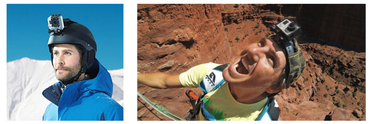
First Person Action Recognition Using Deep Learned Descriptors

Suriya Singh[1], Chetan Arora[2], and C.V. Jawahar[1]
[1] IIIT Hyderabad [2] IIIT Delhi

1. OVERVIEW

First Person Action Recognition is **challenging!**



Unavailability of the actor's pose

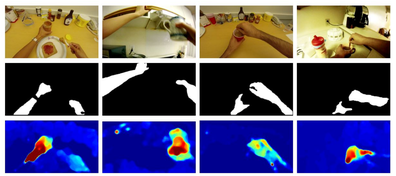
Large camera shakes due to head motion

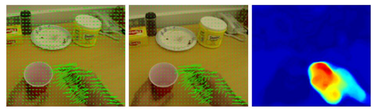
Our contributions

★ Deep learned egocentric features using limited available data
★ Our features are complementary to popular features (DT, iDT, TDD[3])
★ Three-stream architecture: Egocentric, spatial and temporal streams

2. EGO CONVNET

Wearer's hands, head motion and motion saliency are important cues for Egocentric video

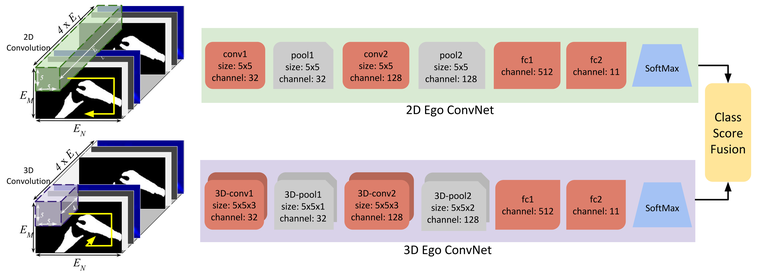




Motion saliency computed after head motion compensation

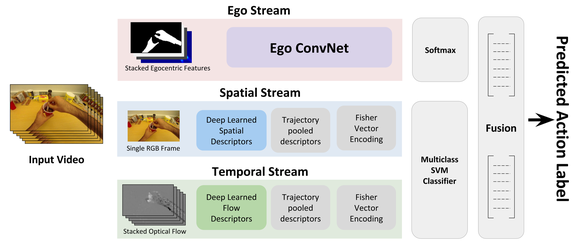
2. EGO CONVNET (CONT'D.)

Pre-processed egocentric cues are stacked and used as input to 2D and 3D CNN.



3. THREE-STREAM ARCHITECTURE

Using deep learned egocentric, spatial and temporal features for first person action recognition



4. RESULTS

Dataset	State of the art	Ours	Ours (cross validated)
GTEA*	47.70[1]	68.50	64.41
Kitchen	48.64[2]	66.23	66.23
ADL	N.A.	37.58	31.62
UTE	N.A.	60.17	55.97

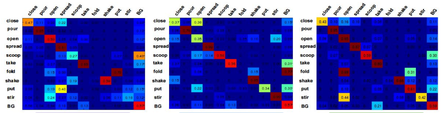
Method	Features	Accuracy
Ego ConvNet 2D	H+C+M	57.61
Ego ConvNet 3D	H+C+M	55.79
TDD[3]	Spatial	58.61
TDD[3]	Temporal	57.12
Combined	**H+C+M+S+T**	**68.50**

Dataset	Accuracy		
	Frame level	Segment level	Chance level
GTEA*	68.50	82.40	11
Kitchen	66.23	71.88	3.4
ADL	37.58	39.02	4.7
UTE	60.17	65.30	4.7

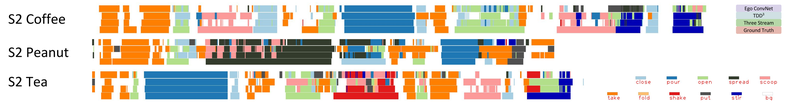
Results are in terms of percentage of accuracy
* GTEA dataset with 11 action classes

Note: For GTEA with 61 classes (action-object), Li *et al* [4], achieved state of the art segment level recognition accuracy of 66.8% (64.0% cross-validated)

Our method perform well across different challenging egocentric video datasets



Ego ConvNet features are complementary with deep learned spatial and temporal features



Error visualization for videos of Subject 2 from GTEA dataset. Most errors occur at the action boundaries!

5. CONCLUSIONS

- CNN for egocentric action recognition can be trained with limited available data.
- Egocentric stream alone can achieve state of the art accuracy.
- Egocentric features are complementary to features from third person video analysis.

REFERENCES

1. Fathi *et. al.* Understanding egocentric activities. ICCV, 2011.
2. Spriggs *et. al.* Temporal segmentation and activity classification from first-person sensing. CVPRW, 2009.
3. Wang *et. al.* Action recognition with trajectorypooled deep-convolutional descriptors. CVPR, 2015.
4. Li *et al.* Delving into egocentric actions. CVPR, 2015.

Codes and Datasets are available on Project Web Page

http://cvit.iiit.ac.in/projects/FirstPersonActions/

IIIT – Hyderabad, India | http://www.iiit.ac.in
IIIT – Delhi, India | http://www.iiitd.ac.in

CVPR 2016

Thanks to Google India and MSR India for the travel grants