template.

	1	2	3	4	5	Avg.	Std.
Dc	0.776	0.709	0.829	0.564	0.746	0.725	0.100
Tr	0.784	0.767	0.829	0.617	0.789	0.757	0.073

5% improvement

Dc : Direct Comparison (W = 1)
Tr : Transferable Distance Function

Confusion Matrix of the Round 2



Clpping vs. Waving

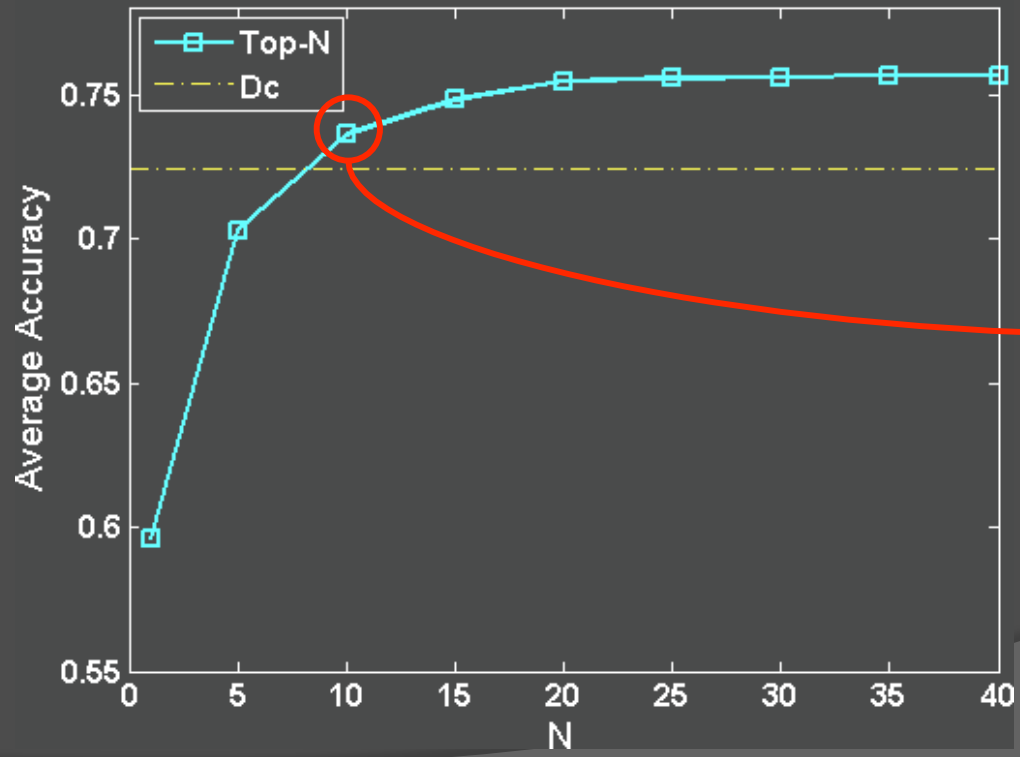
Jogging vs. Running

Direct Comparison
Avg: 70.9%

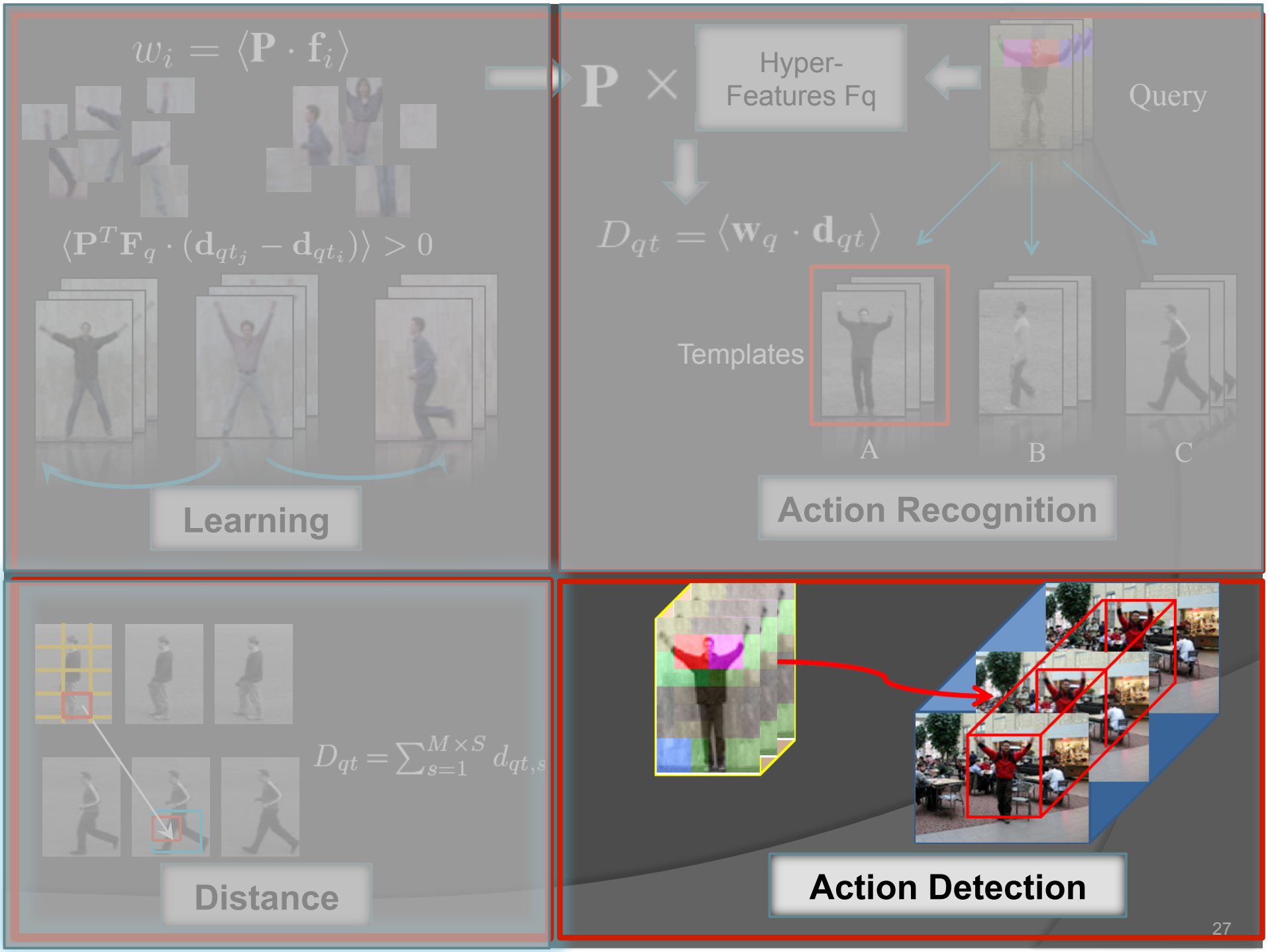
Transfer
Avg: 76.7%

Efficiency

- With the learnt distance function, we can sort the patches on each frame by their saliency.
- Instead of using all patches, we can choose the top N patches with high weights for matching.

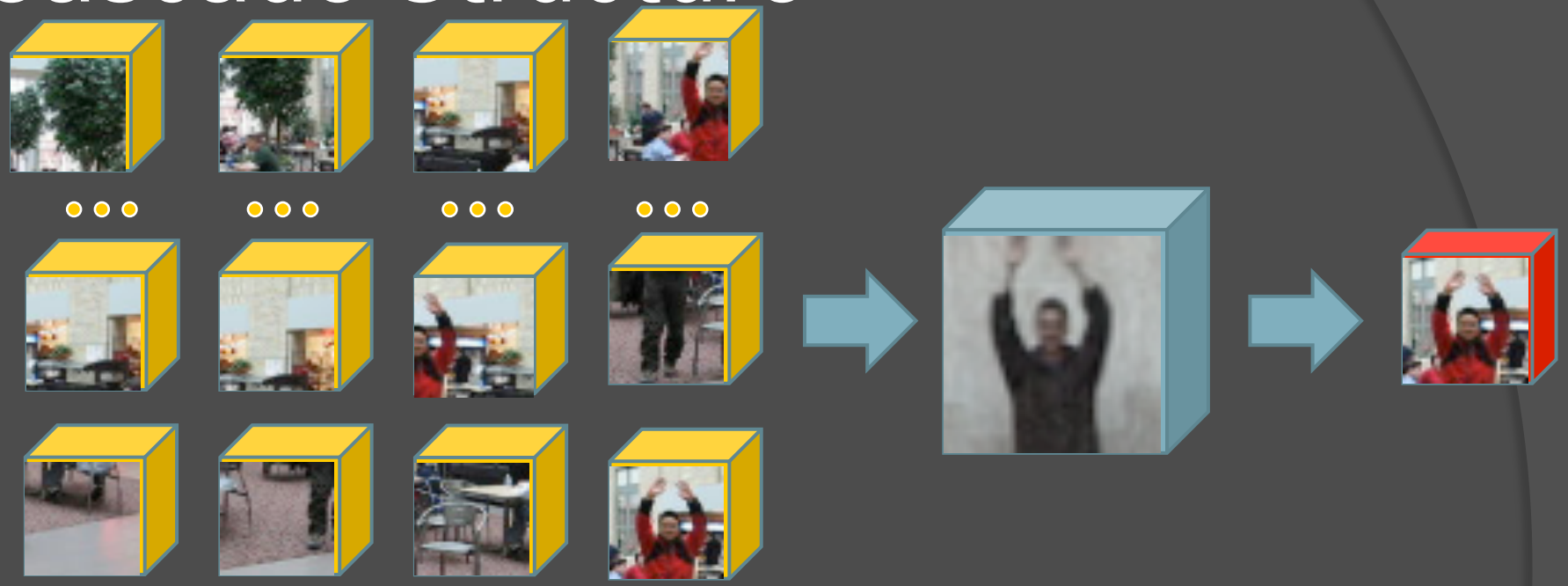


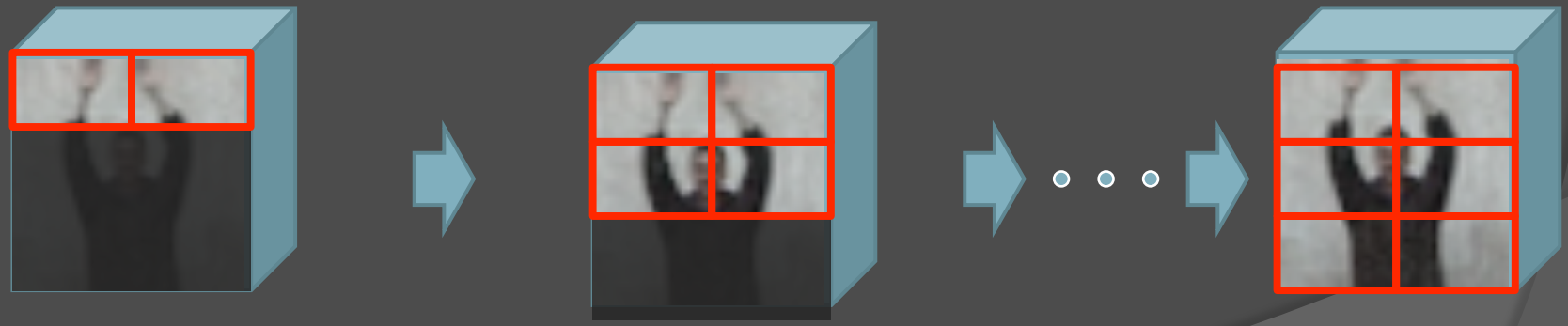
10 Patches on Each Frame


$w_i = \langle \mathbf{P} \cdot \mathbf{f}_i \rangle$
$\langle \mathbf{P}^T \mathbf{F}_q \cdot (\mathbf{d}_{qt_j} - \mathbf{d}_{qt_i}) \rangle > 0$
Learning
$\mathbf{P} \times$
Hyper-Features Fq
Query
$D_{qt} = \langle \mathbf{w}_q \cdot \mathbf{d}_{qt} \rangle$
Templates
A
B
C
Action Recognition
$D_{qt} = \sum_{s=1}^{M \times S} d_{qt,s}$
Distance
Action Detection
27

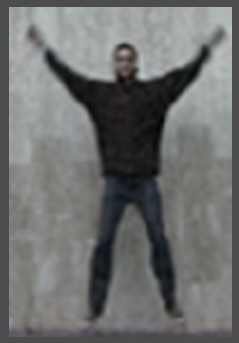
Human Action Detection

Cascade Structure



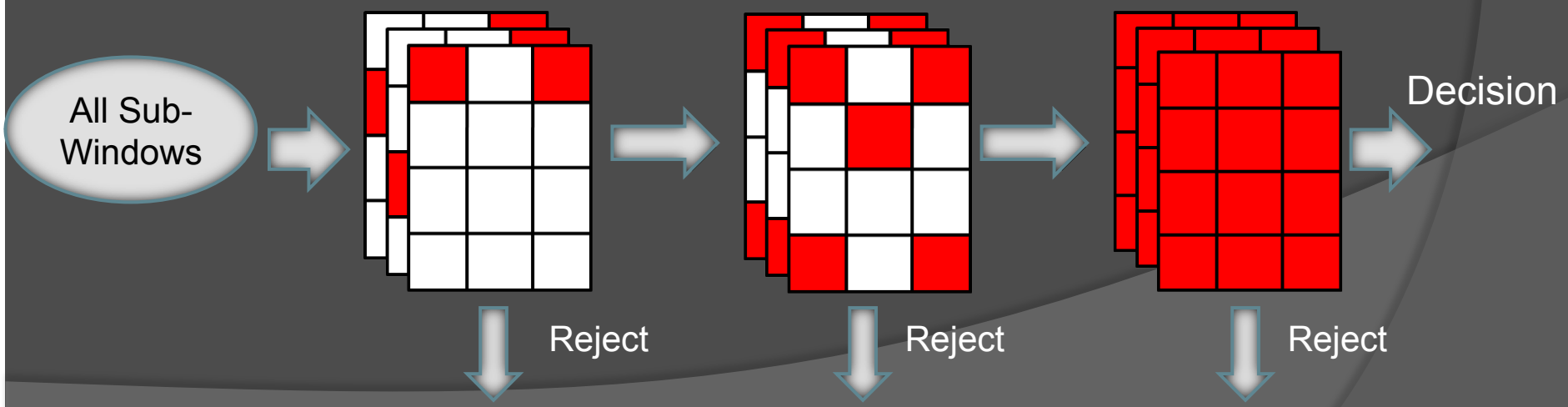


Cascade Structure

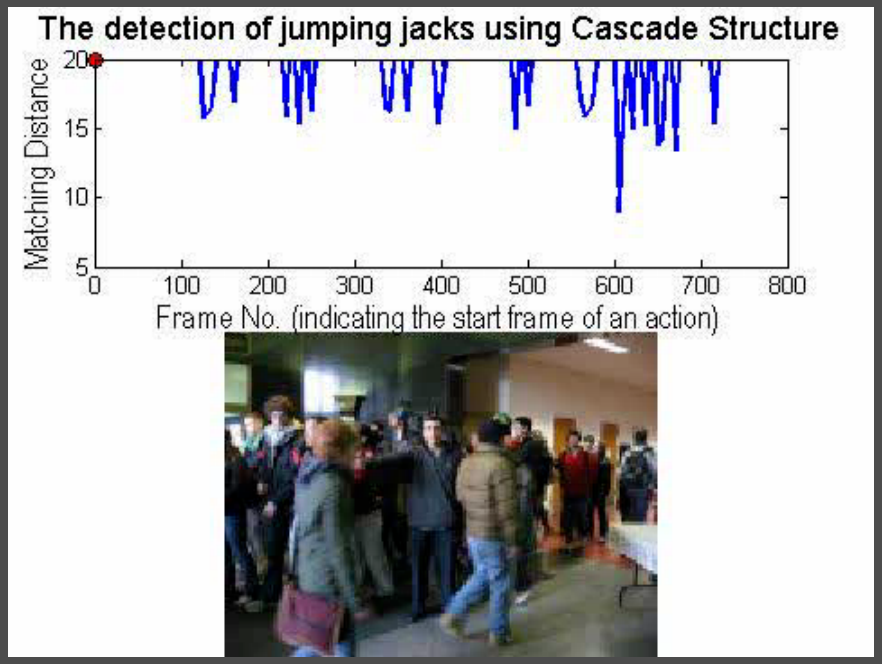


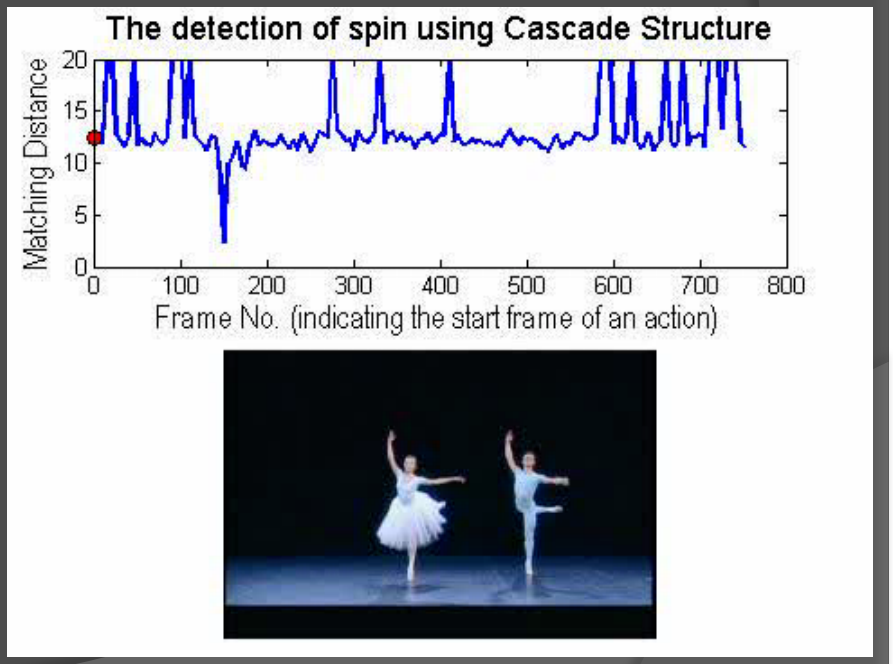
Hyper-Features Fq

$\times \mathbf{P} = \mathbf{w}_q$



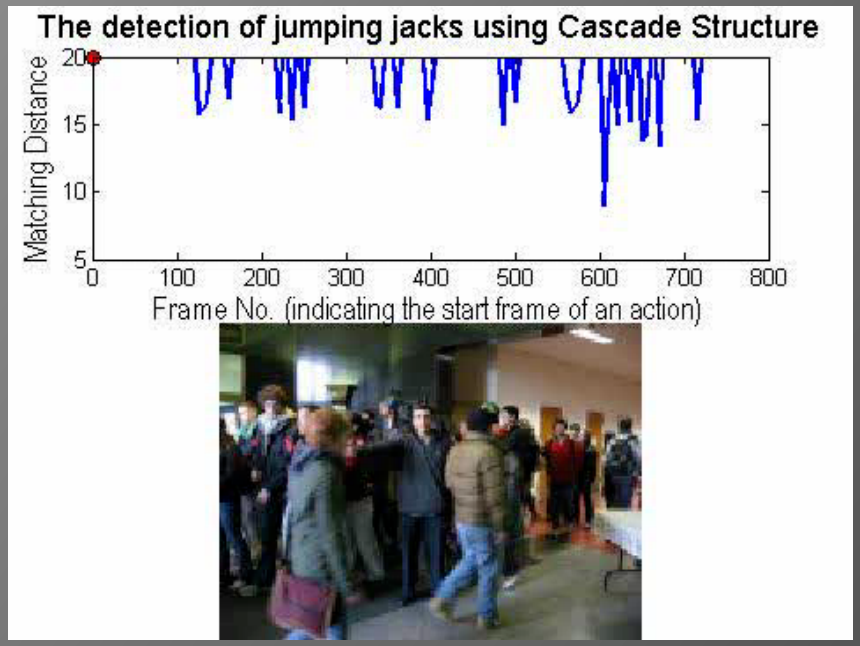
Efficient Action Detection

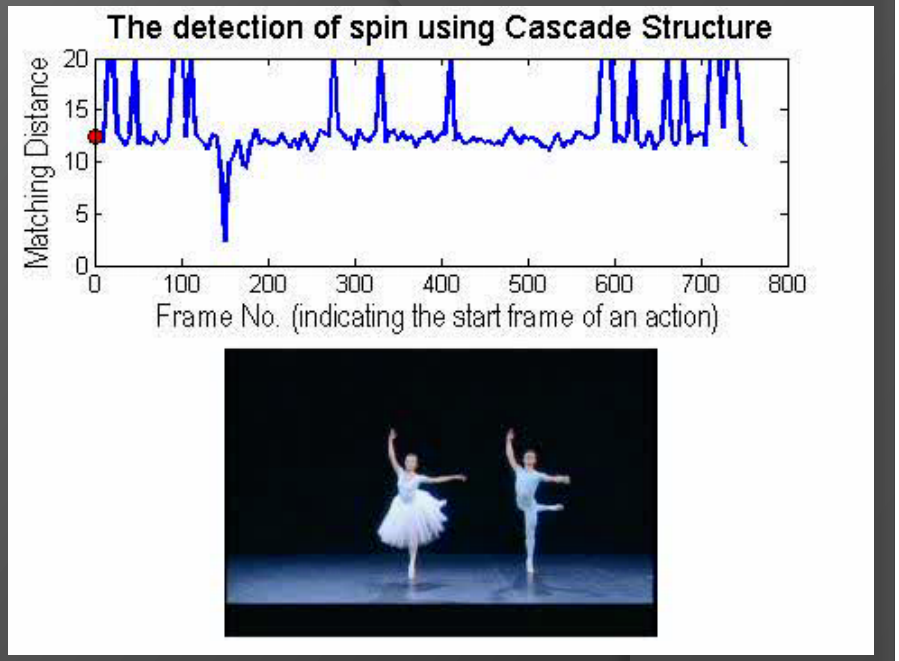




Contributions

- Transferable distance function Learning
  - Hyper-features based on appearance and positions
  - Max-margin Learning framework
- Action recognition from one clip
  - Template Matching based on motion
- Efficient action detection from one clip
  - Cascade structure





Thank You !

SFU

THINKING OF THE WORLD

SFU Vision and Media Lab