Action Recognition In RGB-D Egocentric Videos

Yansong Tang*, Yi Tian, Jiwen Lu, Jianjiang Feng and Jie Zhou

Department of Automation, Tsinghua University, Beijing, 100084, P.R.China

Project Page:
http://ivg.au.tsinghua.edu.cn/dataset/THU_READ.php
*My Website:
http://ivg.au.tsinghua.edu.cn/people/Yansong_Tang

Code of My WeChat

1. OVERVIEW

Motivation

- Egocentric videos: self-generated and embodied, providing richer semantic cues, growing dramatically in recent years.
- RGB-D videos: containing both appearance information and 3D structure of the scenes.
- Exploring the complementary information in egocentric RGB-D videos.

Main Contributions

- An RGB-D egocentric action dataset with diversity and scale.
- A tri-stream convolutional network (TCNet) to take advantage of both the RGB and depth inputs.

2. PROPOSED DATASET

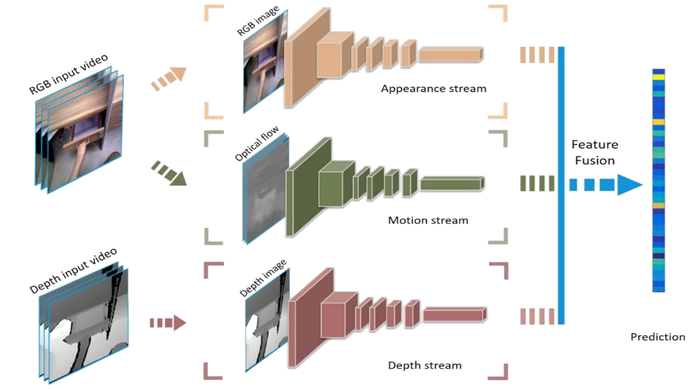
Data collection

- Mounting the RGB-D sensor on a helmet, which was placed on the subject's head.
- Keeping the camera in the same direction with the subject's eyesight.

Dataset	Task	Camera	Frames	Classes
GTEA[16]	Action	RGB	31,253	71
GTEA gaze[15]	Action	RGB	- -	40
GTEA gaze+[15]	Action	RGB	- -	44
UCI ADL[20]	Activity	RGB	93,293	18
WCVS[19]	Activity	RGB-D	- -	20
GUN-71 [21]	Grasp Understanding	RGB-D	12,000	71
Ours	**Action**	**RGB-D**	**343,626**	**40**

- Our dataset shares the advantages on modality, scale and diversity compared with related databases.

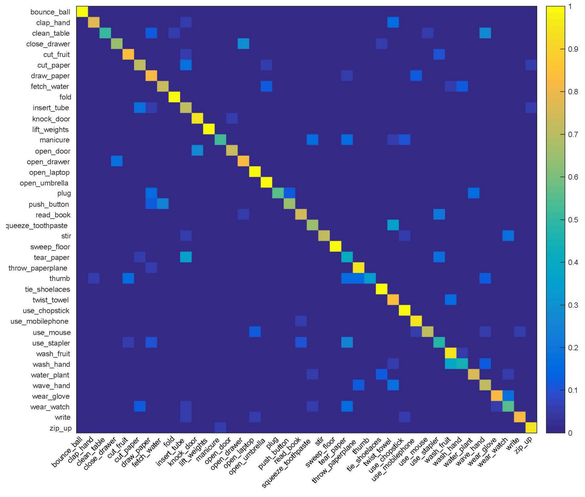
3. APPROACH



- RGB images, optical flows and depth images are sent into the appearance stream, motion stream and depth stream respectively.
- We finally make decision level fusion to predict the action label.

4. EXPERIMENTAL RESULTS

The confusion matrix of the TCNet on our dataset



Method	Accuracy
IDT with higher-dimensional encoding (RGB)	46.9%
IDT with higher-dimensional encoding (Depth)	52.6%
Spatial stream ConvNet	68.4%
Temporal stream ConvNet	40.9%
Depth stream ConvNet	52.7%
Two-stream ConvNet (RGB)	73.3%
MCNet (RGB & Depth)	**76.5%**

- We test some hand-crafted features, deep learning methods and TCNet on our dataset.
- The result shows the complementary information between different modalities, it also demonstrates the efficiency of our TCNet model in comparisons with the state-of-the-arts.

5. FUTURE WORKS

- We will explore to share more semantic information between the RGB and depth modalities, e.g., hand mask and head motion) for action recognition.
- It is desirable to perform more visual tasks like hand-segmentation and human-object interaction on our dataset.

Reference

- Simonyan K, Zisserman A. Two-Stream Convolutional Networks for Action Recognition in Videos. In NIPS2014, pp .568-576.
- Wang H, Schmid C. Action Recognition with Improved Trajectories. In ICCV2013, pp. 3551-3558.