Hacettepe University Department of Computer Engineering

Dynamic Eye Movement Datasets an Learnt Saliency Models for Visual Action Recognition,ECCV 2012

Stefan Mathe and Cristian Sminchisescu

Presenter: Levent Karacan

Topics

1. Introduction
2. Related Work
3. Hollywood-2 and UCF Sports Datasets
4. Static and Dynamic Consistency Analysis
5. Learnt Saliency Model
6. Visual Action Recognition
7. Conclusions
8. Summary

Inroduction

Visual processing operate in **«saccade and fixate»** regime so **interest point** based approach consistent with human visual, but knowledge and methodology are distinct. Three contributions are presented related with this issue.

Image from http://*www.scholarpedia.org/article/Eye_movements*

Inroduction

- Some of the most successful approaches to action recognition employ bag-of words representations based on descriptors computed at spatial-temporal video locations.

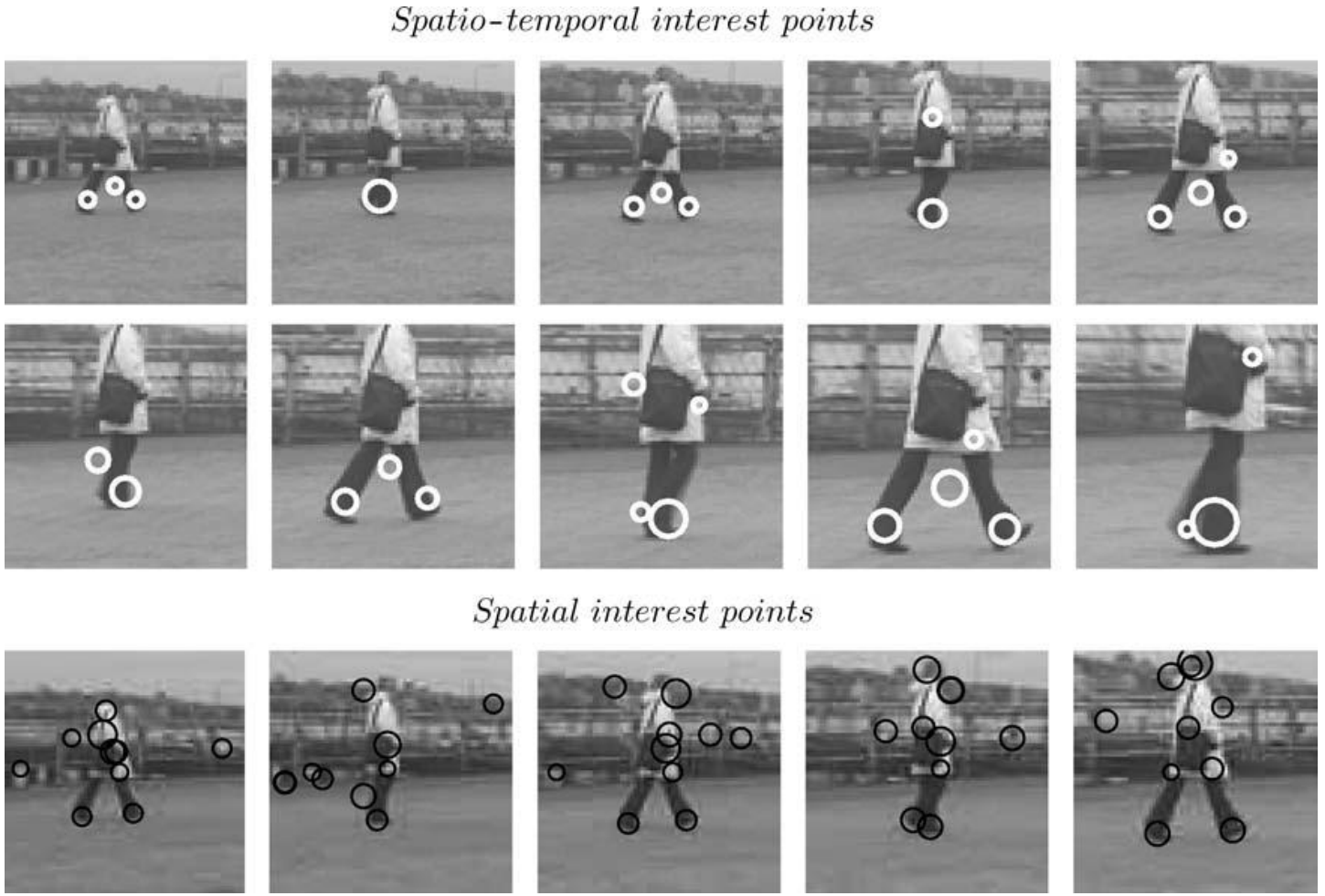
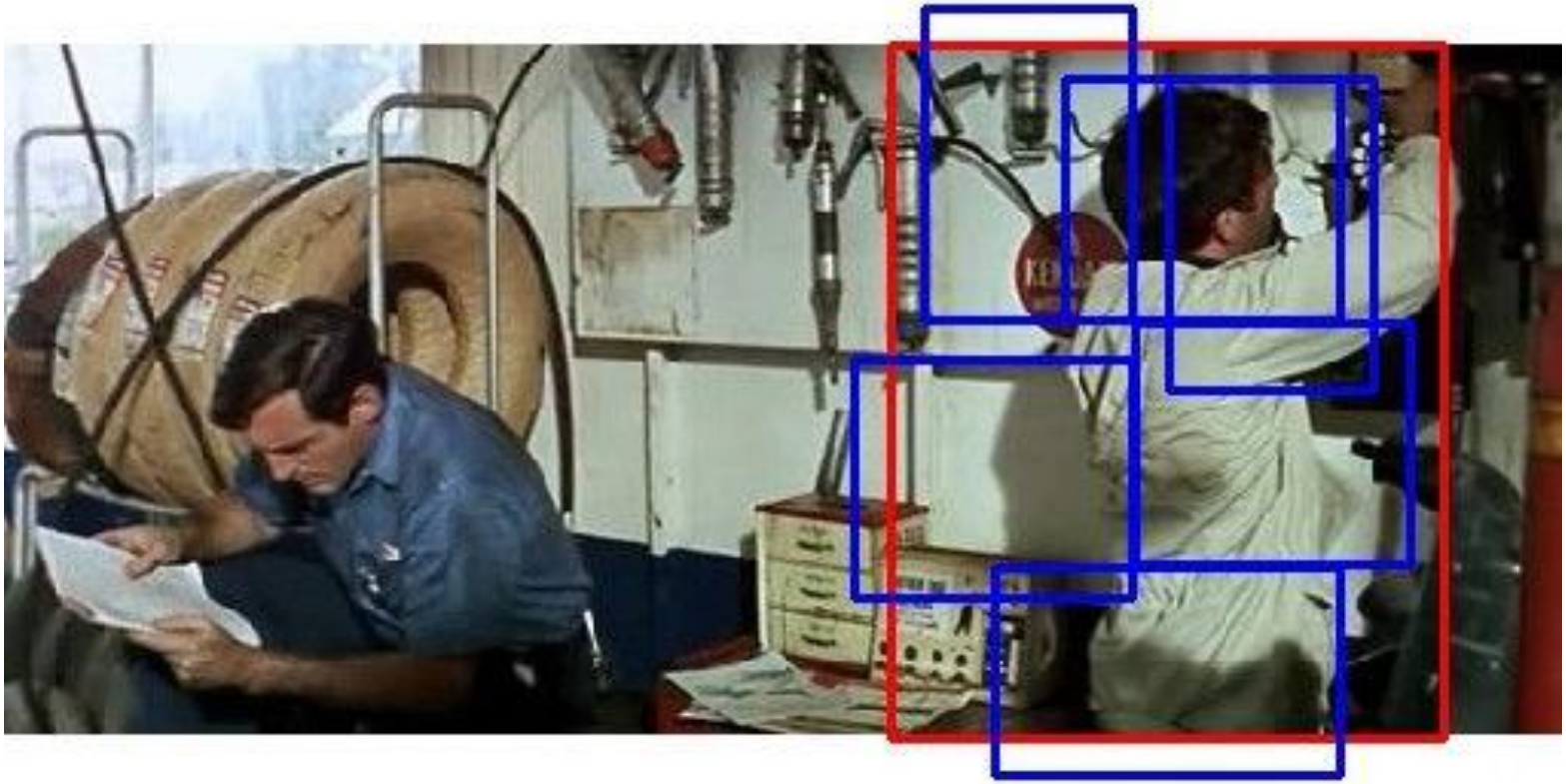


Figure from :Laptev, I.: On space-time interest points. In: IJCV. (2005)

Inroduction

Image representations based on objects and their relations, as well as multiple kernels have been employed with a degree of success.



AnswerPhone Class

Image from :Han, D., Bo, L., Sminchisescu, C.: Selection and context for action recognition.In: ICCV. (2009))

Inroduction

- Can computational insights from a working system like the human visual system can be used to improve performance?
- The effect of using human fixations in conjunction with computer vision algorithms in the context of action recognition have not been yet explored.
- Human eye movement annotations offer pragmatic potential.

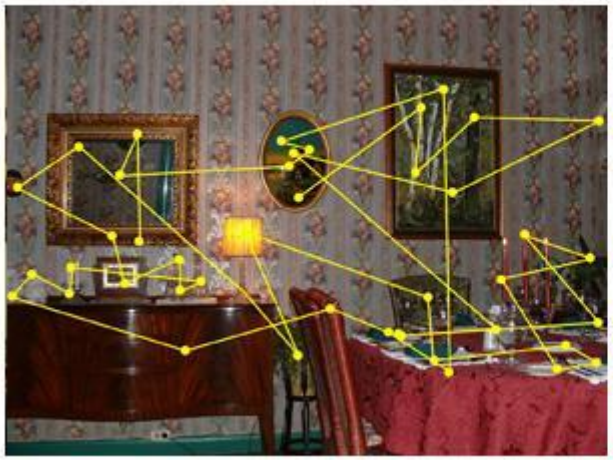
Contributions

- Complementation existing state-of-the art large-scale dynamic computer vision datasets **Hoollywood-2, UCF Sports**
- Novel consistency models and algorithms, as well as relevance evaluation measures adapted for video.
- Trainable computer vision systems based on human eye movements.

Related Works

- Studies carried out by the human vision community use datasets containing human gaze pattern annotations of images.
- These datasets have been designed for small quantitative studies, consisting of at most a few hundred images or videos, usually recorded under **free-viewing**
- The data is provide in this work, which is large-scale, dynamic, and **task controlled**

Memorization Task



Visual Search Task

- Itti, L., Koch, C.: A saliency-based search mechanism for overt and covert shifts of visual attention. Vision Research , (2000)
- Ehinger, K.A., Sotelo, B., Torralba, A., Oliva, A.: Modeling search for people in 900 scenes: A combined source model of eye guidance. Visual Cognition , (2009)
- Judd, T., Durand, F., Torralba, A.: Fixations on low resolution images. ICCV. 2009
- Judd, T., Ehinger, K., Durand, F., Torralba, A.: Learning to predict where humans look. In: ICCV. (2009)
- Kienzle, W., Scholkopf, B., Wichmann, F., Franz, M.: How to nd interesting locations in video: a spatiotemporal interest point detector learned from human eye movements. In: DAGM. (2007)

Image from http://*www.scholarpedia.org/article/Eye_movements*

Related Works

- On eye movement data collection and saliency models for action recognition. Vig, E., Dorr, M., Cox, D.: Space-variant descriptor sampling for action recognition based on saliency and eye movements. In: ECCV. (2012)

- Early biologically inspired recognition system was presented by Kienzle et al:How to find interesting locations in video: a spatiotemporal interest point detector learned from human eye movements. In: DAGM. (2007)

- Currently the most successful systems remain the ones dominated by complex features extracted at interesting locations, bagged and fused using advanced kernel combination techniques. Marszalek, M., Laptev, I., Schmid, C.: Actions in context. In: CVPR. (2009)

- This work a first step in explaining some of the most advanced data and models that human vision and computer vision can offer at the moment.

Datasets

- Hollywood -2 and UCF Sports 21 datasets span approximately hours of video (around 500k frames)
- 12 classes of human actions and 10 classes of scenes distributed over 3669 video clip in Hollywood-2
- 9 classes of human sport actions in UCF Sports

Hoolywood-2

UCF Sports from http://www.cs.ucf.edu/vision/public_html/data.html
Hollywood-2 from http://www.di.ens.fr/~laptev/actions/hollywood2/

Datasets

- Objective of this work is to add additional annotations to these datasets.
- Task control is done as opposed to free-viewing.
- To carry out this objective, two subject groups are included to study: an active group (12 subjects), asked to perform an action recognition task and a free viewing group (4 subjects)
- Studies in the human vision community have advocated a high degree of agreement between human gaze patterns for subjects queried under static stimuli.
- An investigation is done whether this effect extends to dynamic data.

Static Consistency Among Subjects

- How well the regions fixated by human subjects agree on a frame by frame basis?
- **Evaluation Protocol :** For the task of locating people in a static image, one can evaluate how well the regions fixated by a particular subject can be predicted from the regions fixated by the other subjects on the same image[1].

- Shooter bias and center bias affect this measure.

- One can address this issue by checking how well the fixations of a subject on one stimulus can be predicted from those of the other subjects on a different, unrelated, stimulus.

- For videos evaulation is done inter-subject correlation on randomly chosen frames.

Ehinger, K.A., Sotelo, B., Torralba, A., Oliva, A.: Modeling search for people in 900 scenes: A combined source model of eye guidance. Visual Cognition 17 (2009)

Static Consistency Among Subjects

- A probability distribution is generated by adding a **Dirac impulse** at each pixel fixated by the other subjects followed by blurring with a Gaussian kernel.

- The probability at the pixel fixated by the subject is taken as the prediction for its fixation.

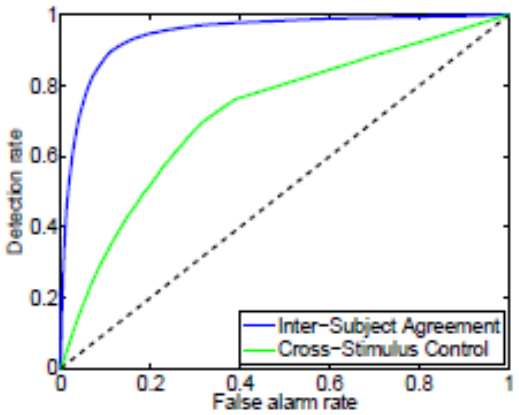
- For control this process is repeated for pairs of frames chosen from different videos and predict the fixation of each subject on one frame from the fixations of the other subjects on the other frame.

consistency measure	agreement	control
static consistency	94.8%	72.3%
temporal AOI alignment	70.8%	51.8%
AOI Markov Dynamics	70.2%	12.7%

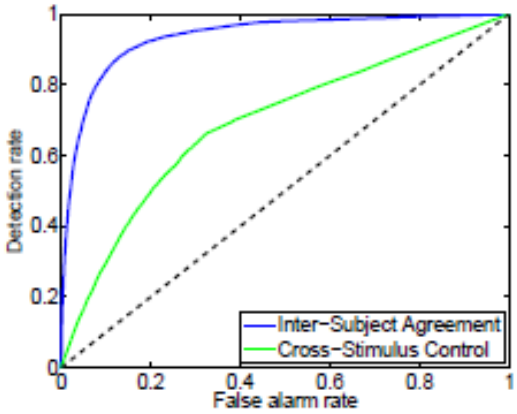
Hollywood2

consistency measure	agreement	control
static consistency	93.2%	69.2%
temporal AOI alignment	65.4%	30.4%
AOI Markov Dynamics	55.5%	12.9%

UCF Sports



Hollywood2

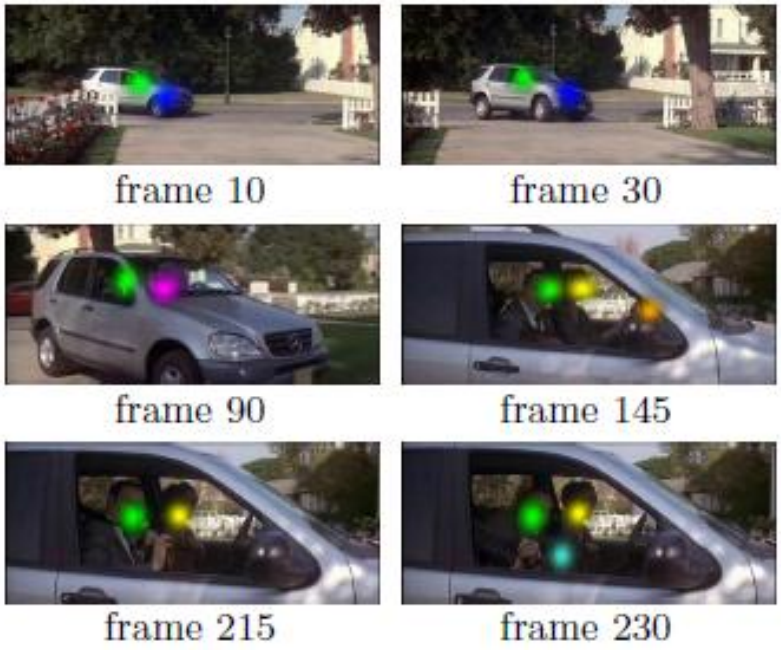


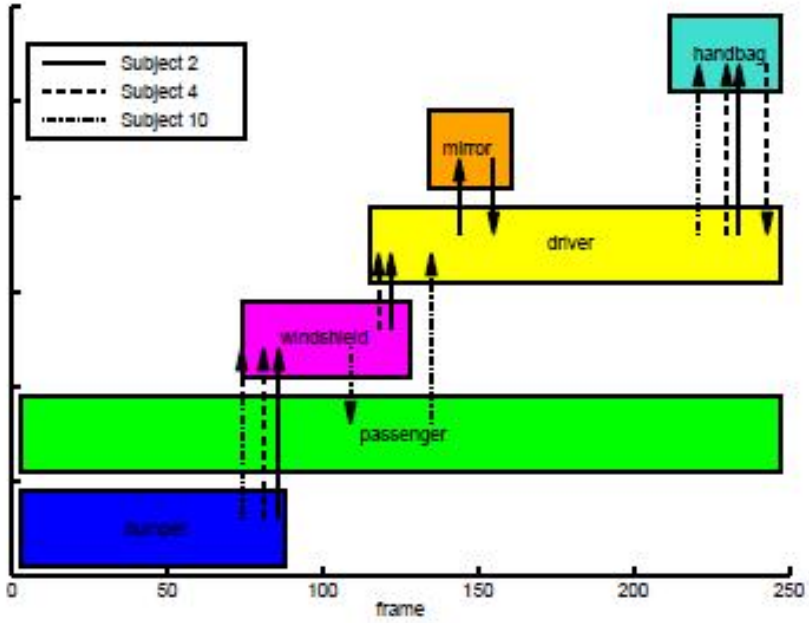
UCF Sports

The Influence of Task on Eye Movements

- For a given video frame and for each free viewing subject, the p-statistic is computed at the fixated location with respect to the probability distribution derived using fixation data from active subjects.
- This process is repeated for 1000 randomly sampled frames and the average p-value is computed for each subject.
- Fixation patterns of free viewers do not deviate significantly from those of active subjects ($p = 0:65$ for Hollywood-2 and $p = 0:58$ for UCF Sports).
- Since in the Hollywood-2 dataset several actions can be present in a video, either simultaneously or sequentially, this rules out initial habituation effects and further neglect(free viewing) to some degree.

Dynamic Consistency Among Subjects





- The spatial distribution of fixations in video is highly consistent across subjects..

- Two metrics AOI Markov dynamics and temporal AOI alignment that are sensitive to the temporal ordering among fixations and evaluate consistency

Learnt Saliency Models for Visual Action Recognition

Human Saliency Map Prediction

- Analysis includes many features derived directly from low, mid and high level image information. In addition, we train our own detector that fires preferentially at the locations fixated by the subjects, using the vast amount of eye movement data available in the dataset.
- We evaluate all these features and their combinations on our dataset under two metrics AUC and KL divergence

Human Saliency Map Predictors

Saliency map predictors on dataset:

- Baselines

  -The uniform saliency map

  -The center bias (CB) feature,

  -Human saliency predictor, which computes the saliency map derived from the fixations made by half of human subjects.

- Static Features(SF):Image Features

  -Low:color, steerable filter subbands, feature maps

  -Mid:(Horizon detector) and

  -High(Object detectors)

Human Saliency Map Predictors

Saliency map predictors on dataset:

- Motion Features:Features maps derived motion or space time information
  - -Flow
  - -Pb with flow
  - -Flow bimodality
  - -Harris
  - -HoG-MBH detector

- Feature Combinations

Saliency Predictor Evaluation with AUC and KL divergence

- When training our HoG-MBH detector, we use 106 training examples, half of which are positive and half of which are negative.
- At each of these locations, spatio-temporal HoG and MBH descriptors are extracted, with various grid configurations.
- These descriptors are concatenated and the resulting vector is lifted into a higher dimensional space by employing an order 3 2 kernel approximation[31].
- Training process is done with a linear SVM to obtain detector.

baselines		
feature	AUC (a)	KL divergence (b)
uniform baseline	0.500	18.63
central bias (CB)	0.840	15.93
human	0.936	10.12
static features (SF)		
color features [4]	0.644	17.90
subbands [32]	0.634	17.75
Itti&Koch channels [10]	0.598	16.98
saliency map [27]	0.702	17.17
horizon detector [27]	0.741	15.45
face detector [29]	0.579	16.43
car detector [30]	0.500	18.40
person detector [30]	0.566	17.13

our motion features (MF)		
feature	AUC (a)	KL divergence (b)
flow magniture	0.626	18.57
pb edges with flow	0.582	17.74
flow bimodality	0.637	17.63
Harris cornerness	0.619	17.21
HOG-MBH detector	0.743	**14.95**
feature combinations		
SF [4]	0.789	16.16
SF + CB [4]	0.861	15.96
MF	0.762	15.62
MF + CB	0.830	15.97
SF + MF	0.812	15.94
SF + MF + CB	**0.871**	15.89

Visual Action Recognition Pipeline

An automatic end-to-end action recognition system based on predicted saliency maps, together with several baselines:

- **Interest Point Operator:** various interest point operators, both computer vision based and biologically derived are used. Each interest point operator takes as input a video and generates a set of spatio-temporal coordinates, with associated spatial and temporal scales
- **Descriptors:** the spacetime generalization of the HoG descriptor are used as well as the MBH descriptor computed from optical flow.

- **Visual Dictionaries:** The resulting descriptors are clustered using k-means into vocabularies of 4000 visual words. For computational reasons, only 500k randomly sampled descriptors are used as input to the clustering step and then each video is represented by the L1 normalized histogram of its visual words.

- **Classifiers:** Multiple Kernel Learning (MKL) framework is used to compute kernel matrices. Training process is done for each action label in a one-vs-all fashion.

Visual Action Recognition Pipeline

Eye Movement Datasets and Saliency Models for Action Recognition

action	interest points						trajectories + interest points		
	Harris corners (a)	uniform sampling (b)	central bias sampling (c)	predicted saliency sampling (d)	ground truth saliency sampling (e)	trajectories only (f)	uniform sampling (g)	predicted saliency sampling (h)	ground truth saliency sampling (i)
AnswerPhone	16.4%	21.3%	23.3%	23.7%	**28.1%**	**32.6%**	24.5%	25.0%	32.5%
DriveCar	85.4%	92.2%	92.4%	92.8%	**57.9%**	88.0%	93.6%	93.6%	**96.2%**
Eat	59.1%	59.8%	58.6%	**70.0%**	67.3%	65.2%	69.8%	**75.0%**	73.6%
FightPerson	71.1%	74.3%	76.3%	76.1%	**80.6%**	81.4%	79.2%	78.7%	**83.0%**
GetOutCar	36.1%	47.4%	49.6%	54.9%	**55.1%**	52.7%	55.2%	**60.7%**	59.3%
HandShake	18.2%	25.7%	26.5%	**27.9%**	27.6%	**29.6%**	29.3%	28.3%	26.6%
HugPerson	33.8%	33.3%	34.6%	**39.5%**	37.8%	**54.2%**	44.7%	45.3%	46.1%
Kiss	58.3%	61.2%	62.1%	61.3%	**66.4%**	65.8%	66.2%	66.4%	**69.5%**
Run	73.2%	76.0%	77.8%	82.2%	**85.7%**	82.1%	82.1%	84.2%	**87.2%**
SitDown	54.0%	59.3%	62.1%	**69.0%**	62.5%	62.5%	67.2%	**70.4%**	68.1%
SitUp	26.1%	20.7%	20.9%	29.7%	**30.7%**	20.0%	23.8%	**34.1%**	32.9%
StandUp	57.0%	59.8%	61.3%	**63.9%**	58.2%	65.2%	64.9%	**69.5%**	66.0%
Mean	49.1%	52.6%	53.7%	57.6%	57.9%	58.3%	58.4%	61.0%	**61.7%**

Conclusion

- Comprehensive human eye-tracking annotations for Hollywood-2 and UCF Sports.
- The paper introduce novel saliency detectors and show that they can be trained effectively to predict human fixations as measured under both average precision (AP), and Kullblack-Leibler spatial comparison measures.
- Accurate saliency operators can be effectively trained based on human fixations.
- Automatic saliency predictors can be used within end-to-end computer-based visual action recognition systems to achieve state of the art results.

Summary

- Dataset are used in literature may have some constraints.
- A model must be tested on appropriate dataset.
- Human visual system is excellent working system, we can benefit more our computer vision problems.
- Fixations provide a sufficiently high-level signal that can be precisely registered with the image stimuli, for testing hypotheses and for training visual feature extractors and recognition models quantitatively.
- A saliency model can be learnt from eye movement informations as well as image information.
- Visual action recognition can be done succesfully via a good saliency detector.

THE END

Thanks for your attention