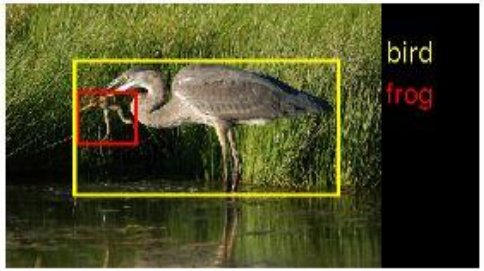
Actions in the Eye: Dynamic Gaze Datasets and Learnt Saliency Models for Visual Recognition

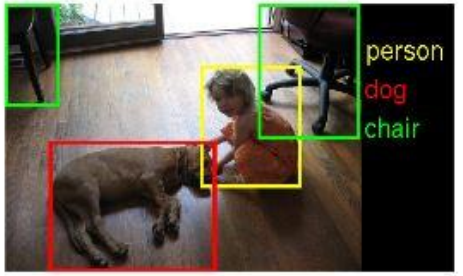
Stefan Mathe, Cristian Sminchisescu

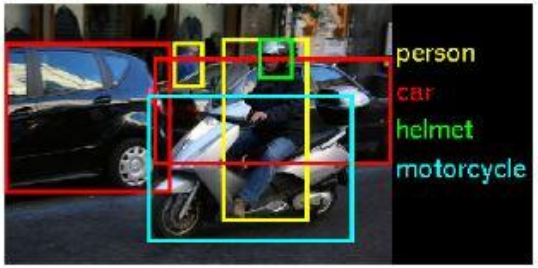
Presented by Mit Shah

Motivation...

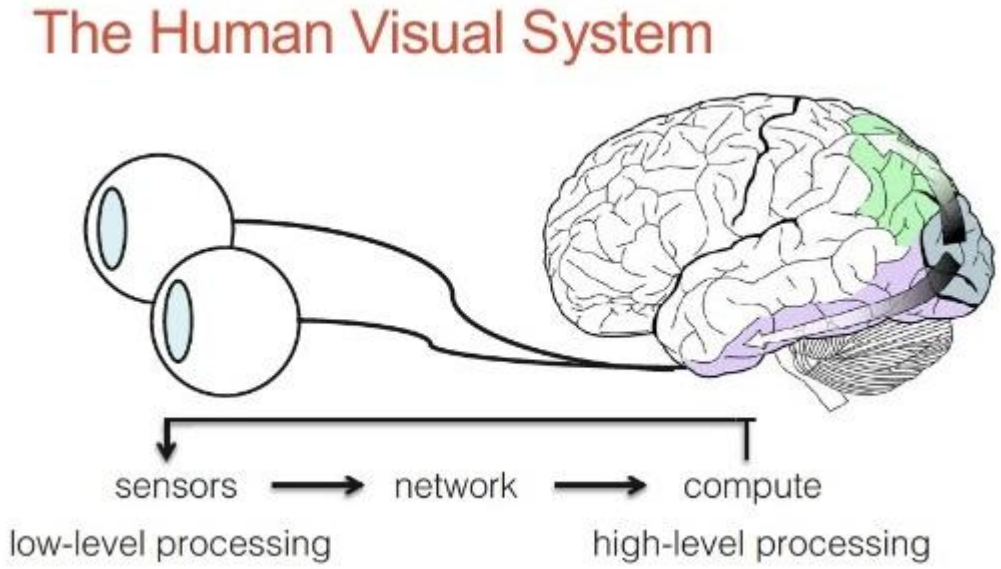
- Current Computer Vision
  - Annotations subjectively defined
  - Intermediate levels of computation??







Motivation...



- Lack of large scale datasets that provide recordings of the workings of the human visual system

Previous Work...

- Study of Gaze patterns in Humans

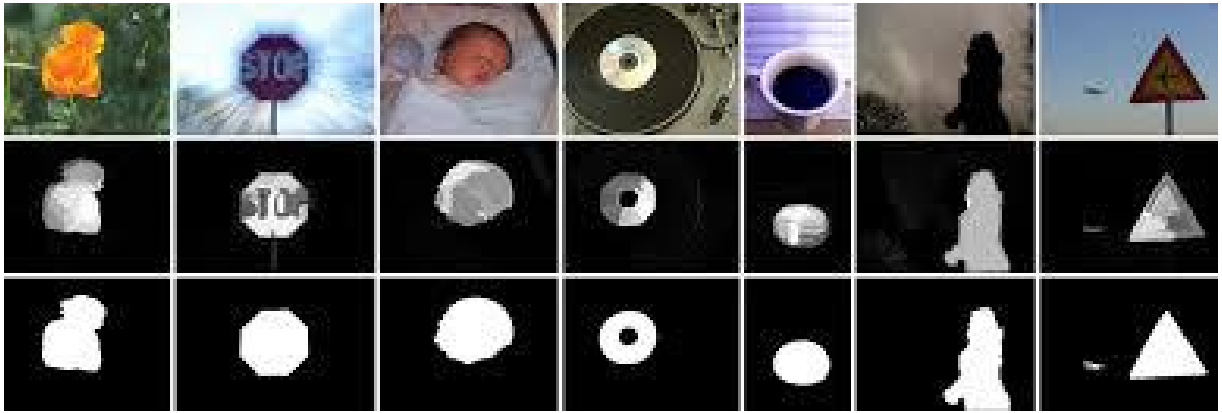
A person browsing reddit with the F-shaped pattern

Previous Work...

- Study of Gaze patterns in Humans
  - **Inter-observer consistency**

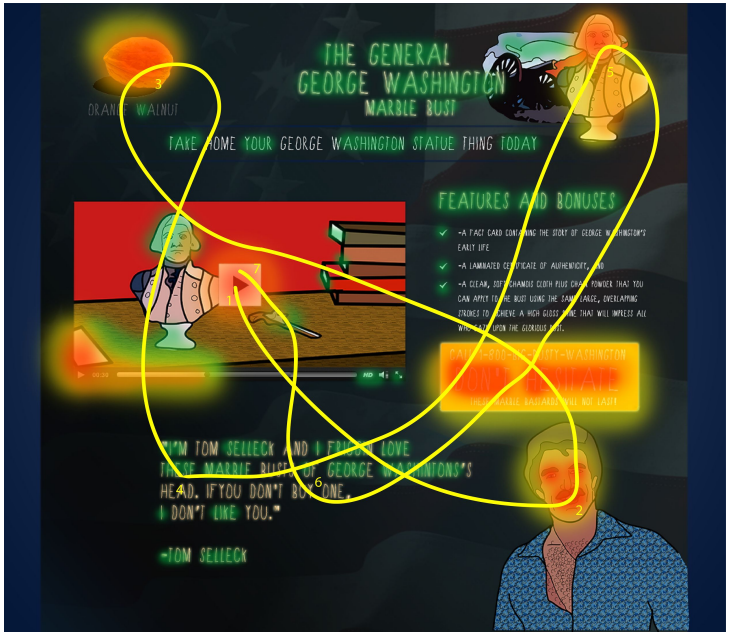
Previous Work...

- Study of Gaze patterns in Humans
  - Inter-observer consistency
  - **Bottom-up Features**



Previous Work...

- Study of Gaze patterns in Humans
  - Inter-observer consistency
  - Bottom-up Features
  - **Human Fixations**

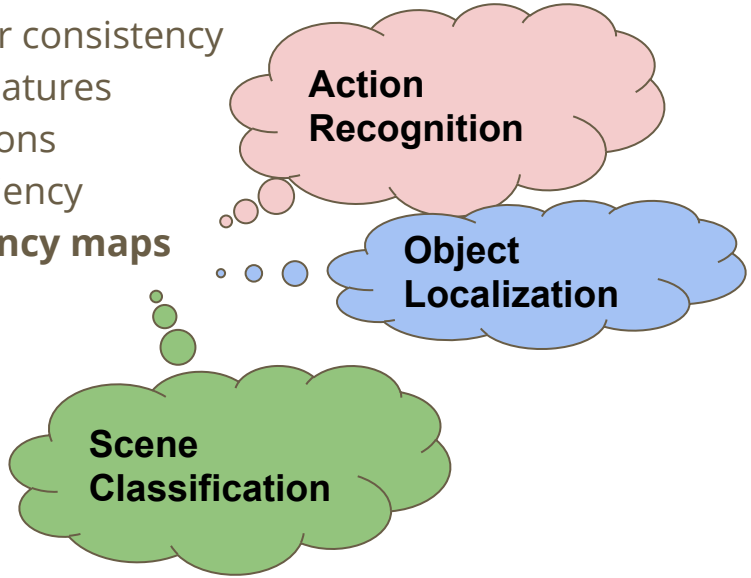


Previous Work...

- Study of Gaze patterns in Humans
  - Inter-observer consistency
  - Bottom-up Features
  - Human Fixations
  - **Models of saliency**

Previous Work...

- Study of Gaze patterns in Humans
  - Inter-observer consistency
  - Bottom-up Features
  - Human Fixations
  - Models of saliency
  - **Uses of Saliency maps**



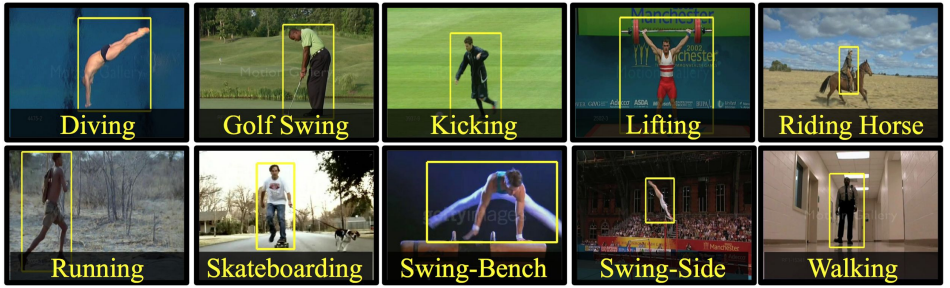
Previous Work...

- Study of Gaze patterns in Humans
  - Inter-observer consistency
  - Bottom-up Features
  - Human Fixations
  - Models of saliency
  - Uses of Saliency maps
  - **Previous data sets**

At most few hundred videos recorded under **free viewing conditions**

Contributions… (1)

- Extended existing large scale datasets **Hollywood-2** and **UCF Sports**



Contributions… (2)

- Dynamic consistency and alignment measures

AOI Markov Dynamics

Temporal AOI Alignment

Contributions... (3)

- Training an **End-to-End automatic visual action recognition system**

Data Collection...

Hollywood-2 Movie Dataset

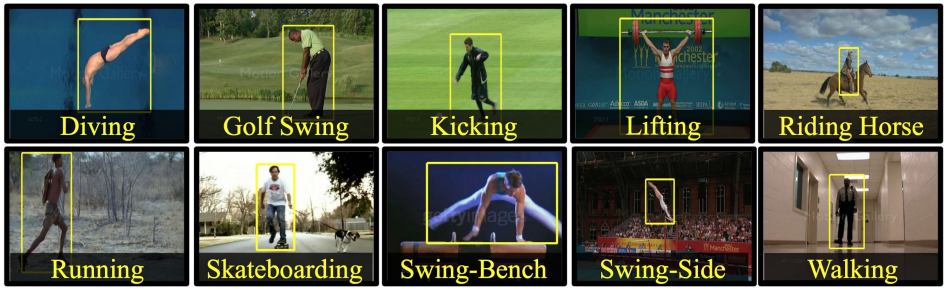
Largest and Most challenging dataset

12 classes
69 movies
823/884 split
487k frames
20 hr

Answering phone, driving a car, eating, fighting, etc.

Data Collection...

UCF Sports Action Dataset



- Broadcast of television channels
- 150 videos covering 9 sports action classes
- Diving, golf swinging, kicking, etc..

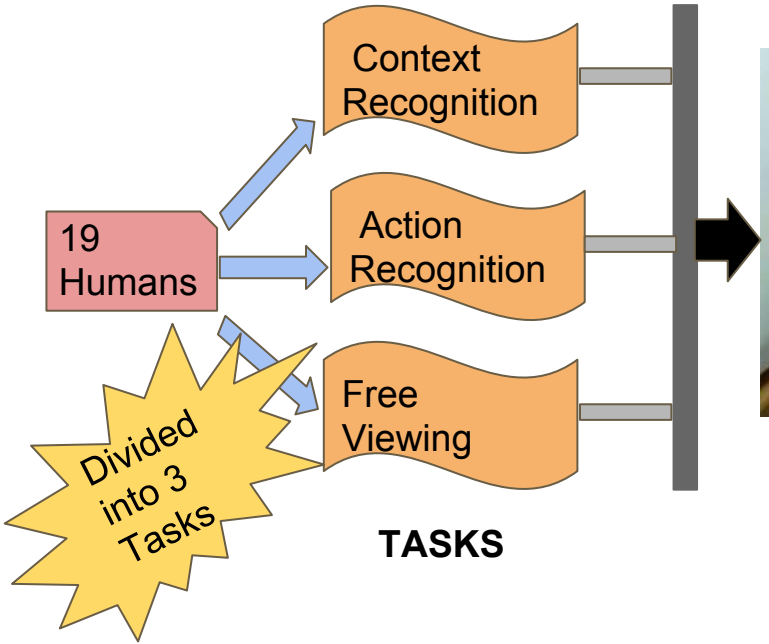
Data Collection...

Extending the two data sets

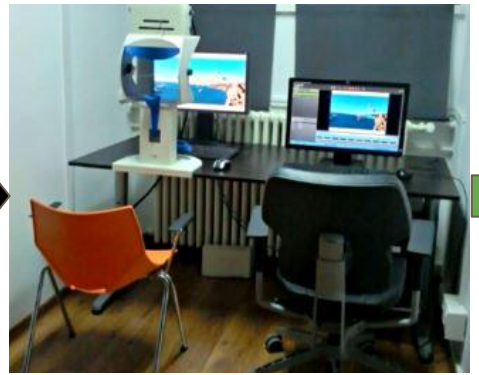
Many other Specifications

Timings/Durations & Breaks

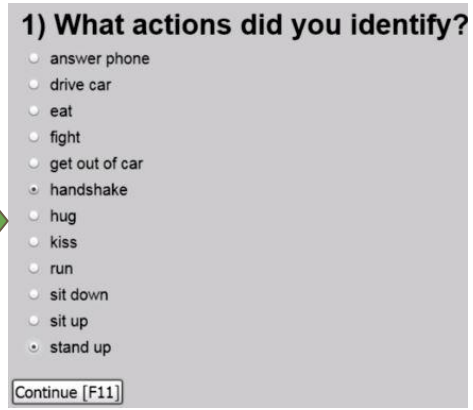
[SMI iView X HiSpeed 1250 Tower-Mounted Eye Tracker]



TASKS



Recording Environment

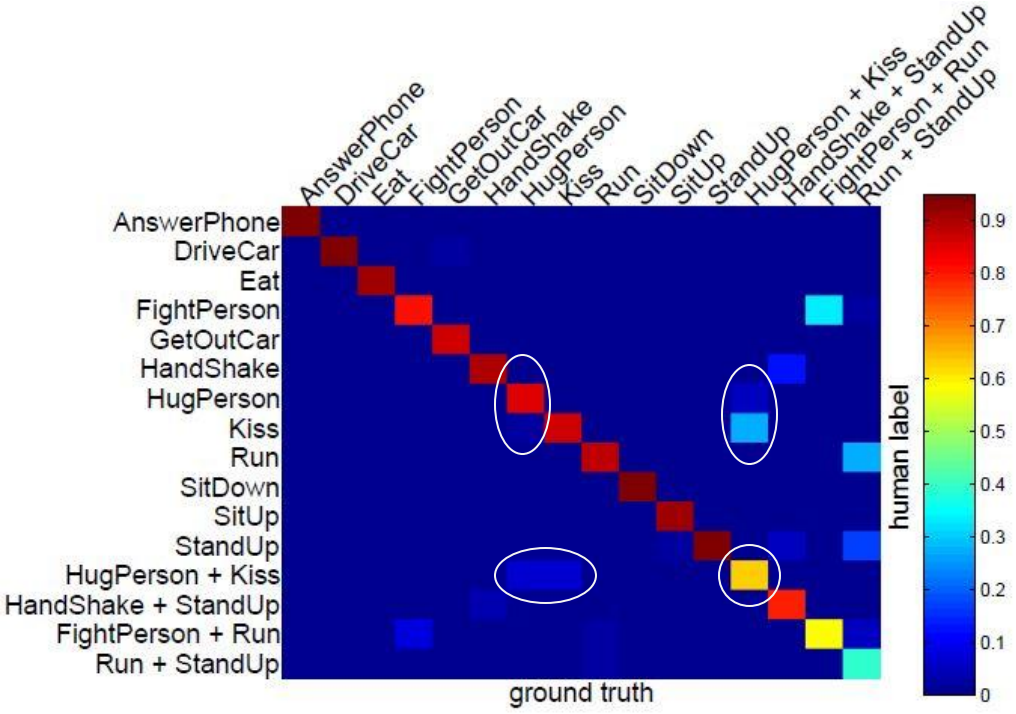


Recording Protocol

Static & Dynamic Consistency

Action Recognition by Humans

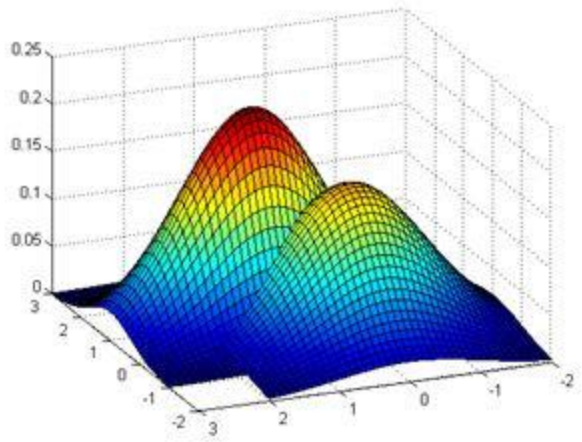
- Goal & Importance
- Human errors
  - Co Occurring Actions
  - False Positives
  - Mislabeling Videos



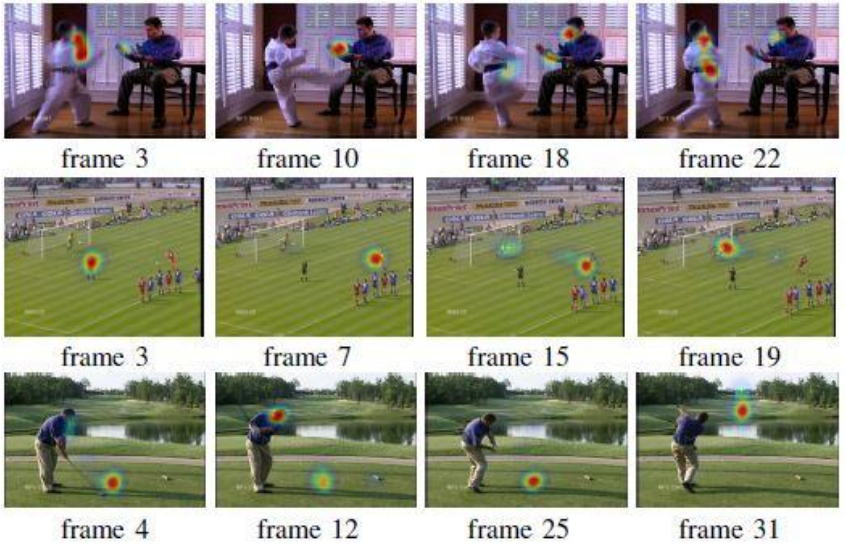
Static Consistency Among Subjects

- How well the regions fixated by human subjects agree on a frame by frame basis?

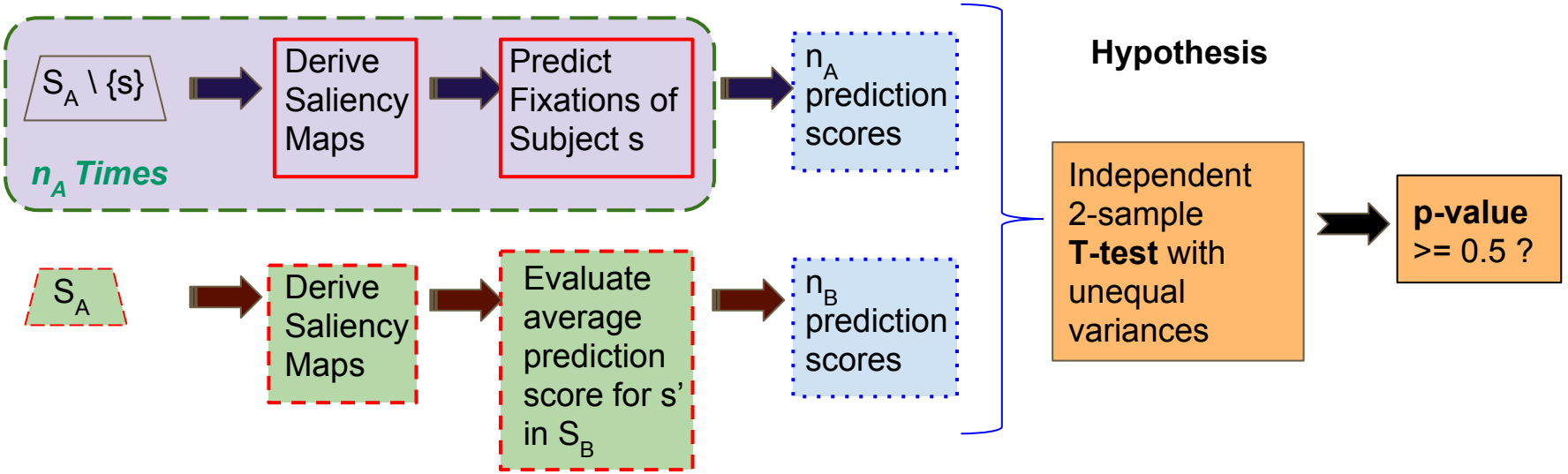
- Evaluation Protocol



Static Consistency Among Subjects



The Influence of Task on Eye Movements



The Influence of Task on Eye Movements

Results -

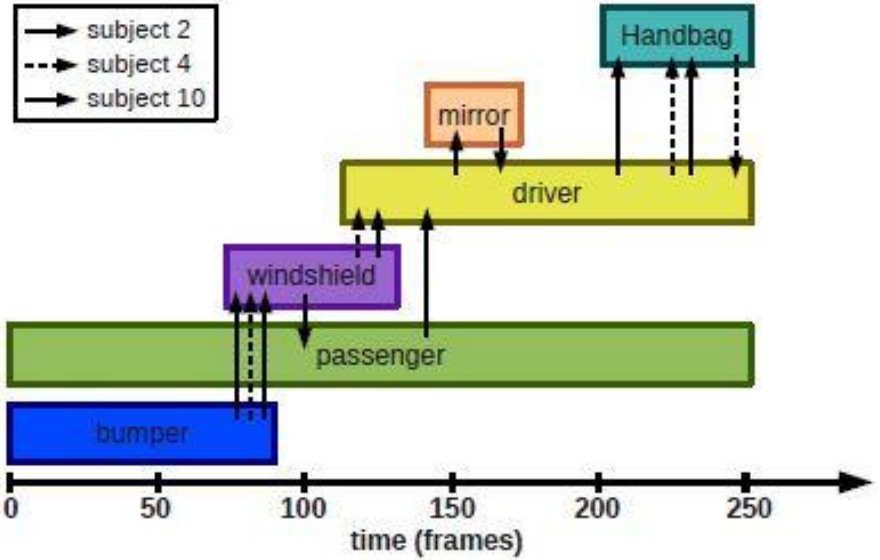
dataset	task A	task B	p-value	significant ($p < 0.05$)
Hollywood-2	action recognition	free viewing	0.14	no
Hollywood-2	action recognition	context recognition	0.01	yes
UCF Sports	action recognition	free viewing	0.75	no
UCF Sports	action recognition	context recognition	0.04	yes

Dynamic Consistency Among Subjects

- Spatial distribution - highly consistent
- Significant consistency in the order also??
- Automatic Discovery of AOIs & 2 metrics
  - AOI Markov dynamics
  - Temporal AOI alignment

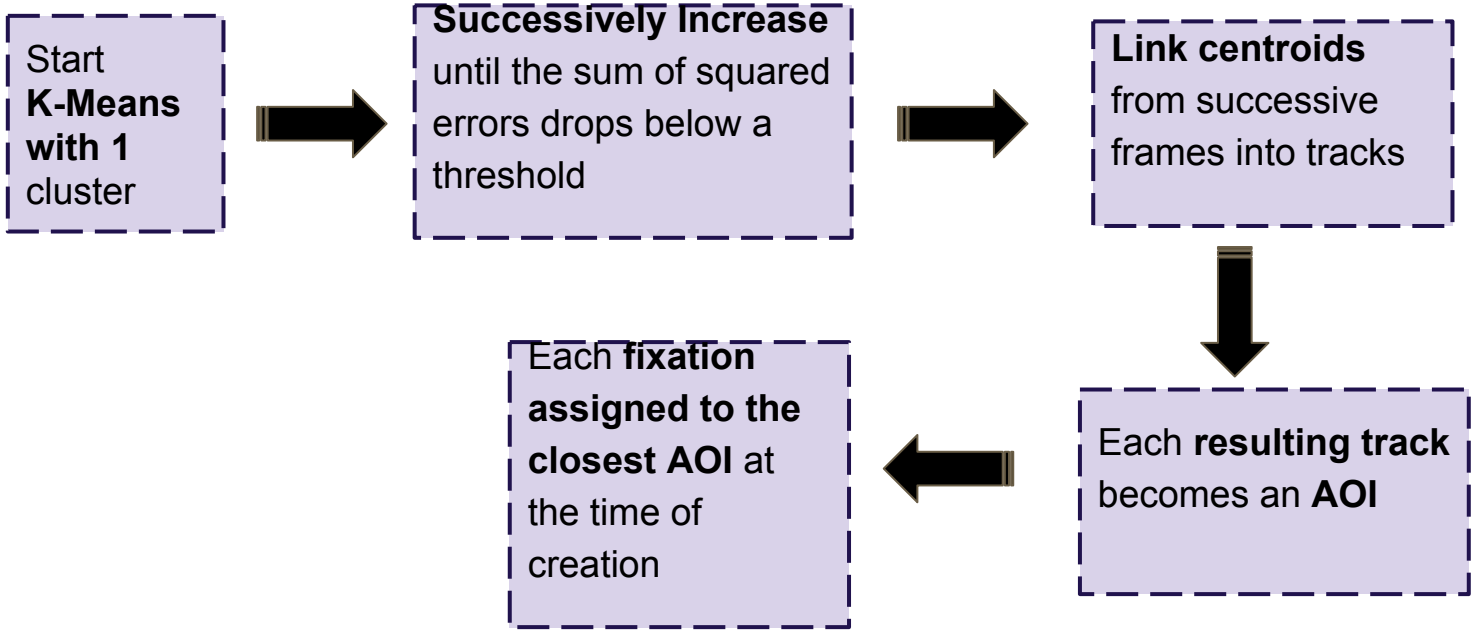
Scanpath representation

- Human fixations - tightly clustered
- Assigning to closest AOI
- Trace the scan path



Automatically Finding AOIs

- Clustering the fixations of all subjects in a frame



Automatically Finding AOIs

frame 10

frame 30

frame 90

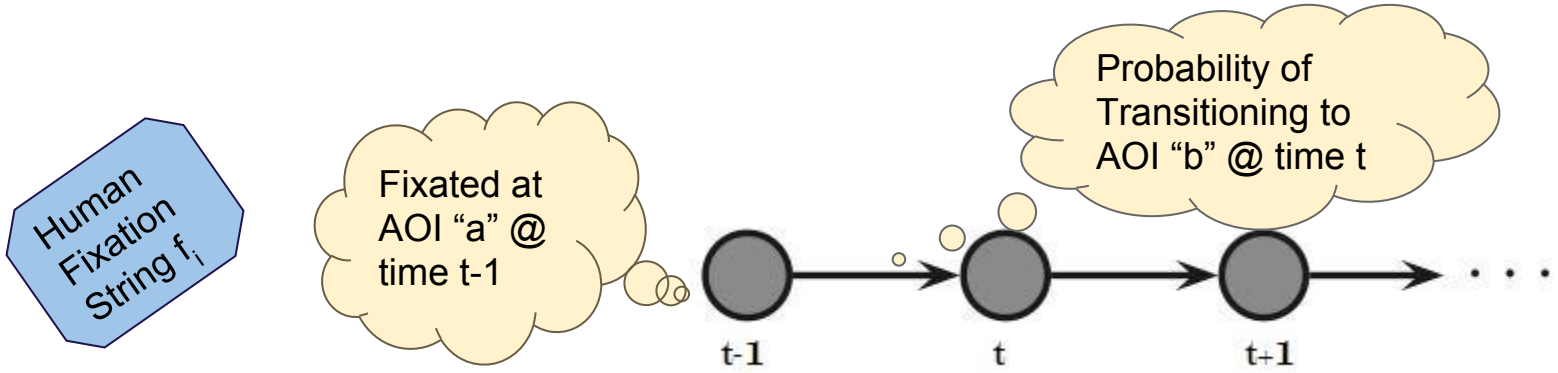
frame 145

frame 215

frame 230

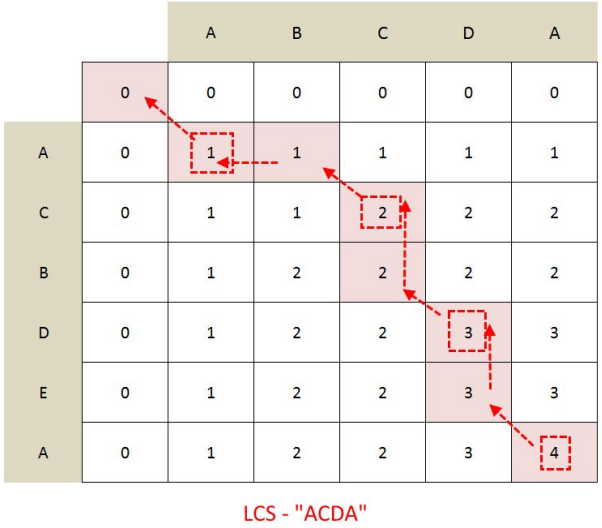
AOI Markov Dynamics

- Transitions of human visual attention between AOIs by..



Temporal AOI Alignment

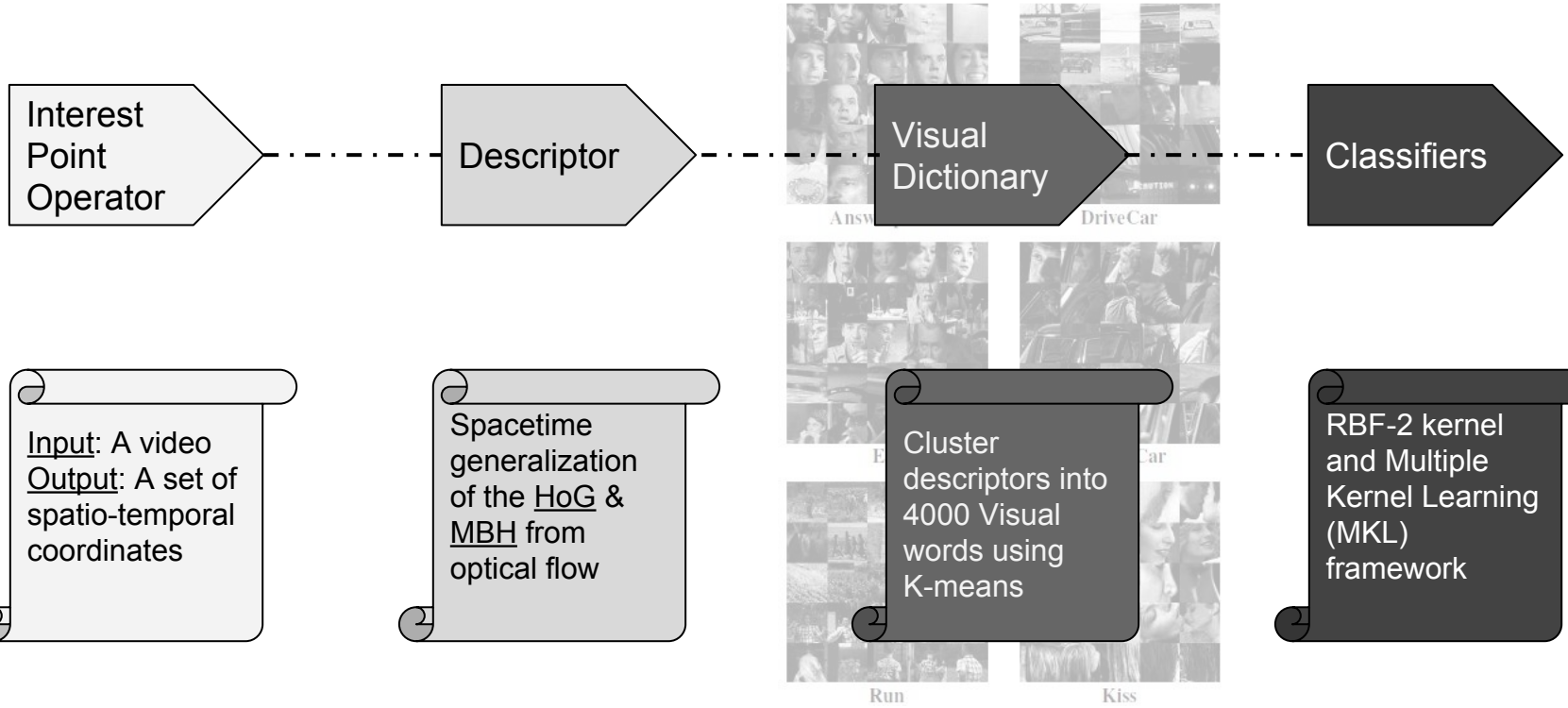
- Longest Common Subsequence??
- Able to handle gaps and missing elements



		A	B	C	D	A
	0	0	0	0	0	0
A	0	1	1	1	1	1
C	0	1	1	2	2	2
B	0	1	2	2	2	2
D	0	1	2	2	3	3
E	0	1	2	2	3	3
A	0	1	2	2	3	4

LCS - "ACDA"

Evaluation Pipeline



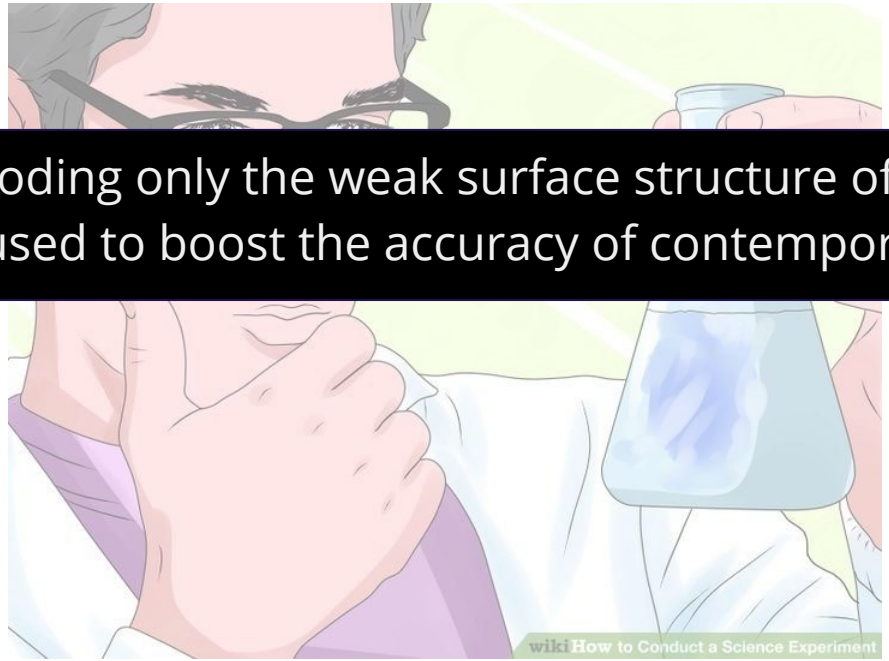
Human Fixation Studies

Human vs. Computer Vision Operators

- Fixations as interest point detector
- Findings
  - Low correlation
  - Why??

action	percent of human fixated spacetime Harris corners (a)
AnswerPhone	6.2%
DriveCar	5.8%
Eat	6.4%
FightPerson	4.6%
GetOutCar	6.1%
HandShake	6.3%
HugPerson	4.6%
Kiss	4.8%
Run	6.0%
SitDown	6.2%
SitUp	6.3%
StandUp	6.0%
Mean	5.8%

Impact of Human Saliency Maps for Computer Visual Action Recognition



Saliency maps encoding only the weak surface structure of fixations (no time ordering), can be used to boost the accuracy of contemporary methods

Saliency Map Prediction

Static Features

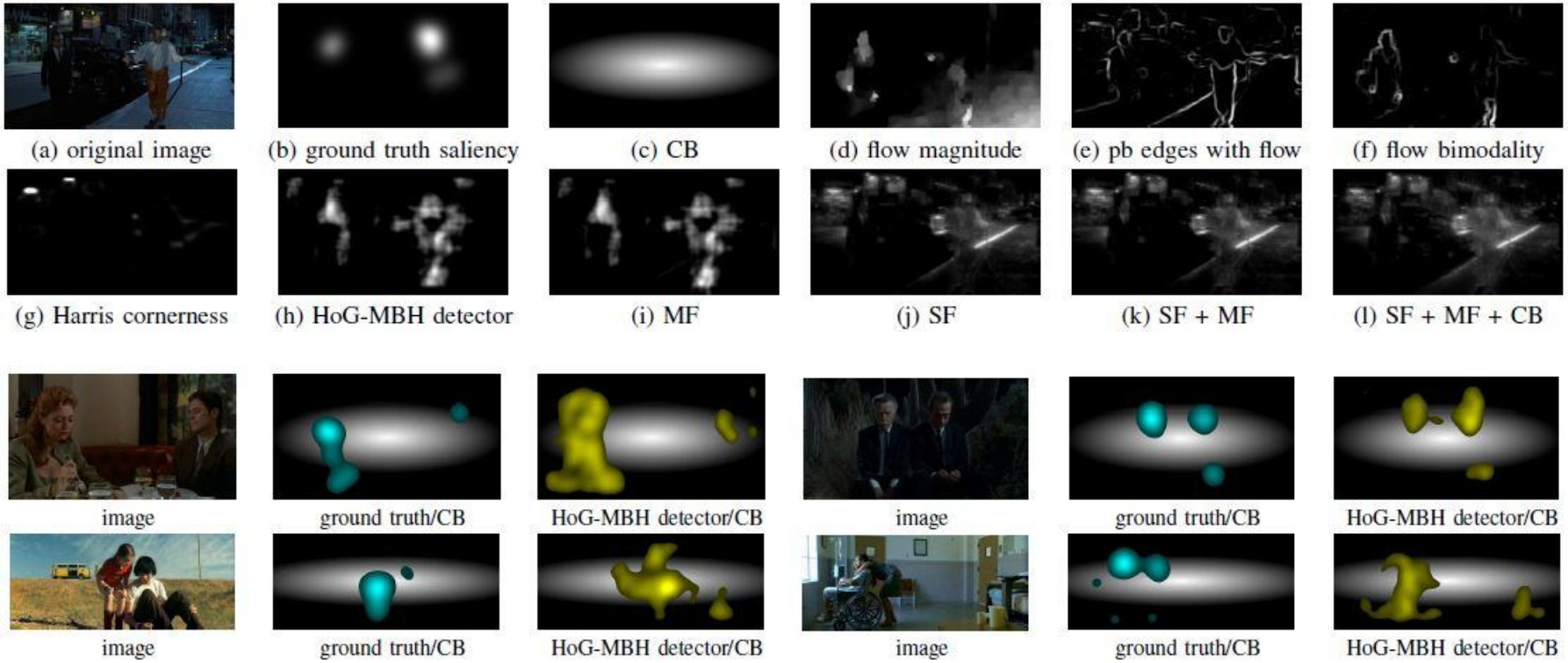
Motion Features

AUC & Spatial KL Divergence

baselines		
feature	AUC (a)	KL (b)
uniform baseline	0.500	18.63
central bias (CB)	0.840	15.93
human	0.936	10.12
static features (SF)		
color features [5]	0.644	17.90
subbands [64]	0.634	17.75
Itti&Koch channels [18]	0.598	16.98
saliency map [56]	0.702	17.17
horizon detector [56]	0.741	15.45
face detector [58]	0.579	16.43
car detector [59]	0.500	18.40
person detector [59]	0.566	17.13

our motion features (MF)		
feature	AUC (a)	KL (b)
flow magnitude	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 [5]	0.789	16.16
SF + CB [5]	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

Automatic Visual Action Recognition



Conclusions

- Combining Human + Computer Vision
- Extending Dataset
- Evaluating Static & Dynamic Consistency
- Human Fixations -> Saliency Maps
- End-to-End Action Recognition System

Thanks!