Modeling the Temporality of Visual Saliency and Its Application to Action Recognition

Luo Ye

2018-01-24

Content

1. Background

2. Modeling the Temporality of Video Saliency

3. Actionness-assisted Recognition of Actions

Content

1. **Background**
   1. Categorization of Visual Saliency Estimation Methods
   2. Existing Video Saliency (VS) Estimation Methods
   3. Our First Effort on Handling Temporality of Salient Video Object (SVO)

2. Modeling the Temporality of Video Saliency

3. Actionness-assisted Recognition of Actions

I. Categorization of Visual Saliency Methods

① ***Bottom-up*** VS. Top-down

② Image Saliency VS. ***Video Saliency***

or Static Saliency VS. ***Dynamic Saliency***

③ Deep learning based VS. ***Non-deep-learning based***

……

Problems Left Unsolved

From Image Saliency to Video Saliency

I. Features used at the Temporal Dimension: Motion

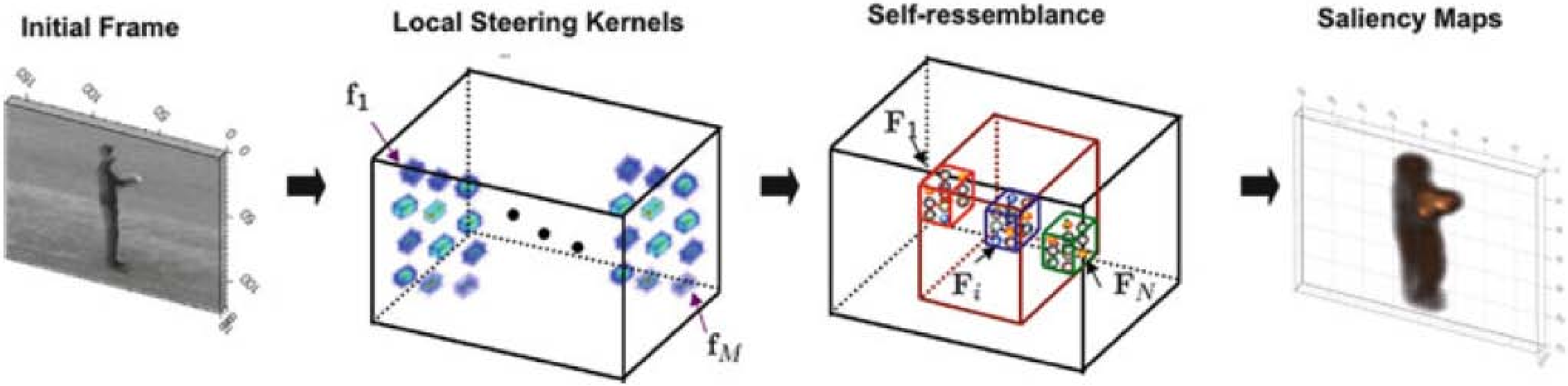
II. The way to watch (plenty of time v.s. limited time)

III. Memory effect

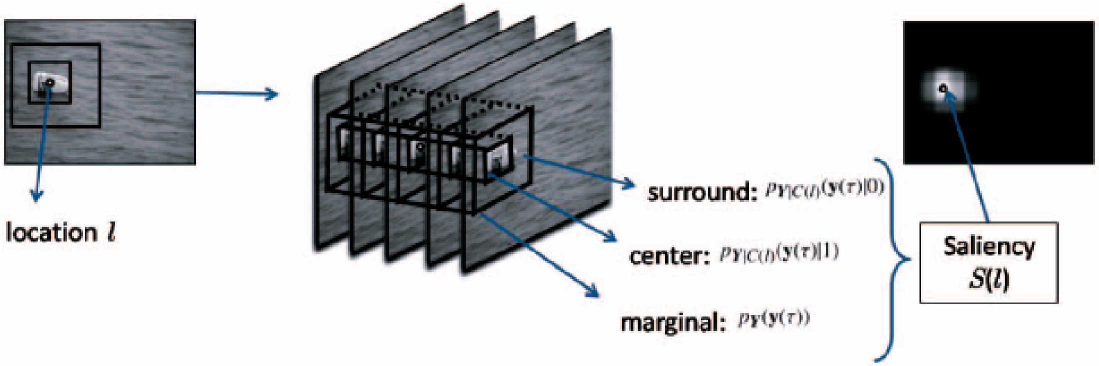
“ attention can also be guided by top-down, memory-dependent, or anticipatory mechanisms, such as when looking ahead of moving objects or sideways before crossing streets. ” *from wikipedia.org*

II. Existing VS Estimation Methods

1. Extension of 2D model (i.e. static saliency model)



Seo, H.J.J., Milanfar, P.: Static and space-time visual saliency detection by self-resemblance,Journal of Vision 2009


location l
surround: $p_{Y|C(l)}(\mathbf{y}(\tau)|0)$
center: $p_{Y|C(l)}(\mathbf{y}(\tau)|1)$
marginal: $p_Y(\mathbf{y}(\tau))$
Saliency
$S(l)$

Mahadevan V, Vasconcelos N. Spatiotemporal Saliency in Dynamic Scenes[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2010, 32(1):171.

II. Existing VS Estimation Methods Cont.

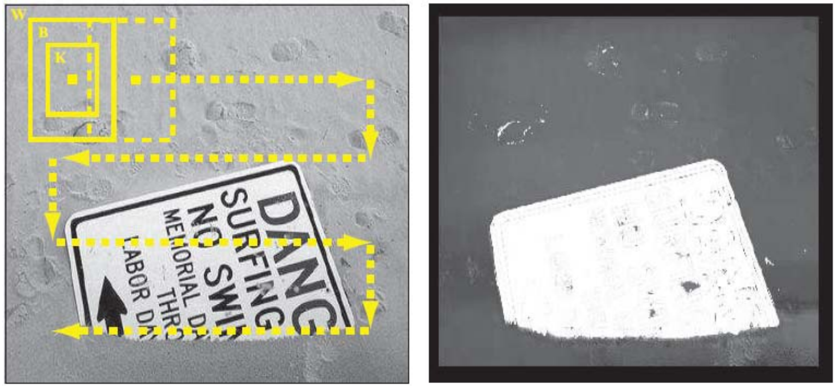
2. Static Saliency + Dynamic Saliency or Image Feature + Motion Features

$$q(t) = M(t) + RG(t)\mu_1 + BY(t)\mu_2 + I(t)\mu_3$$

$$I(t) = \frac{(r(t) + g(t) + b(t))}{3}$$

$$M(t) = |I(t) - I(t-\tau)|$$

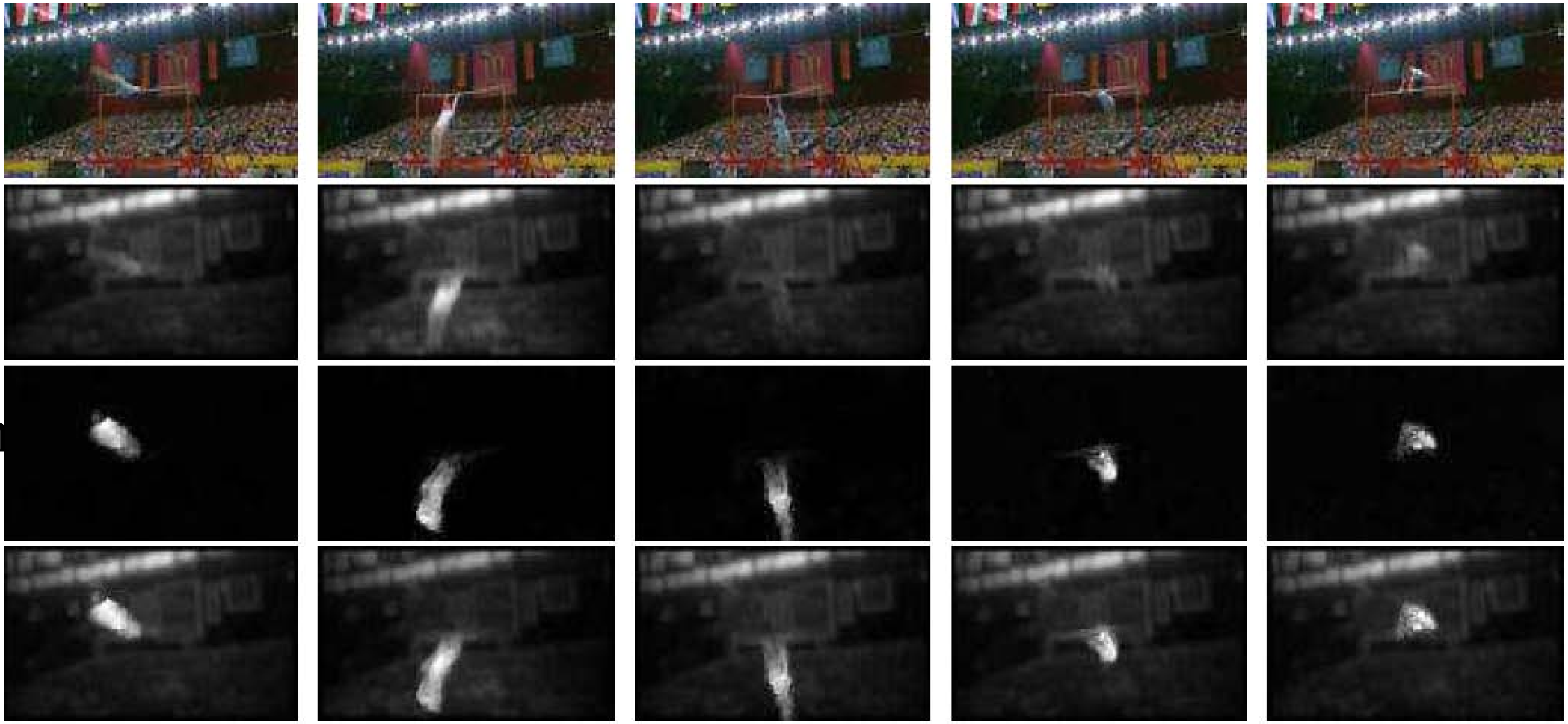
Guo, C., Zhang, L.: A novel multiresolution spatiotemporal saliency detection model and its applications in image and video compression. TIP 57 (2010) 1856-186



CIELab color values + the magnitude of optical flow

Rahtu, E., Kannala, J., Salo, M., Heikkil□a, J.: Segmenting salient objects from images and videos. In: ECCV. (2010)

III. Our First Effort on VS Temporality



Frames

S_image [1]

S_motion

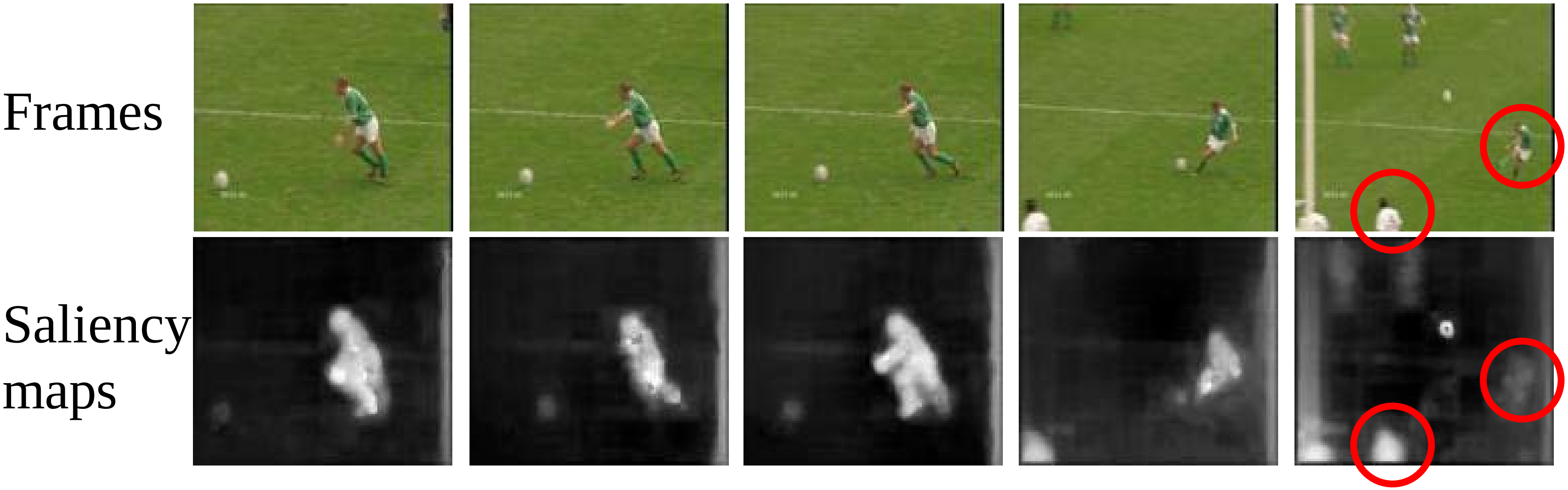
S_fused

$$M = N(S_M) + N(S_I),$$

$$S_M = \sqrt{(V_x - G_x)^2 + (V_y - G_y)^2}$$

[1] S. Goferman, L. Zelnik-Manor, and A. Tal. Context-aware saliency detection. In CVPR, 2010.

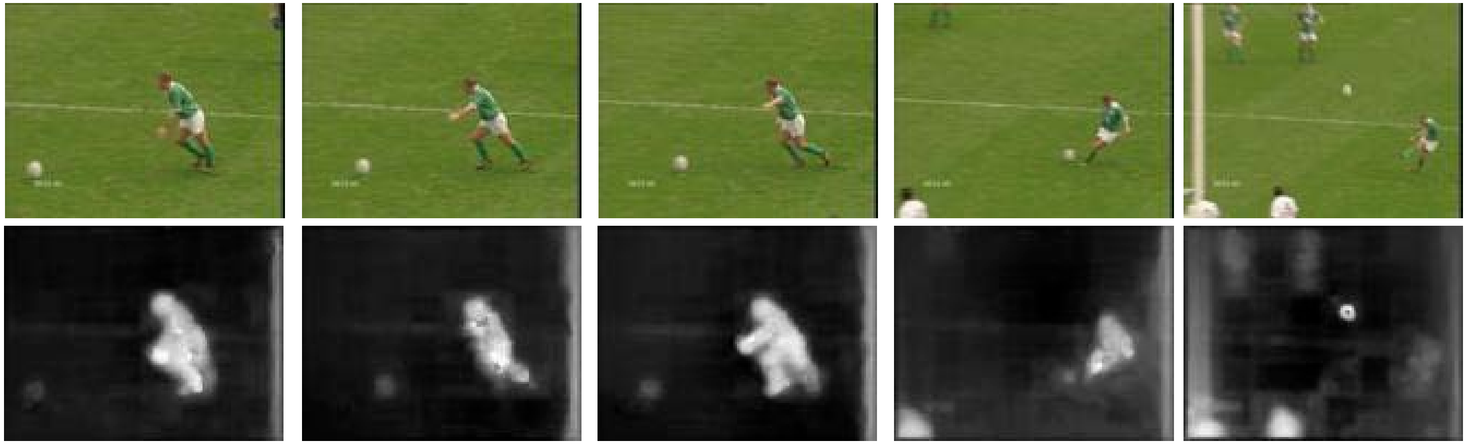
Problems of Existing VS method

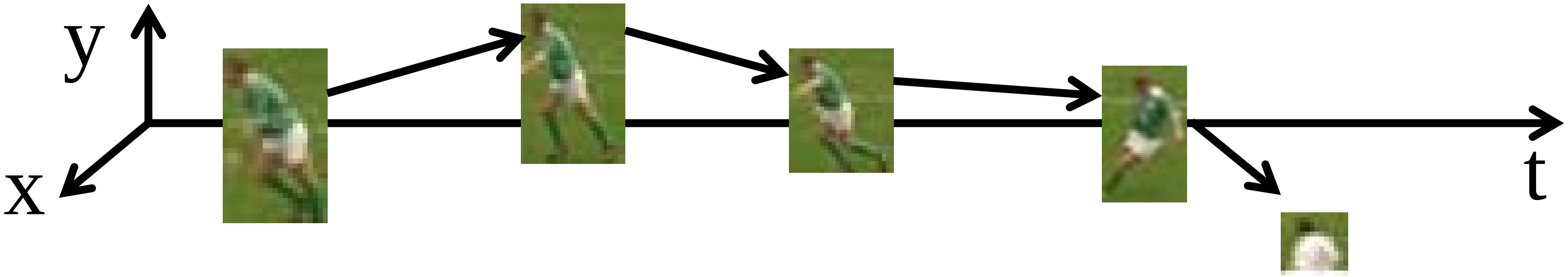


Observations:

1. Objects (including salient objects) in a video share strong temporal coherence.
2. Saliency estimation methods usually do not consider it, e.g. the detection of the coach instead of the football player.
3. A relatively long-term temporal coherence without memory affected is needed to estimate video saliency (VS).

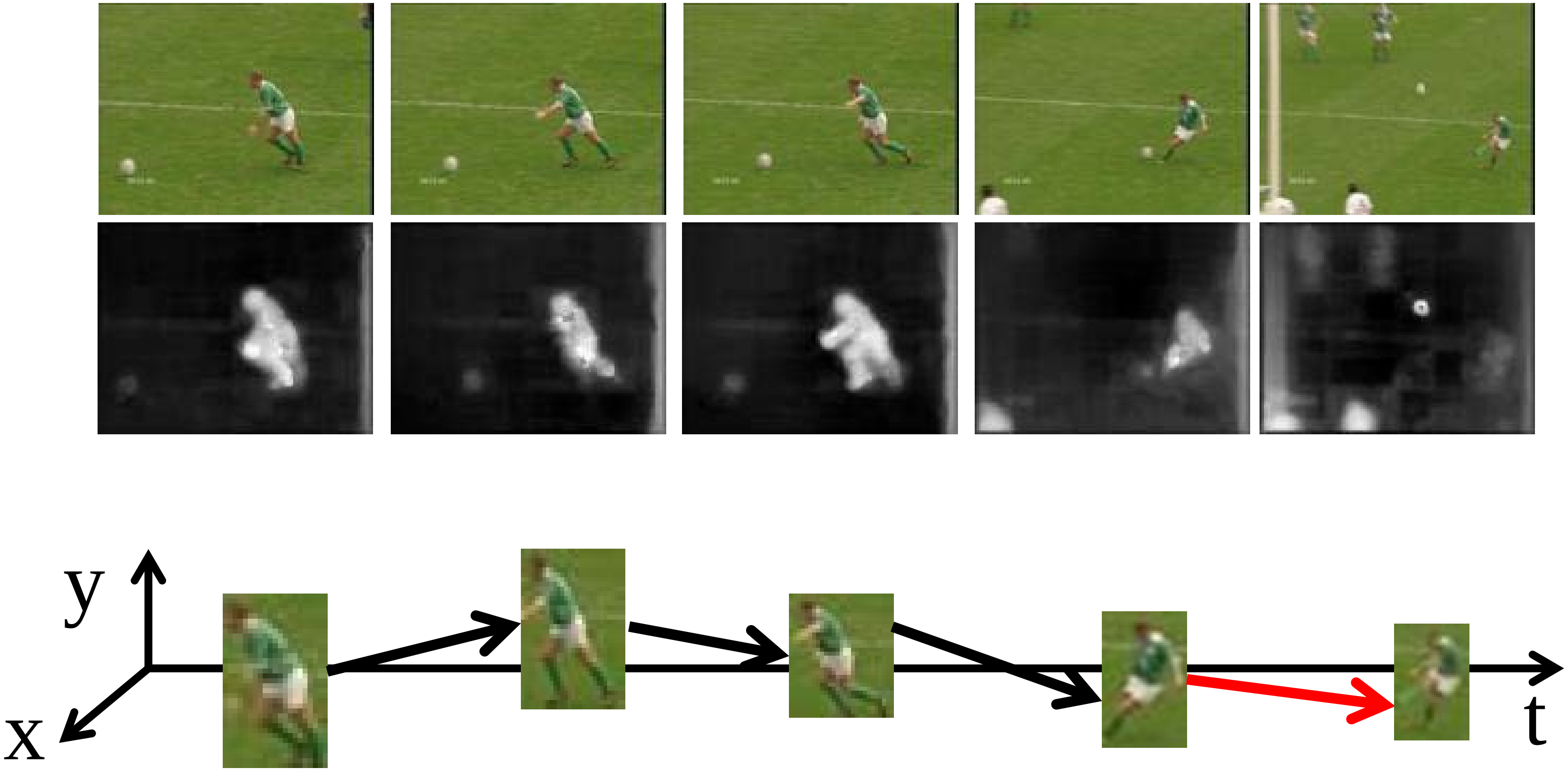
Without Temporal Coherence





Results by detecting the most salient object in each frame as the Salient Object of the Video (SVO)

Temporal Coherence Enhanced



Results of the Salient Object of the Video (SVO) when considering the long-term temporal coherence.

Our Method via Optimal Path Discovery[1]

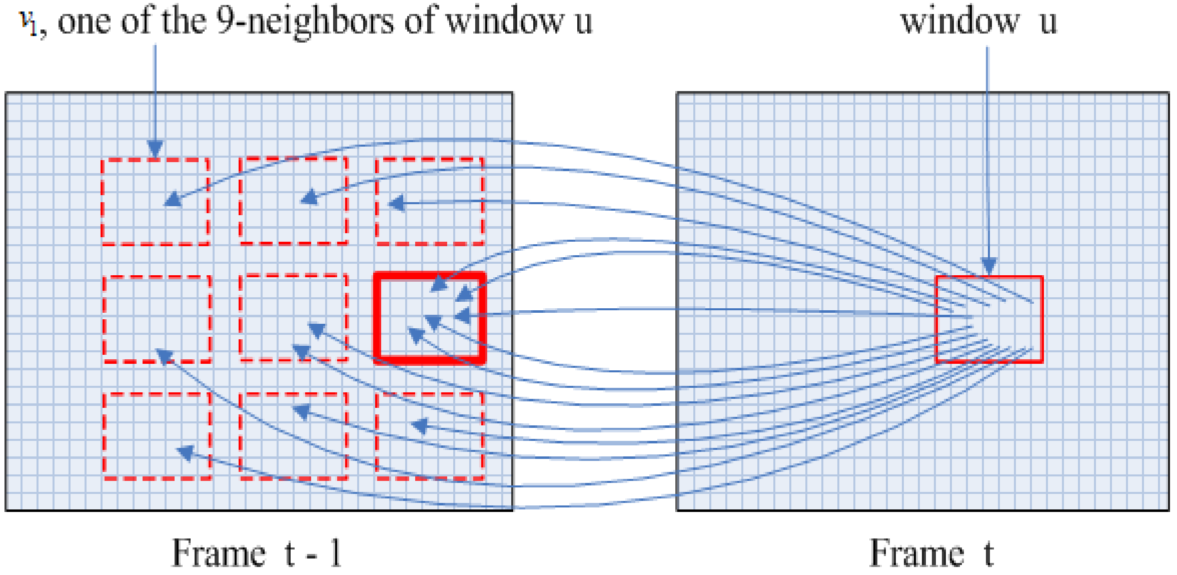
1. **Objective function:** salient video objects can be detected by finding the optimal path which has the largest accumulated saliency density in a video.

$$p^{*} = \arg\max_{p \in path(\mathbb{Q})} (D(p)),$$

Where $D(p) = \sum_{(x_s, y_s, t_s)}^{(x_e, y_e, t_e)} d(x, y, t)$, and ***d*** is the saliency density of a searching window centered at (x, y, t) , and ***p*** is a path starting from the starting point to the end point.

[1]***Ye Luo,*** Junsong Yuan and Qi Tian, “Salient Object Detection in Videos by Optimal Spatial-temporal Path Discovery”, ACM multimedia 2013, pp. 509-512.

2. Handling Temporal Coherence:



The temporal coherence of two windows centred at $u = (x, y)$ and v can be calculated as:

$$w_{(u,t)}(v, t\text{-}1) = \frac{N_i}{N}$$

The objective function of our salient video object detection becomes:

$$D(p) = \sum_{u,t} w_{(u,t)}(v, t\text{ - }1) \times d(u, t)$$

3. Dynamic Programming Solution

Every pixel in a frame is scanned with a searching window and a path is associated with it.

The path is elongated from $(v^*,t-1)$ to(u,t) on the current frame and the accumulated score along the path is updated as:

$$v^* = \max_{v\in N(\mathrm{u})}\{A(v,\mathrm{t}-1)+w_{(\mathrm{u},\mathrm{t})}(v,\mathrm{t}-1)\times \mathrm{d}(\mathrm{u},\mathrm{t})\}$$

$$A(u,t) = A(v^*,\mathrm{t}-1)+w_{(\mathrm{u},\mathrm{t})}(v^*,\mathrm{t}-1)\times \mathrm{d}(\mathrm{u},\mathrm{t})$$

To adapt to the size and the position changes of the salient objects, multi-scale searching windows are used.

Experiment Settings

Two datasets:

1. **UCF-Sports**: 150 videos of 10 action classes
2. **Ten-Video-Clips:** 10 videos of 5 to 10 seconds each

Compared Methods:

1. Our previously proposed **MSD**[13]
2. Optimal Path Discovery (**OPD**) Method[17]

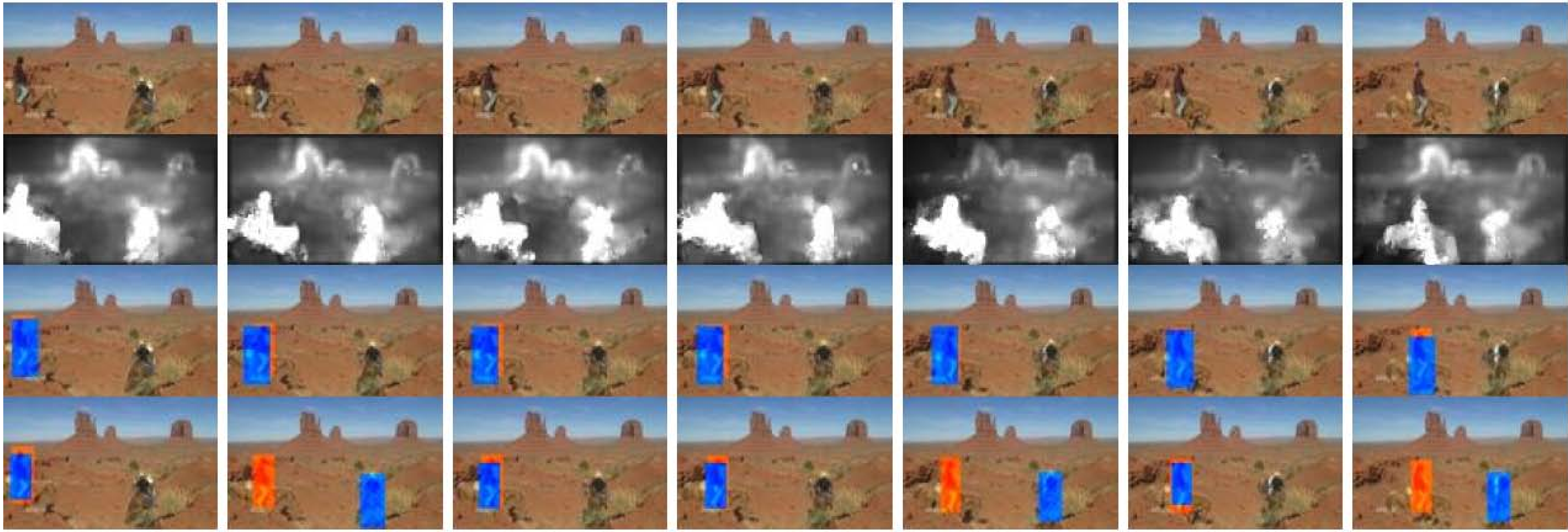
Evaluation Metrics:

$$\text{pre} = \frac{\sum S_g \times S_d}{\sum S_d}, \ \text{rec} = \frac{\sum S_g \times S_d}{\sum S_g}, \ \text{F-measure} = \frac{(1+\alpha)\times pre \times rec}{\alpha \times pre + rec}$$

[13] ***Ye Luo***, Junsong Yuan, Ping Xue and Qi Tian, “Saliency Density Maximization for Efficient Visual Objects Discovery”, in IEEE TCSVT, Vol. 21, pp. 1822-1834, 2011.

[17] D. Tran and J. Yuan. Optimal spatio-temporal path discovery for video event detection. In CVPR, 2011.

Experiments on UCF-Sports Dataset



First row: original frames; Second row: video saliency maps
Third row: our method ; Fourth row: MSD[1].
The blue mask indicates the detected results while the orange ones are the ground truth.

[1]***Ye Luo***, Junsong Yuan, Ping Xue and Qi Tian, "Saliency Density Maximization for Efficient Visual Objects Discovery", in IEEE TCSVT, Vol. 21, pp. 1822-1834, 2011.

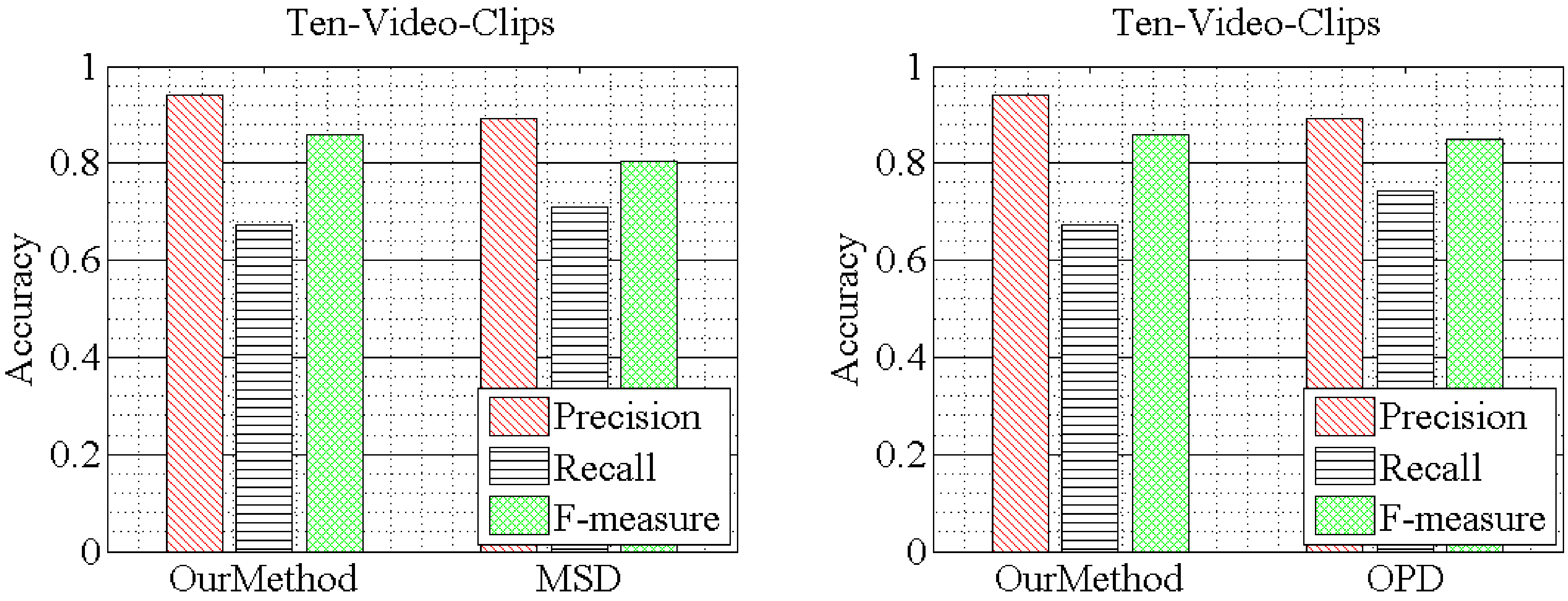
Experiments on UCF-Sports Dataset

Table. Averaged *F-measure* (%) ± *Standard Deviation* for ten types of action videos in UCF-sports dataset.

	Ride	Run	Kick	SwingSide	Lift
Ours	71.04±0.10	61.12±0.28	64.11±0.22	54.47±0.18	88.48±0.02
OPD[17]	68.77±0.23	55.57±0.10	64.10±0.23	37.29±0.15	87.63±0.01
MSD[13]	56.52±0.20	53.55±0.32	60.02±0.26	34.13±0.10	83.86±0.02
	Skate	Diving	Golf	SwingBench	Walk
Ours	46.33±0.35	69.76±0.13	62.88±0.26	59.06±0.18	54.17±0.26
OPD[17]	42.41±0.34	68.62±0.10	56.32±0.26	58.98±0.20	50.67±0.22
MSD[13]	40.69±0.34	61.50±0.13	52.22±0.23	58.62±0.19	45.74±0.20

[13] ***Ye Luo***, Junsong Yuan, Ping Xue and Qi Tian, “Saliency Density Maximization for Efficient Visual Objects Discovery”, in IEEE TCSVT, Vol. 21, pp. 1822-1834, 2011.
[17] D. Tran and J. Yuan. Optimal spatio-temporal path discovery for video event detection. In CVPR, 2011.

Experiments on Ten-Video-Clips Dataset



Precision, recall and F-measure comparisons for our method to MSD and OPD on Ten-Video-Clips dataset.

Content

1. Background

2. **Modeling the Temporality of Video Saliency**

3. Actionness-assisted Recognition of Actions

Motivation

1. **Conspicuity based models lack explanatory power for fixations in dynamic vision**

   Temporal aspect can significantly extend the kind of meaningful regions extracted, without resorting to higher-level processes.

2. **Unexpected changes and temporal synchrony indicate animate motions**

   Temporal synchronizations indicate biological movements with intentions, and thus meaningful to us.

The Proposed Method

1. Definition of our video saliency:

   Video Saliency = Abrupt Motion Changes + Motion Synchronization + Static Saliency

2. A hierarchical framework to estimate saliency in videos from ***three levels***:
   - The intra-trajectory level saliency
   - The inter-trajectory level saliency
   - Spatial static saliency[1]

3. The basic processing unit: a super-pixel trajectory[2]

$$Tr = \{R^s, \cdots, R^k, \cdots, R^e\},\ R \text{ is a superpixel}$$

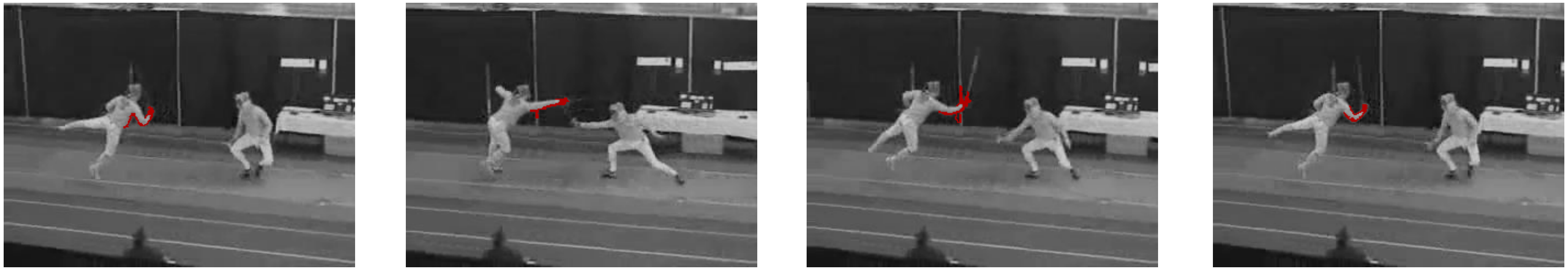
[1] Harel, J., Koch, C., Perona, P.: Graph-based visual saliency. In: NIPS. (2007) 545–552

[2] Chang, J., Wei, D., III, J.W.F.: A video representation using temporal superpixels. In: CVPR. (2013) 2051-2058

1. The intra-trajectory level saliency

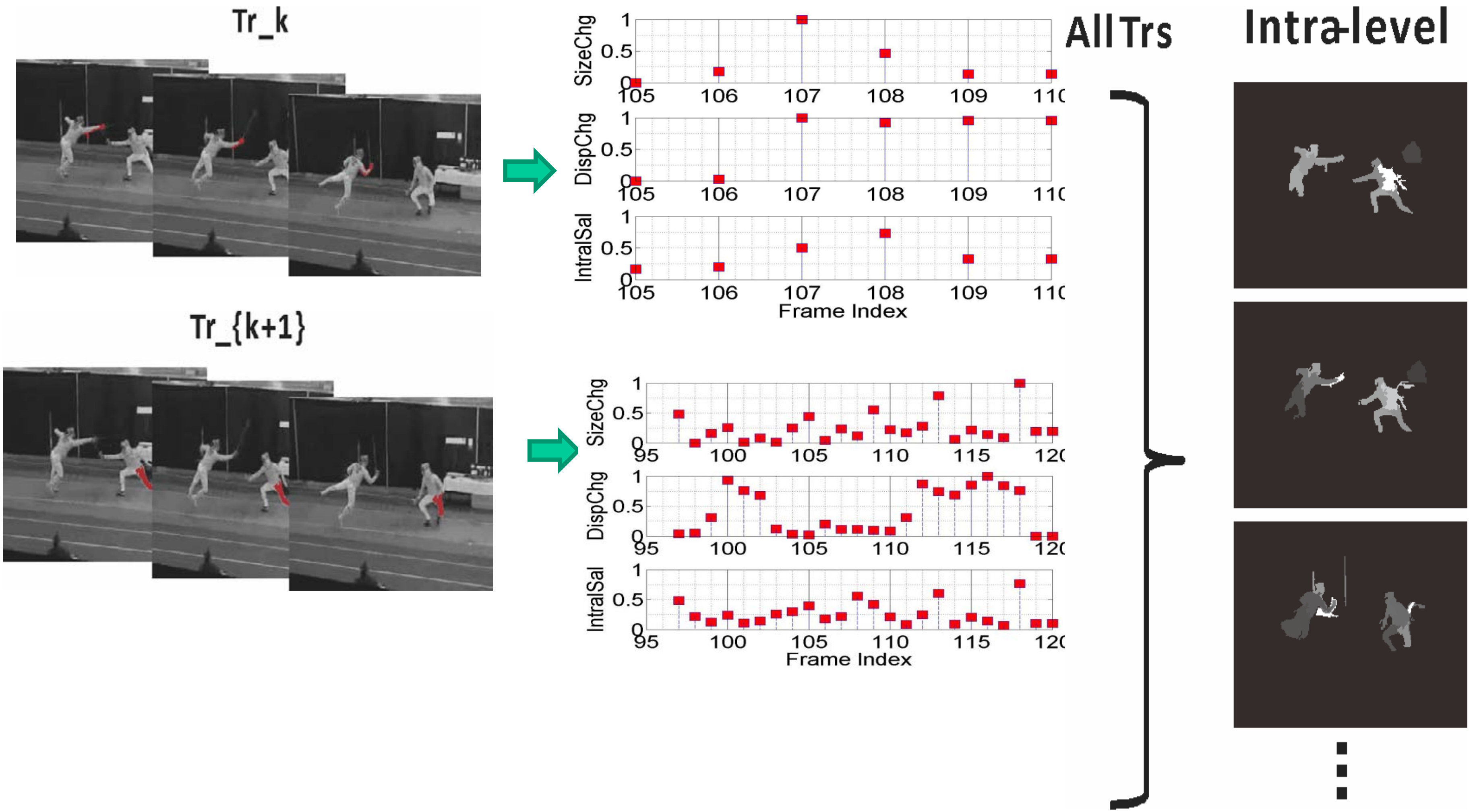
capturing the change of a super-pixel along a trajectory to measure the onset/offset phenomenon and sudden movement

$$S_{\mathrm{int}\,ra}(\mathrm{R}_i^k) = \begin{cases} \frac{1}{2}\left(\frac{\Delta R_{sz}^k}{\Delta R_{sz}^{\max}} + \frac{\Delta R_{disp}^k}{\Delta R_{disp}^{\max}}\right) & t_i^s < k < t_i^e \\ 1 & k = t_i^s \text{ or } k = t_i^e \end{cases}$$



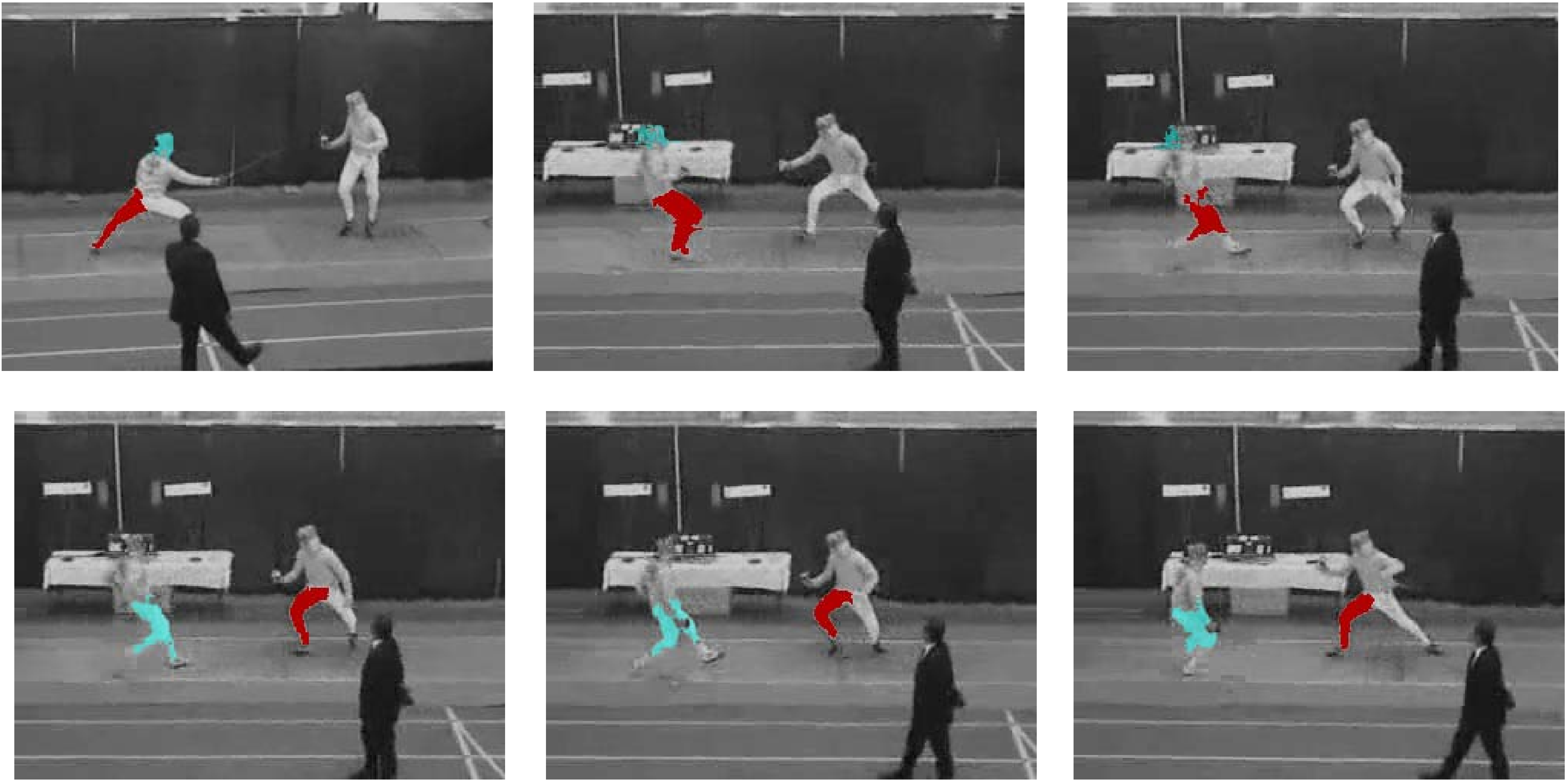
The size and the displacement changes of a super-pixel along time axis

1. The intra-trajectory level saliency cont.



2. The inter-trajectory level saliency

Synchronized motions existing between different parts of human bodies.

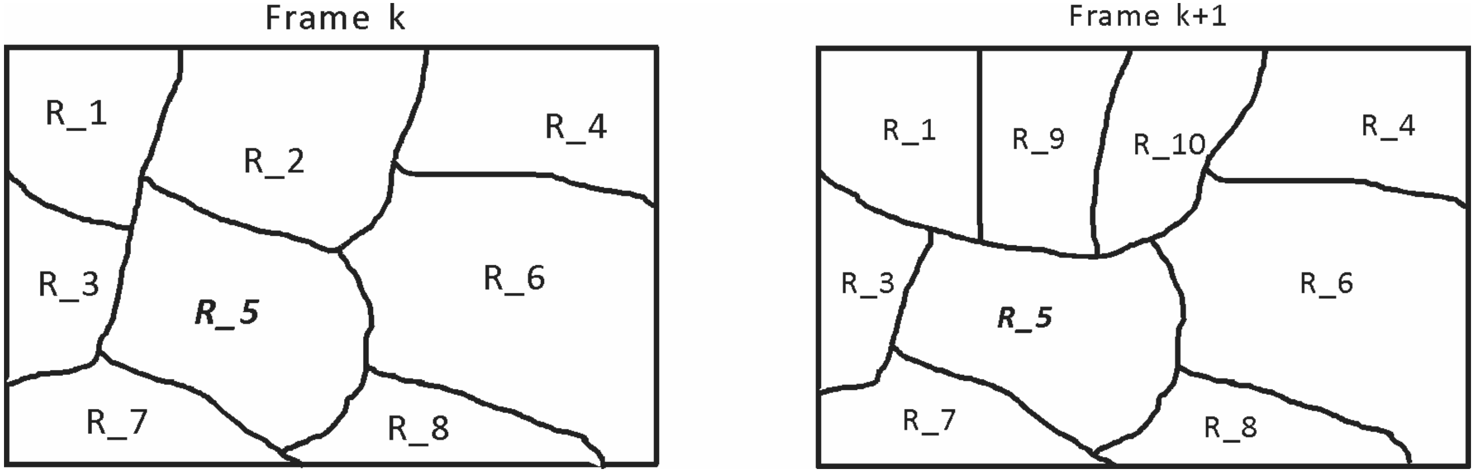


2. The inter-trajectory level saliency

using mutual information to measure the synchronization between two trajectories

$$MI(Tr_i,Tr_j)=\begin{cases}\frac{1}{2}\log\frac{|C_{ii}|\cdot|C_{jj}|}{|C|} & Tr_j\notin \mathrm{N}(Tr_i)\text{ and }\left|\left\{t^s,\cdots,t^e\right\}\right|\geq 3\\ 0 & Otherwise\end{cases}$$

$$S_{inter}(\mathrm{R}_i^k)=S_{inter}(Tr_i)=\max{}_j(\mathrm{MI}(Tr_i,Tr_j))\times \mathrm{H}_i$$



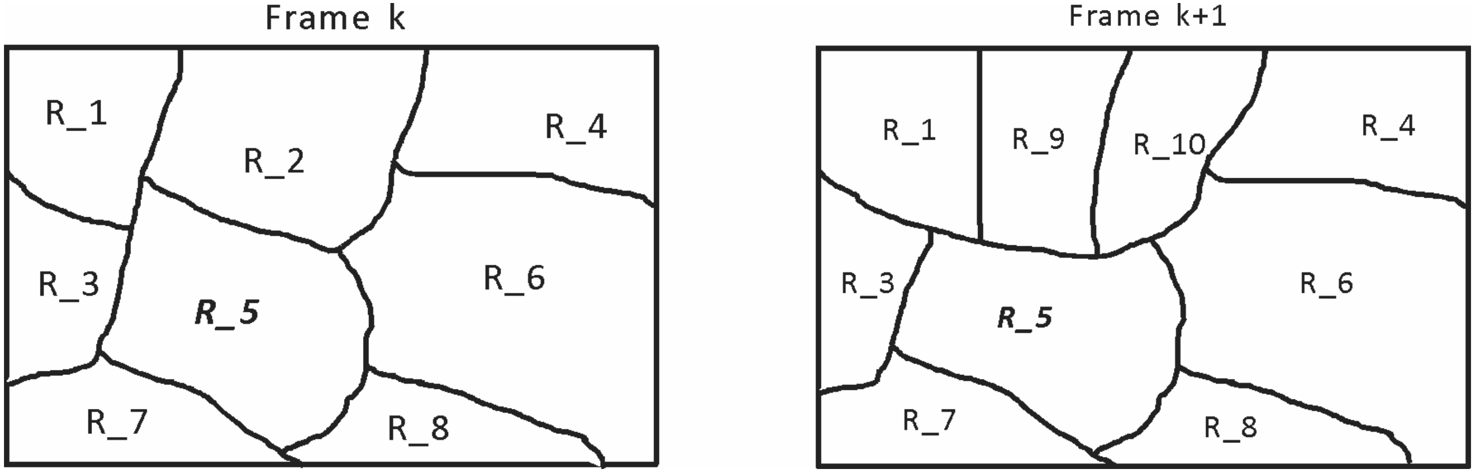
The spatial-temporal neighbors of Tr5 (i.e. R_5) at frame k and frame k + 1.

2. The inter-trajectory level saliency

using mutual information to measure the synchronization between two trajectories

$$MI(Tr_i, Tr_j) = \begin{cases} \frac{1}{2}\log\frac{|C_{ii}|\cdot|C_{jj}|}{|C|} & Tr_j \notin \mathrm{N}(Tr_i) \text{ and } \left|\{t^s, \cdots, t^e\}\right| \geq 3 \\ 0 & \textit{Otherwise} \end{cases}$$

$$S_{inter}(\mathrm{R}_i^k) = S_{inter}(Tr_i) = \max_j(\mathrm{MI}(Tr_i, Tr_j)) \times \mathrm{H}_i$$



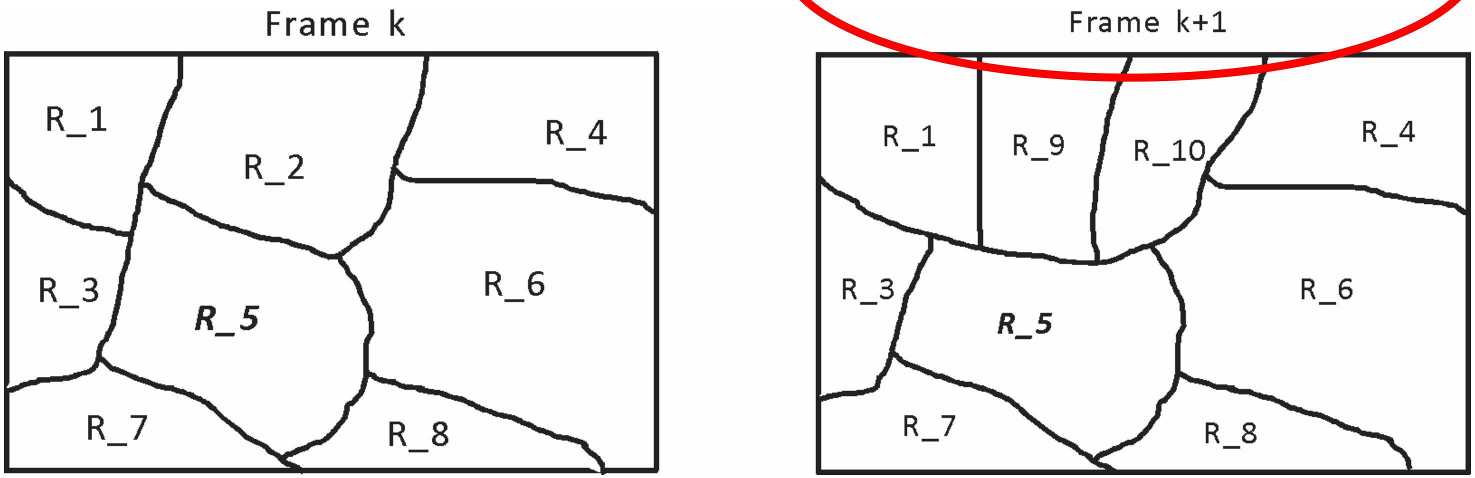
The spatial-temporal neighbors of Tr5 (i.e. R_5) at frame k and frame k + 1.

2. The inter-trajectory level saliency

using mutual information to measure the synchronization between two trajectories

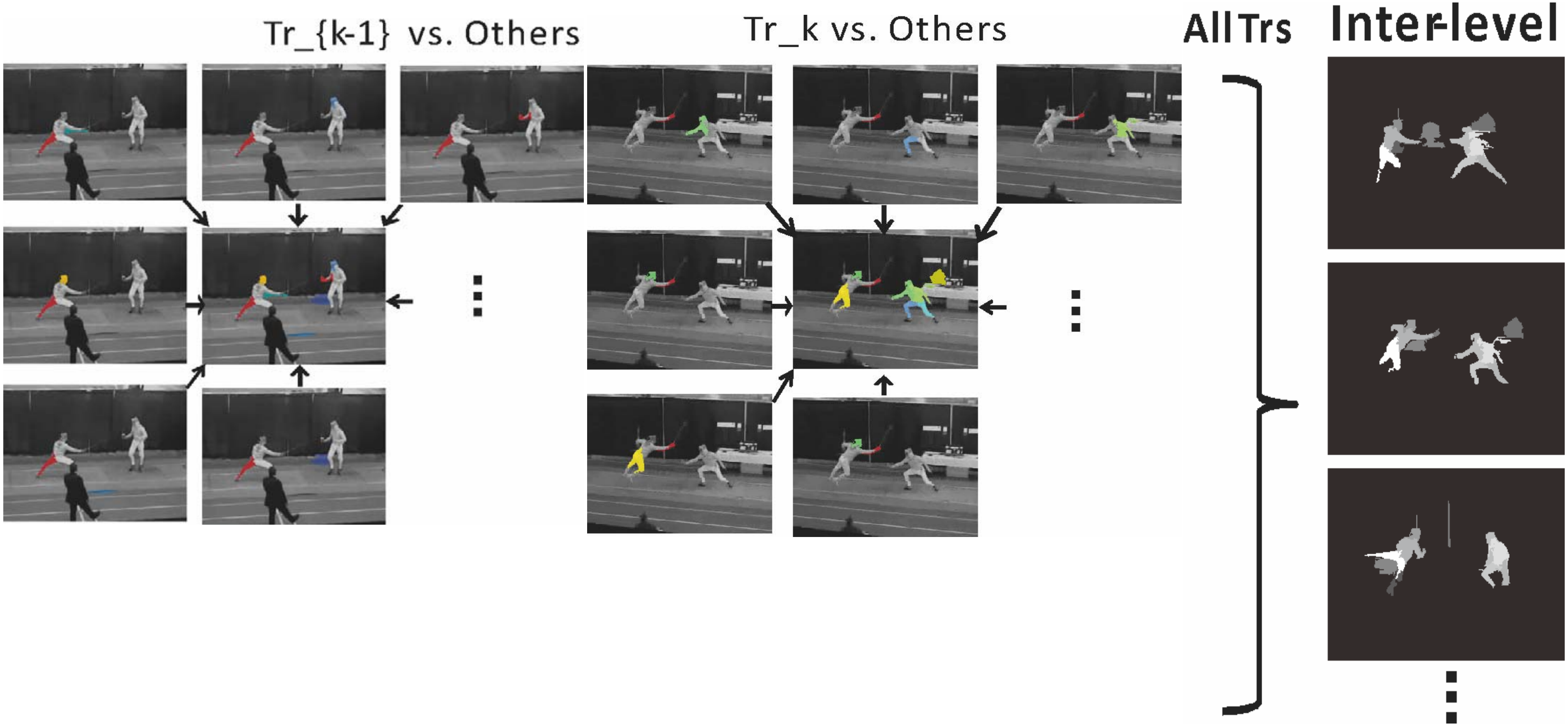
$$MI(Tr_i,Tr_j)=\begin{cases}\frac{1}{2}\log\frac{|C_{ii}|\cdot|C_{jj}|}{|C|} & Tr_j\notin \mathrm{N}(Tr_i) \text{ and } \left|\{t^s,\cdots,t^e\}\right|\geq 3\\ 0 & Otherwise\end{cases}$$

$$S_{inter}(\mathrm{R}_i^k)=S_{inter}(Tr_i)=\max_j(\mathrm{MI}(Tr_i,Tr_j))\times \mathrm{H}_i$$



The spatial-temporal neighbors of Tr5 (i.e. R_5) at frame k and frame k + 1.

2. The inter-trajectory level saliency cont.



The super-pixel (in red) has different levels of synchronization to other super-pixels (in other colors) which are corresponding to various parts of both fencers.

3. Fusing Scheme and Others:

1. Normalization
   - Spatial level: normalized into [0,1] per frame
   - Intra-level and inter-level: normalized into [0,1] per video

2. Fusion scheme for each super-pixel on frame k

$$S\left(R_i^k\right)=\frac{1}{3}\left(S_{static}(\mathrm{R}_i^k)+\mathrm{S}_{\mathrm{int}\,ra}(\mathrm{R}_i^k)+\mathrm{S}_{\mathrm{int}\,er}(\mathrm{R}_i^k)\right)$$

3. Camera Motions: RANSAC, homograph estimation, and motion compensation

4. Inhibition-of-Return: Not considered in this paper

Experimental Settings

Four datasets:

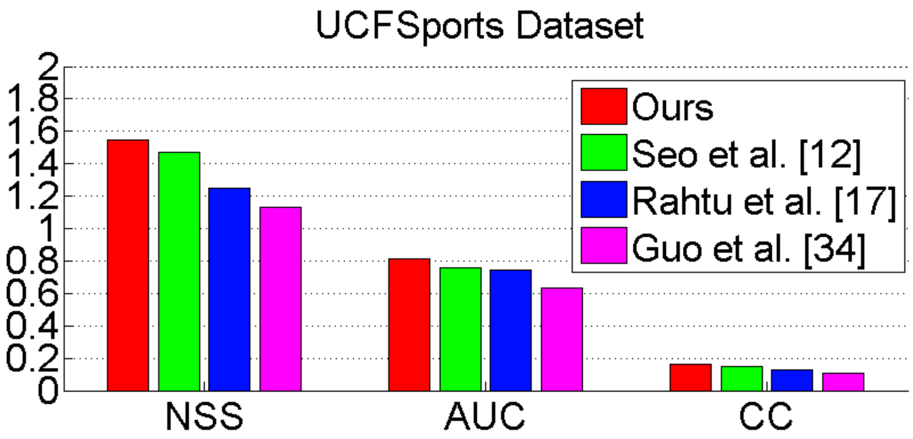
- UCF-sports: eye tracking data
- ASCMN: eye tracking data
- Ten-video-clip: human labeled mask
- Interaction dataset: self-collected dataset with human labeled masks provided

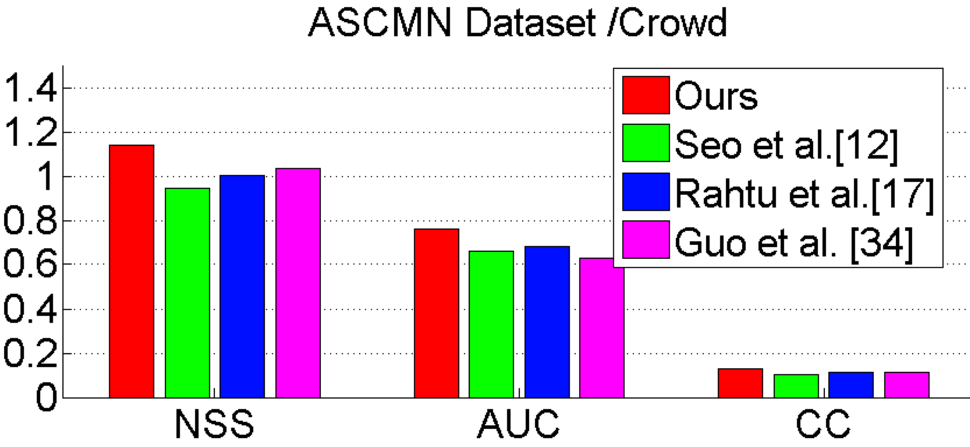
Four evaluation metrics

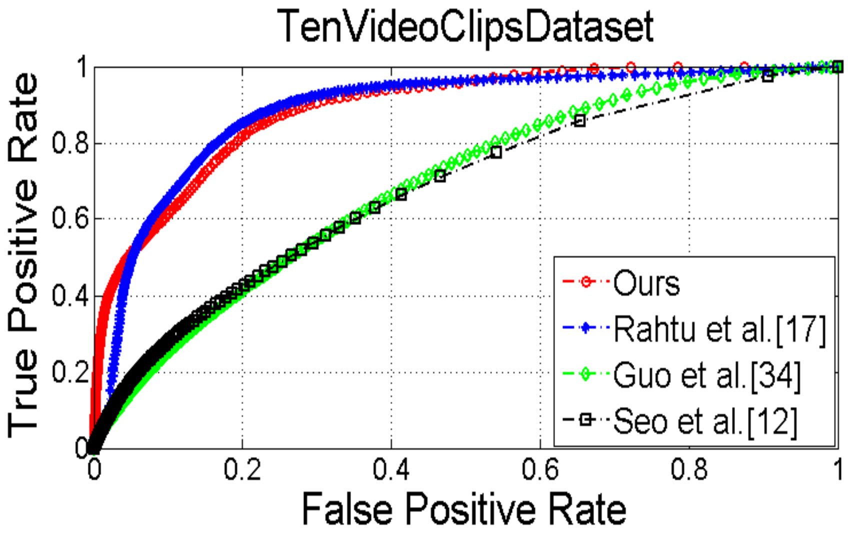
- Area under Receiver Operating Characteristics Curve (ROC-AUC)
- Normalized Scanpath Saliency (NSS)
- Linear Correlation Coefficients (CC)
- True positive rate vs. false positive rate curve

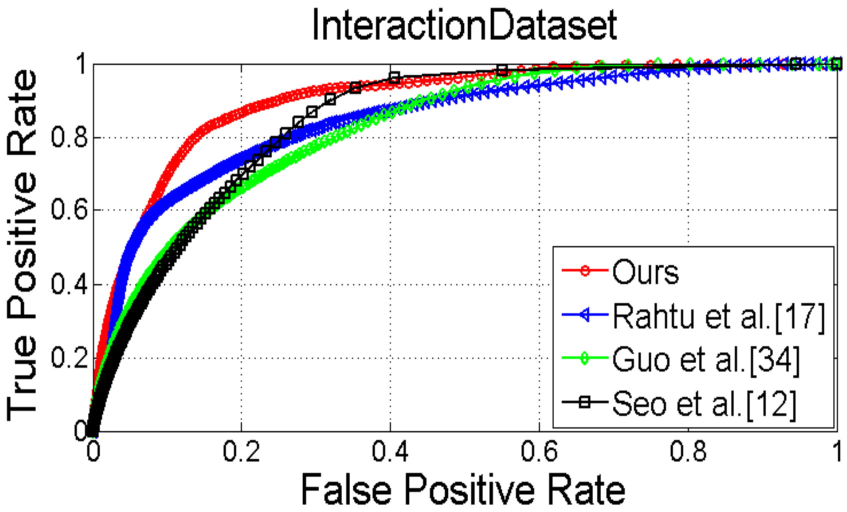
Experimental Results

1. Comparisons with 3 methods on employed four datasets


UCFSports Dataset

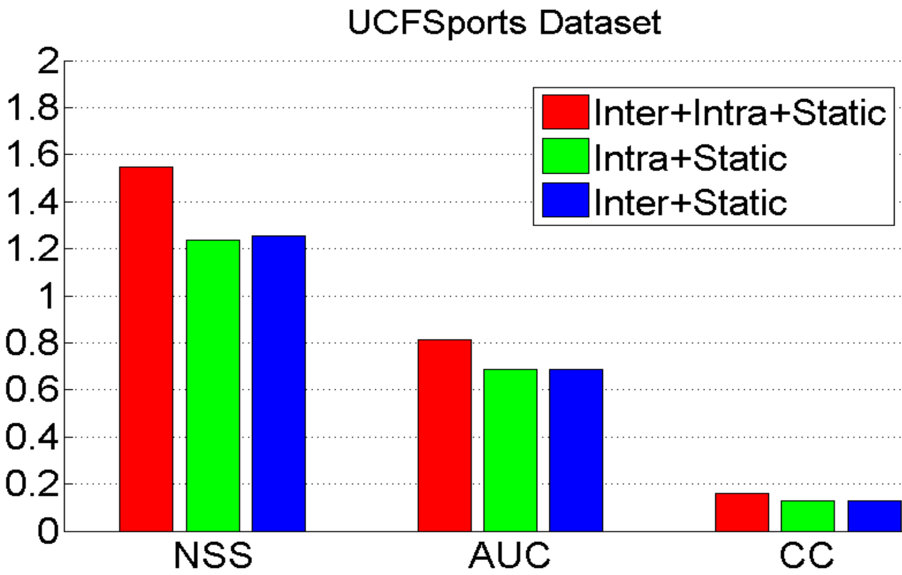

ASCMN Dataset /Crowd


TenVideoClipsDataset


InteractionDataset

Experimental Results cont.

2. Performance of individual component of our method

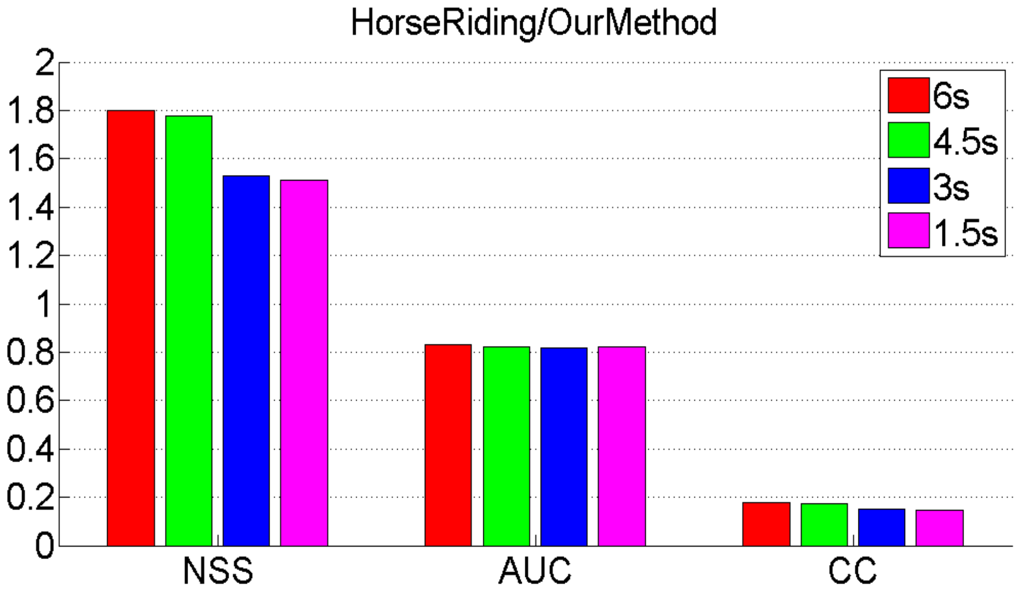


Findings:

1. Marginally improvements are obtained: inter-level saliency + the static saliency or the intra-level saliency + static saliency.
2. All three levels together, there is a substantial increase in performance

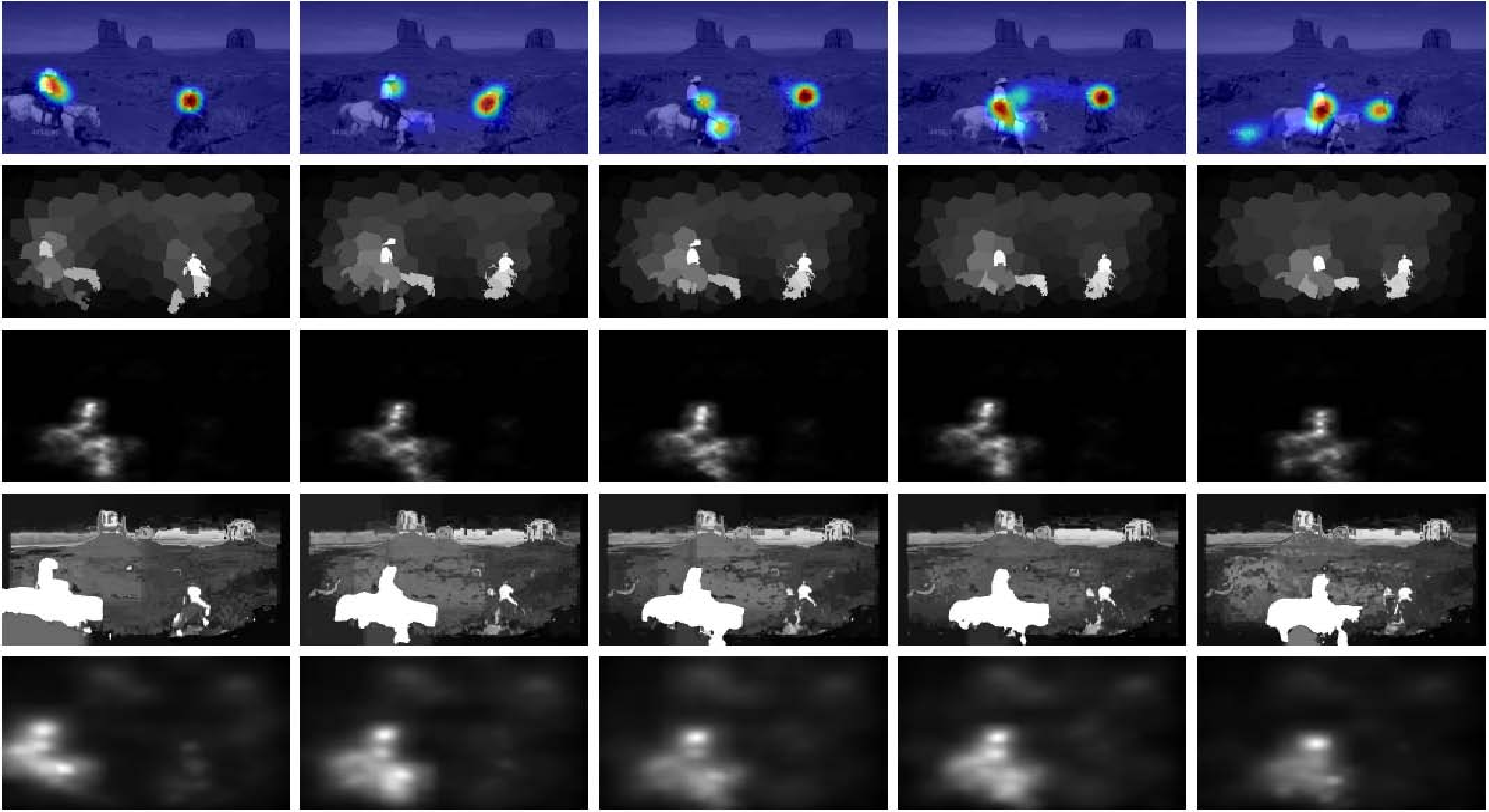
Experimental Results cont.

3. Video clip length vs. performance



Findings:

1. In accordance with human's short-term memory, there is a upper-limit of the length of the video clip used in our method, e.g. 6 second.
2. Under the upper limit of the video length, longer time durations generally improves the performance

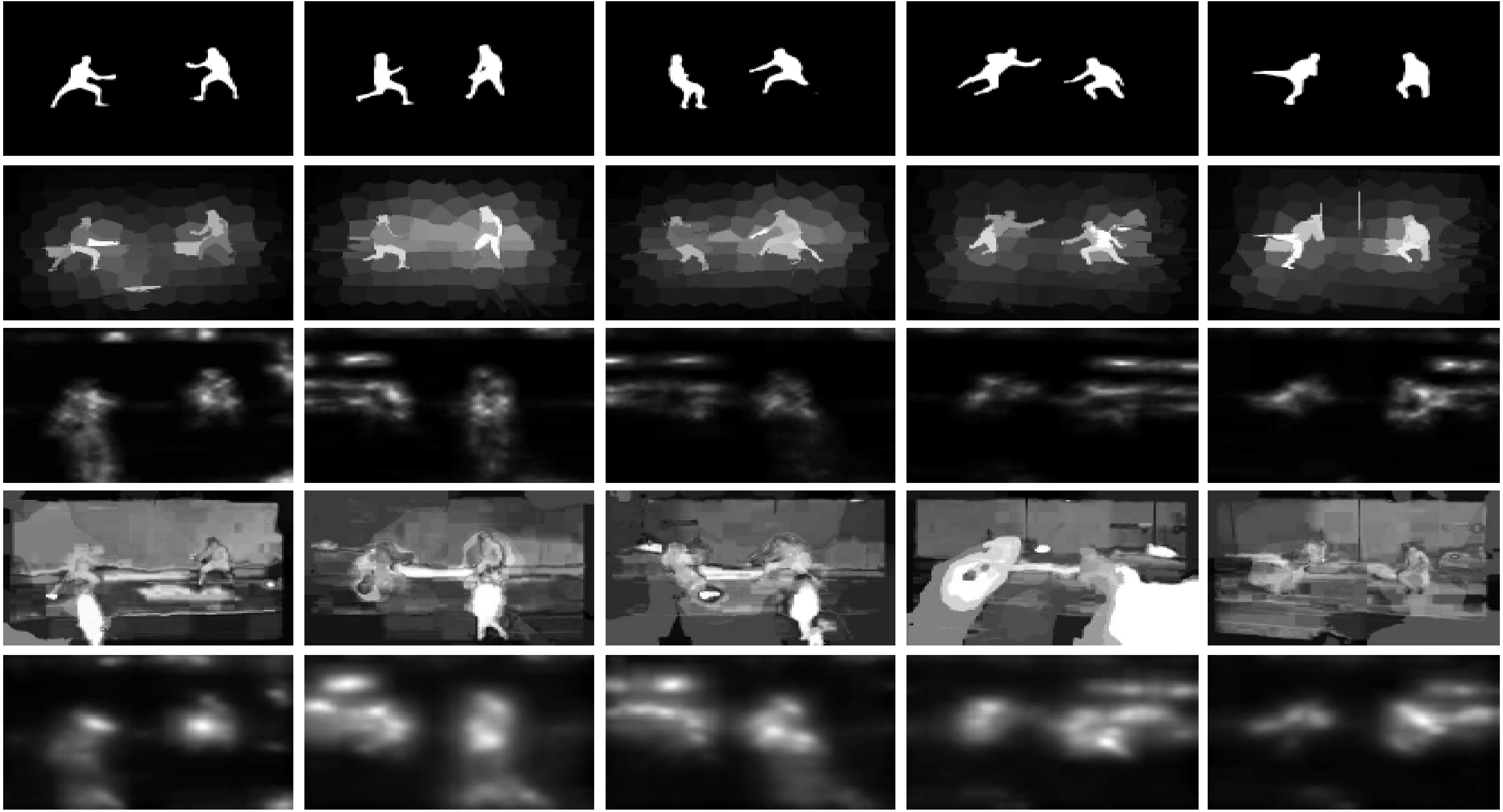


First row: fixation maps; Second row: our results; Third row: results of [12]; Fourth row: results of [17] and the fifth row: results of [34]. Our results better fit to the human fixations than other methods.

[12] Seo, H.J.J., Milanfar, P.: Static and space-time visual saliency detection by self-resemblance,Journal of Vision 2009

[17] Rahtu, E., Kannala, J., Salo, M., Heikkil a, J.: Segmenting salient objects from images and videos. In: ECCV. (2010)

[34] Guo, C., Zhang, L.: A novel multiresolution spatiotemporal saliency detection model and its applications in image and video compression. TIP 57 (2010) 1856-186



First row: human labeled masks; Second row: our results; Third row: results of [12]; Fourth row: results of [17] and the fifth row: results of [34]. Our results correctly detect the two fencers instead of the judge passing by.

[12] Seo, H.J.J., Milanfar, P.: Static and space-time visual saliency detection by self-resemblance,Journal of Vision 2009

[17] Rahtu, E., Kannala, J., Salo, M., Heikkilä, J.: Segmenting salient objects from images and videos. In: ECCV. (2010)

[34] Guo, C., Zhang, L.: A novel multiresolution spatiotemporal saliency detection model and its applications in image and video compression. TIP 57 (2010) 1856-186

Experimental Results

Demo

Content

1. Background

2. Modeling the Temporality of Video Saliency

3. **Actionness-assisted Recognition of Actions**

Motivation

- Simple spatial pooling method such as grids does not keep the pertinent structure of various actions.

- Current saliency assisted models lack the explanatory power for the intention of an action and the ability to differentiate animated from inanimated motions.

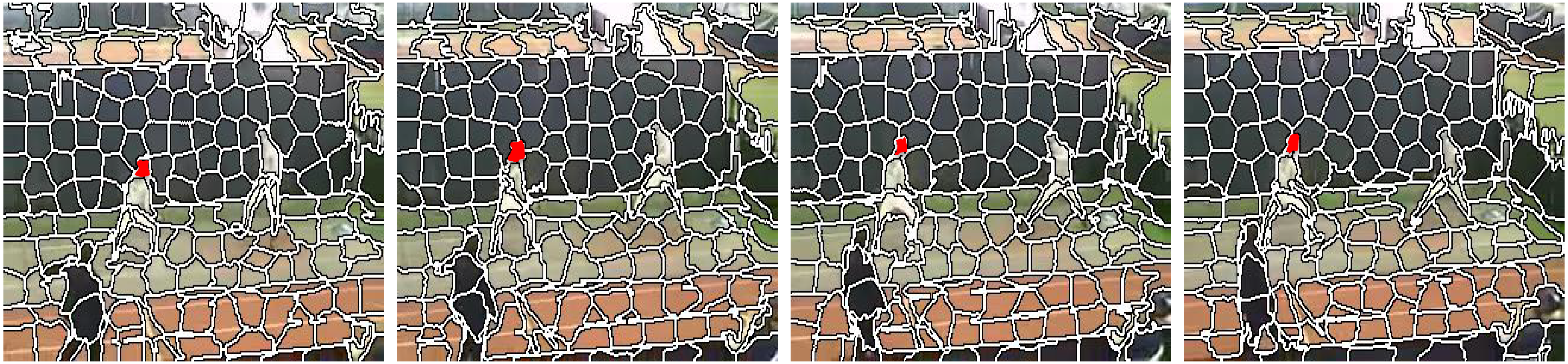
- Some generic low-level features exist and can make various actions stand out of the background.

Main Idea

1. Basic processing unit: a super-pixel trajectory[1]

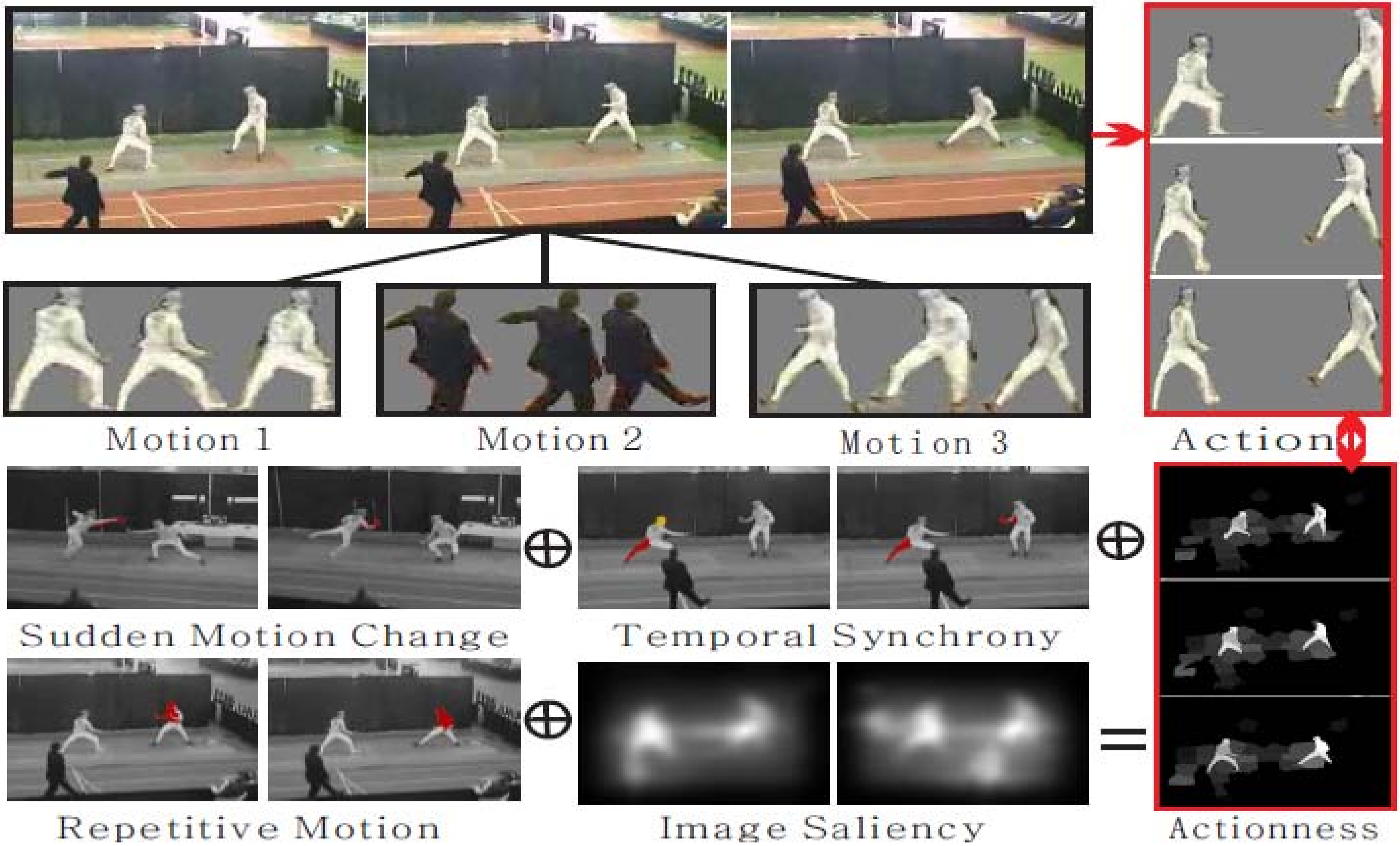
$$Tr = \{R^s, \cdots, R^k, \cdots, R^e\}$$

R is the superpixel (e.g. the red head).



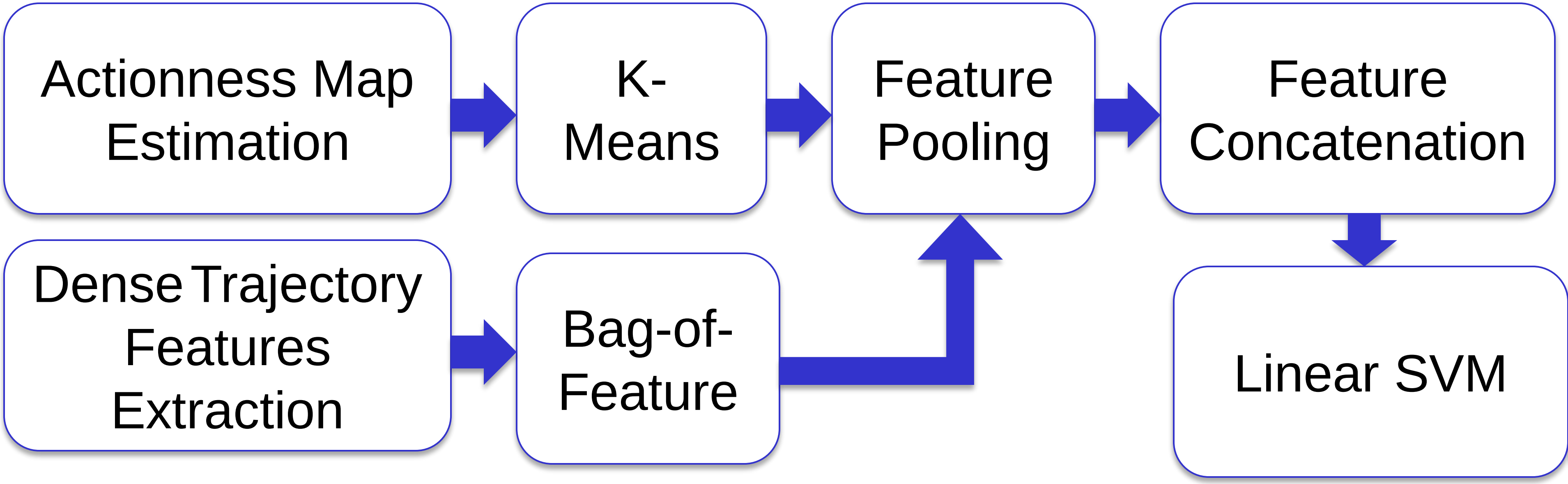
[1] J. Chang, D. Wei, and J. W. Fisher III. A video representation using temporal superpixels. In *CVPR*,2013.

Main Idea: 2. Actionness Map Generation



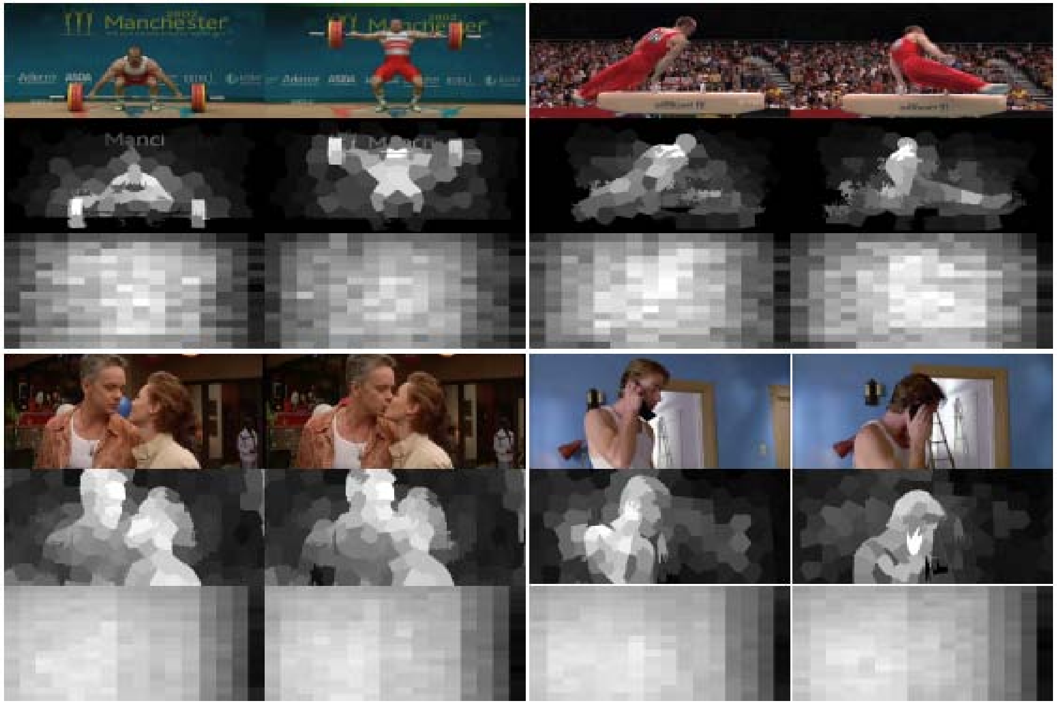
Main Idea Cont.

3. The pipeline of our actionness-driven pooling scheme on action recognition



Experimental Results

1. <u>Action Detection</u>: Mean average precision (mAP) of action Detection on the UCF-Sports and HOHA datasets



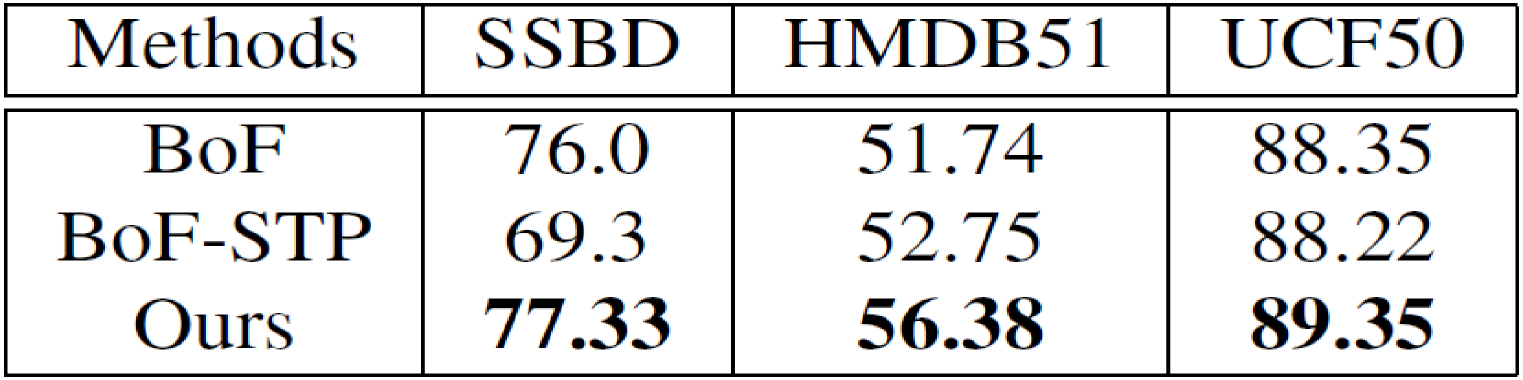
	Our Method	L-CORF [9]	DPM [11]
UCF-Sports	**66.81**	60.8	54.9
HOHA	**70.16**	68.5	60.8

[9] W. Chen, C. Xiong, R. Xu, and J. Corso. Actionness ranking with lattice conditional ordinal random fields. In *CVPR*, 2014.

Experimental Results Cont.

2. Action Recognition

- Comparison with Two Baseline Methods: method with BoF[35] and BoF with Spatial-Temporal pyramid Pooling (BoF-STP) [21].



Methods	SSBD	HMDB51	UCF50
BoF	76.0	51.74	88.35
BoF-STP	69.3	52.75	88.22
Ours	**77.33**	**56.38**	**89.35**

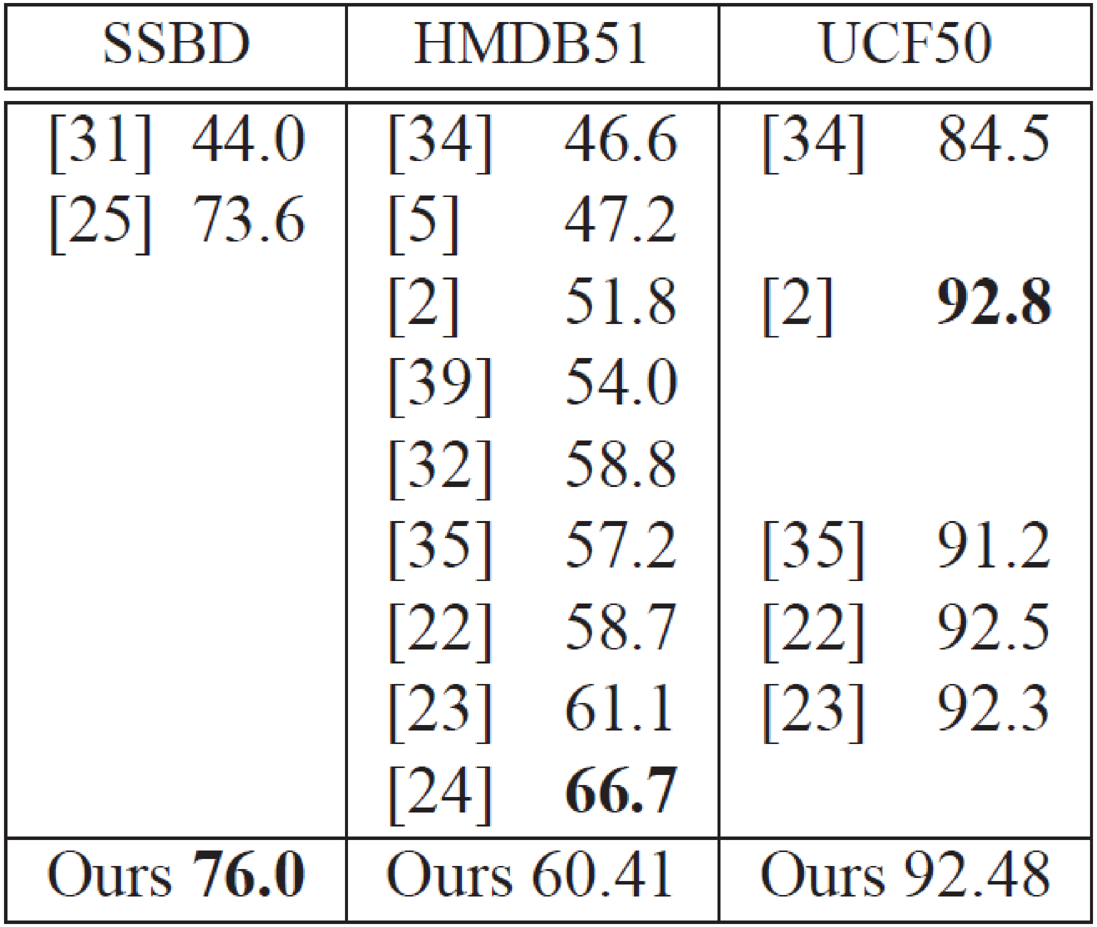
[21] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld. Learning realistic human actions from movies. In *CVPR*, pages 1–8, 2008.

[35] H.Wang and C. Schmid. Action Recognition with Improved Trajectories. ICCV, 2013

Experimental Results Cont.

2. Action Recognition

- Comparison with the State-of-the-art Methods



SSBD		HMDB51		UCF50	
[31]	44.0	[34]	46.6	[34]	84.5
[25]	73.6	[5]	47.2		
		[2]	51.8	[2]	**92.8**
		[39]	54.0		
		[32]	58.8		
		[35]	57.2	[35]	91.2
		[22]	58.7	[22]	92.5
		[23]	61.1	[23]	92.3
		[24]	**66.7**		
Ours	**76.0**	Ours	60.41	Ours	92.48

[2]N. Ballas, Y. Yang, Z.-Z. Lan, B. Delezoide, F. Preteux, and A. Hauptmann. Space-time robust representation for action recognition. In *ICCV, 2013.*

[5] H. Boyraz, S. Masood, B. Liu, M. Tappen, and H. Foroosh.Action recognition by weakly-supervised discriminative region localization. In *BMVC,2014.*

[22] S. Narayan and K. Ramakrishnan. A cause and effect analysis of motion trajectories for modeling actions. In CVPR, 2014.

[23] X. Peng, L. Wang, X. Wang, and Y. Qiao. Bag of Visual Words and Fusion Methods for Action Recognition: Comprehensive Study and Good Practice. ArXiv , 2014.

[24] X. Peng, C. Zou, Y. Qiao, and Q. Peng. Action recognition with stacked fisher vectors. In ECCV, 2014.

[25] S. S. Rajagopalan and R. Goecke. Detecting self-stimulatory behaviours for autism diagnosis. In ICIP, 2014.

[31] S. Sundar Rajagopalan, A. Dhall, and R. Goecke. Selfstimulatory behaviours in the wild for autism diagnosis. In *ICCV Workshops, 2013*

[32]E.Taralova, F de la Torre, and M.Hebert.Motion words for videos. *ECCV, 2014.*

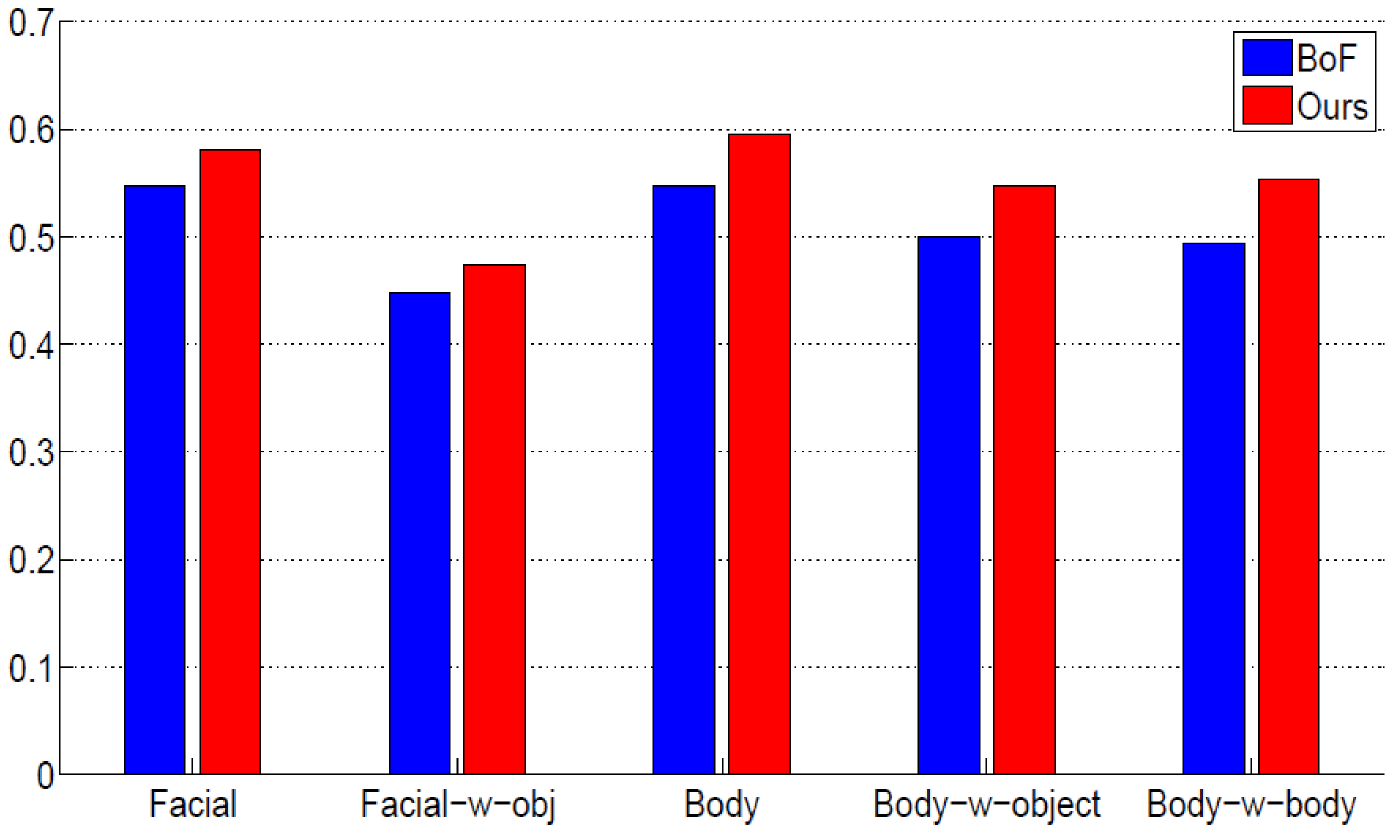
[34]H. Wang, A. Kl¨aser, C. Schmid, and C.-L. Liu. Dense trajectories and motion boundary descriptors for action recognition. *IJCV, 2013.*

[35]H.Wang and C. Schmid. Action Recognition with Improved Trajectories. ICCV, 2013

[39] J. Zhu, B. Wang, X. Yang, W. Zhang, and Z. Tu. Action recognition with actons. In ICCV, 2013

Experimental Results Cont.

3. Performance on Different Types of Actions

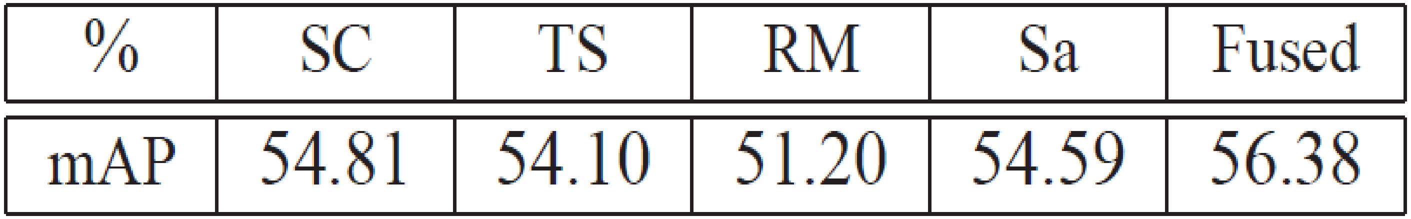


Accuracy Comparisons within HMDB51

Experimental Results Cont.

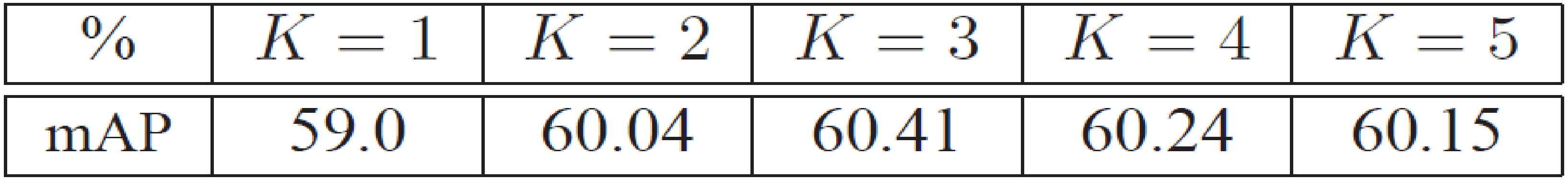
4. Others

Individual attributes comparisons in HMDB51



%	SC	TS	RM	Sa	Fused
mAP	54.81	54.10	51.20	54.59	56.38

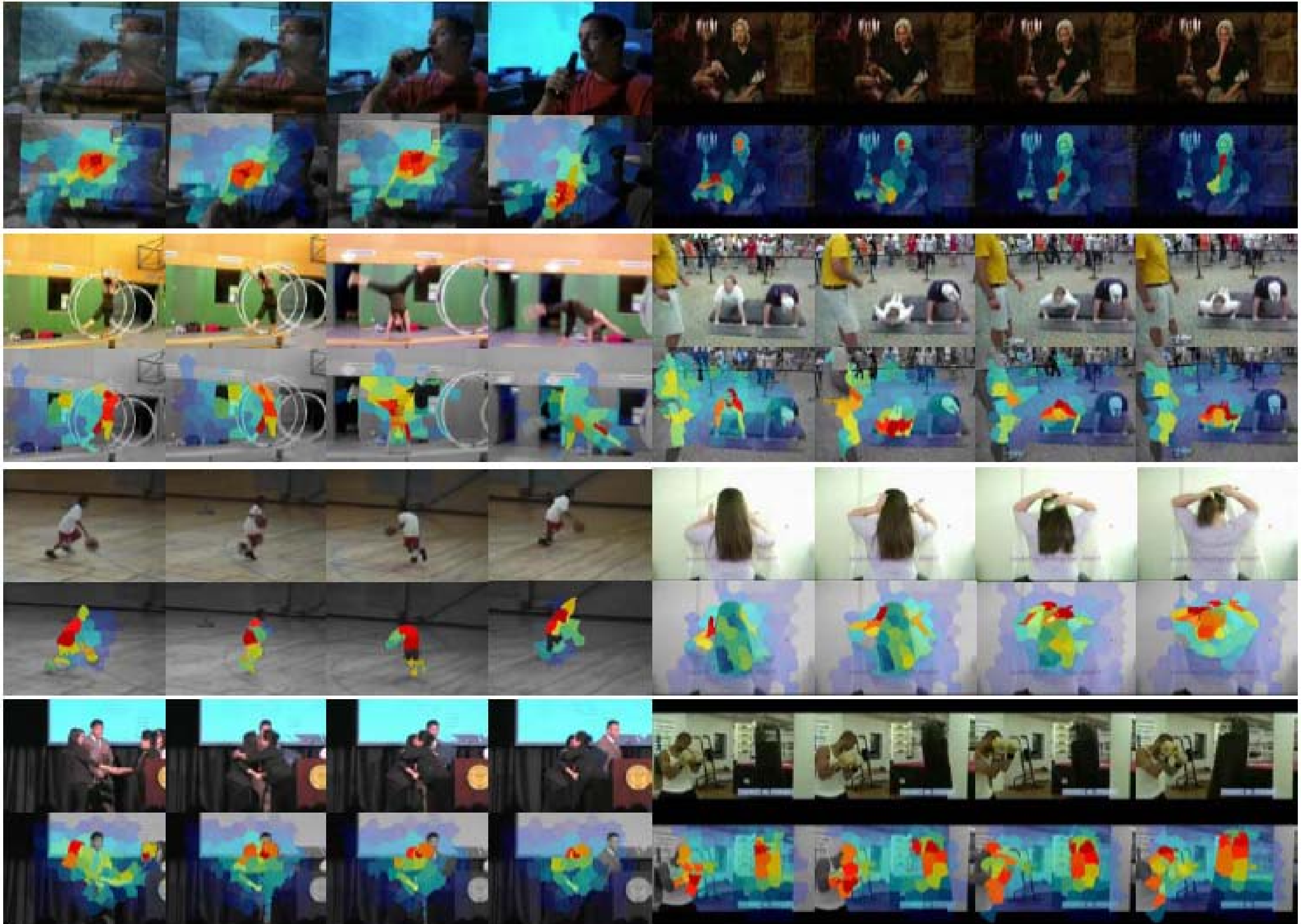
Sensitivity analysis of *K* in HMDB51



%	$K = 1$	$K = 2$	$K = 3$	$K = 4$	$K = 5$
mAP	59.0	60.04	60.41	60.24	60.15

Experimental Results Cont.

Actionness Maps for Various Actions in HMDB51

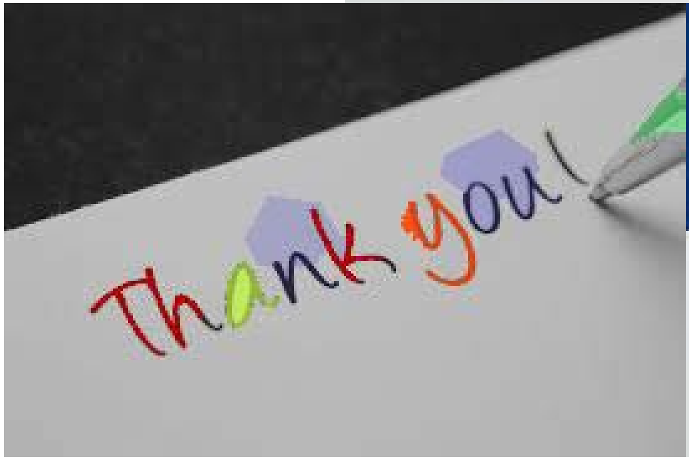


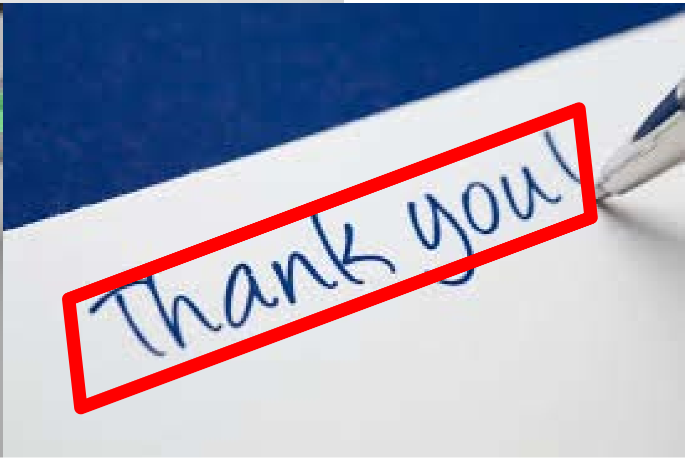
Subjective Results

Demo

Original Image

Saliency Map





Salient Object