ACTION RECOGNITION IN RGB-D EGOCENTRIC VIDEOS

Yansong Tang[1,2,3], *Yi Tian*[1], *Jiwen Lu*[1,2,3,*], *Jianjiang Feng*[1,2,3], *Jie Zhou*[1,2,3]

[1]Department of Automation, Tsinghua University, Beijing, 100084, China
[2]State Key Lab of Intelligent Technologies and Systems, Beijing, 100084, China
[3]Tsinghua National Laboratory for Information Science and Technology (TNList), Beijing, 100084, China
{tys15,tianyi15}@mails.tsinghua.edu.cn; {lujiwen,jfeng,jzhou}@tsinghua.edu.cn

ABSTRACT

In this paper, we investigate the problem of action recognition in RGB-D egocentric videos. These self-generated and embodied videos provide richer semantic cues than the conventional videos captured from the third-person view for action recognition. Moreover, they contain both appearance information and 3D structure of the scenes from the RGB modality and depth modality respectively. Motivated by these advantages, we first collect a video-based RGB-D egocentric dataset (THU-READ) with diverse types of daily-life actions. Then we evaluate several approaches including hand-crafted features and deep learning methods on THU-READ. To improve the performance, we further develop a tri-stream convolutional network (TCNet) method, which learns to exploit the fuse with both the RGB and depth modalities for action recognition. Experimental results show that our model achieves competitive performance with state-of-the-art methods.

Index Terms— Action recognition, RGB-D, egocentric videos

Fig. 1. **An overview of our egocentric dataset.** We sample pair-wise representative frames from each type of action video. The RGB-based frames are at the top half, while their counterparts of depth modality are at the bottom half.

1. INTRODUCTION

Action recognition [1–5] is a broadly researched field, which attempts to discriminate the action category in a video. Researchers have proposed different methods to learn spatio-temporal descriptors for action representation. These descriptors can be divided into two catagories: *hand-crafted features* [1,4,6,7] and *deep-learned features* [8–10]. Most hand-crafted features describe the local visual patterns based on salient region of actions, while most deep-learned features automatically learn the global representation using deep neural network trained from huge quantity of labeled videos.

With the development of wearable cameras such as GoPro [11], Google Glass [12] and Pivothead [13], the number of egocentric videos is growing dramatically in recent years. Different from the videos captured from the third-person view, these videos are self-generated and embodied [14], which provide richer semantic cues. For these reasons, increasing works have been proposed to investigate the egocentric videos from different aspects, like video summarization [15–17], visual recognition [18–21], social interaction [22], gaze detection [23] and many others. Among these problems, action recognition in egocentric videos is a valuable research issue which presents significant importance for some real-world applications like life logging [24], virtual reality [25] and tele-rehabilitation [21, 26]. However, conventional methods [20, 21, 27, 28] have been presented based on the RGB modality, which lack the utilization of the depth modality and lose 3D structural information of the scenes. Moreover, to our best knowledge, there are few RGB-

* Corresponding author. This work was supported in part by the National Key Research and Development Program of China under Grant 2016YFB1001001, the National Natural Science Foundation of China under Grants 61672306, 61572271, 61527808, 61373074 and 61373090, the National 1000 Young Talents Plan Program, the National Basic Research Program of China under Grant 2014CB349304, the Ministry of Education of China under Grant 20120002110033, and the Tsinghua University Initiative Scientific Research Program.

Fig. 2. **The equipment and method of data collection.** We mounted the RGB-D sensor (left top) on a helmet (left bottom) to make an egocentric equipment (middle). The subject was looking at his hand, plants and the bottle while performing the *water plant* action, which was recorded by the egocentric camera on his head (right).

D video-based datasets for egocentric action recognition.

To address these limitations, we first collect an RGB-D egocentric dataset, which consists of diverse types of daily-life actions as shown in Fig. 1. In order to incorporate RGB and depth modalities, we propose a tri-stream convolutional network (TCNet) which learns to fuse with the complementary information extracted from both modalities. Experimental results show that our model achieves competitive performance on our proposed dataset with state-of-the-art methods.

2. RGB-D EGOCENTRIC ACTION DATASET

In this section, we describe our RGB-D egocentric action dataset (THU-READ) collected in Tsinghua University.

2.1. Data collection

We collected our THU-READ by using the Primesense Carmine camera [29], which has the capability of recording RGB-based and depth-based video sequences simultaneously at 30 fps. Resolutions of these two modalities are both 640 $\times$ 480. Fig. 2 shows the equipment and method of data collection. We mounted the RGB-D sensor on a helmet in the same direction with the subject's eyesight so as to simulate the real conditions. We encouraged the subjects to perform the actions as naturally as possible, which brought greater challenges of shifting backgrounds and various motion speeds to the task of action recognition. For the depth modality, the sensor captured the image frames ranging from 0.3 m to 5 m effectively, covering the space where the subjects performed the actions from the first-person view. We collected our dataset in 5 different scenarios: lab, bathroom, conference room, dormitory and restaurant. In order to balance the data distribution, we asked 8 subjects (6 males and 2 females, height ranging from 162 cm to 185 cm) to repeat performing the action of each class for the same N times (here we chose N = 3). Finally, we obtained 1920 video clips, where

$$1920 = 8\,(\text{subjects}) \times 2\,(\text{modalities}) \times 40\,(\text{classes}) \times 3\,(\text{times})$$

Table 1. A detailed list of all the actions that appear in THU-READ. We classify them according to two criteria: 1. the number of hands in the scenes (single-handed/double-handed) and 2. whether the hands interact with other object (hand-object/non-hand-object).

	single-handed	**double-handed**
hand-object	bounce_ball, clean_table close_drawer, insert_tube knock_door, lift_weight water_plant, open_drawer use_mobilephone, open_door push_button, use_chopstick sweep_floor, use_mouse throw_paperplane	cut_fruit, cut_paper, draw_paper fetch_water, manicure, open_laptop plug, read_book, squeeze_toothpaste stir, tear_paper, tie_shoelaces twist_tower, fold, open_umbrella wash_fruit, wear_glove, wear_watch use_stapler, write, zip_up
non-hand-object	thumb, wave_hand	clap_hand, wash_hand

Table 2. Publicly released egocentric datasets

Dataset	**Task**	**Camera**	**Frames**	**Classes**
GTEA [20]	Action	RGB	31,253	71
GTEA gaze [28]	Action	RGB	52,260	40
GTEA gaze+ [28]	Action	RGB	—	44
UCI ADL [21]	Activity	RGB	93,293	18
WCVS [31]	Activity	RGB-D	—	20
GUN-71 [30]	Grasp Understanding	RGB-D	12,000	71
THU-READ	**Action**	**RGB-D**	**343,626**	**40**

We sample pair-wise representative frames from each type of action videos and present them in Fig. 1. Table 1 shows the list of 40 actions which appear in our dataset in detail. For one thing, our dataset is "all-about-hand", which is classified into two classes of "single-handed" and "double-handed". For another, we divide our dataset into two categories: "hand-object" and "non-hand-object", according to whether "the hands interact with other objects".

We summarize some major statistics of our dataset and the existing related egocentric datasets in Table 2, which shows the advantages on modality, scale and diversity of our dataset. Besides, all of these datasets are video-based except GUN-71 [30], which consists of image sets of different grasp actions.

2.2. Data Preprocessing

Since the sensor is sensitive to illumination, and its estimation algorithm is not robust enough, a few depth frames are especially dark compared to others and thus have to be removed. Having removed an estimated 5% of depth frames and their RGB counterparts, we have 343,626 valid frames in aggregate. The length of each action video instance varies from 34 to 859, depending on the natural lasting time of the action. On average, there are 179 frames per instance. We have re-

leased our dataset at http://ivg.au.tsinghua.edu.cn/dataset/THU_READ.php.

3. TRI-STREAM CONVOLUTIONAL NETWORKS

Having obtained the dataset, we introduce a model to take advantage of both the RGB-based and the depth videos effectively. To address this problem, we propose a tri-stream convolutional network (TCNet) method by including additional stream to learn depth cues and adopt the similar way of fusing each stream as described in [3]. In this part, we will introduce our TCNet model for RGB-D egocentric action recognition.

3.1. Network Input

We prepare pair-wise RGB-D egocentric videos for the TCNet as we mentioned in section 2.2. The frames are resized to 224 $\times$ 224 to fit the input of our network. The RGB-based videos are first decomposed into spatial and temporal components as RGB images and optical flow. Then, we adopt data augmentation used in [32]. After that, the RGB images are sent into the appearance stream. And the optical flow, which is a 3D volume of 224 $\times$ 224 $\times$ 2L (L is the number of stacking flows), is sent into the temporal stream to capture dynamic motion. Simultaneously, the depth-based videos are extracted into depth frames of the size 224 $\times$ 224 $\times$ 3, and are sent into the depth stream. In order to extract optical flow fields as we mention above, we adopt the TVL1 algorithm [33] due to its easy usage and promising performance. We empirically set L = 10 optical flow as the input of the motion stream.

3.2. Network Architecture

Fig. 3 shows the architecture of our proposed TCNet model, which contains three streams in order to utilize the static appearance, dynamic motion and depth information respectively. For the appearance stream and the motion stream, we choose the Two-stream ConvNet model [3], where we replace each stream, which is VGG-M model employed in [34], to VGG-16 model adopted by [35]. The VGG-16 model consists of 13 convolutional layers, 5 max-pooling layers and 3 fully-connected layers. For the depth stream, we employ the same network architecture as the other two streams. In order to match the input size, we modify the input layers to 20 for the motion stream. Also, at the end of each stream, the dimension of the output layers is changed from 1000 to 40 (our dataset consists of 40 types of action). As a result, we obtain a 40-dimension action confidence score $\mathbf{s}_i$ (i = 1, 2, 3) for each modality. Finally, we combine the three streams together by averaging their classifier scores and obtain the prediction score $\mathbf{s}$, where the index of the max element indicates the final action category.

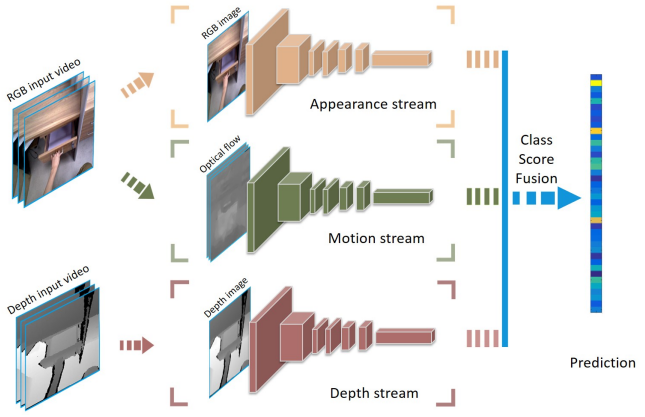

Fig. 3. **An overview of our TCNet model.** We use an egocentric sensor to capture RGB-D data as the input of the TCNet. The RGB videos are first decomposed into spatial and temporal components as RGB images and optical flow, which are sent into the appearance stream and motion stream respectively, while the corresponding depth images are fed into the depth stream. We finally fuse the class score of different modalities at the end of each stream to predict the action label.

3.3. Training Procedure

We used the MatConvnet [36] toolbox and 3 Nvidia Tesla K80 GPUs to train our TCNet model. We employed the VGG-16 model pre-trained on ImageNet [37] and fine-tuned them on the training set of our proposed dataset. The batchsize of each stream was equally set to 96 and the momentum was set to 0.9. We chose a fixed learning rate of 0.001, and the training procedure stopped at the 50th iteration.

4. EXPERIMENTS AND ANALYSIS

4.1. Experimental Setup

For each action class, we randomly sampled about 30% video clips for training and used the rest clips for testing. We employed some existing methods and our TCNet model on our dataset, after which we evaluated them by classification accuracy as well as class confusion matrix. The size of this matrix M is 40 $\times$ 40, and each element M_{ij} represents the percentage of the i-th class testing samples classified into the j-th class.

4.2. Results and Analysis

Hand-crafted Features: We first employed IDT [4] features, due to its better performance compared with exsiting hand-crafted spatio-temporal descriptors [6, 38]. We obtained the HOG [39], HOF [2] and MBH [40] features, which were extracted based on the trajectories of IDT, as well as their combination on both RGB-based and depth-based videos on our egocentric dataset. Then, we employed the higher-dimensional

Table 3. Recognition results on RGB-based and depth-based videos of our egocentric action dataset on 3 descriptors used in IDT [4] and their combination, respectively.

Method	RGB	Depth
HOG [39]	24.61%	64.40%
HOF [2]	26.32%	63.30%
MBH [40]	28.88%	63.44%
Combined Feature [4]	**42.67%**	**66.29%**

Table 4. Main accuracy of several deep learning models of different modalities and their combination on our egocentric action dataset.

Method	Accuracy
Appearance Stream	68.4%
Motion Stream	40.9%
Depth Stream	52.7%
Two-stream ConvNet (RGB) [3]	73.3%
TCNet (RGB & Depth)	**76.5%**

encodings methods [41], with gmmSize empirically set to 256 to generate good performance. We tested several encoding methods, and finally chose the super vector coding (SVC) [42] due to its higher performance than other encoding algorithms mentioned in [41]. Table 3 shows the classification accuracy. On one hand, the hand-crafted features on depth-based videos achieve much more promising performance than that on RGB-based videos, *i.e.* , they are respectively 39.79%, 36.98%, 34.56% higher on the 3 descriptors and 23.62% higher on their combined feature. The results demonstrate the importance of depth input of our dataset, and reveal that, the 3D structural cues of depth data are more effective for trajectory-based hand-crafted features extracting. On the other hand, the combination of these three descriptors on RGB-based videos improved the accuracy by about 16% than using them separately, while that of the depth-based combined features only improved about 3%.

Deep Learning Methods: We have also evaluated each single stream (*i.e.*, the appearance stream, the motion stream and the depth stream) of our proposed model and the Two-stream ConvNet [3]. Table 4 reports their performances. For each single stream, using appearance information achieves the best value (68.4%) of the three. This indicates the virtual importance of the appearance input. The performances of the motion stream and the depth stream are 40.9% and 52.7% respectively, relatively poorer than the appearance stream. Two-stream ConvNet [3], which combines the depth stream and the motion stream together, achieves 73.3% recognition accuracy on our dataset and performs better than adopting these two streams separately.

TCNet model: To evaluate the effectiveness of the TCNet model, which integrates the three recognition streams, we finally tested it on our proposed dataset. Table 4 reports its classification accuracy which attains the performance of 76.5%

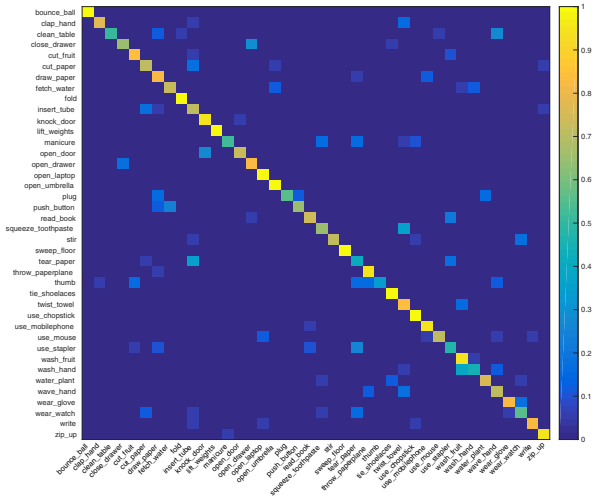

Fig. 4. **The confusion matrix of the TCNet model on our dataset.** The ground truth label is displayed on the vertical axis, while the predicted label is shown as the horizontal axis.

(highest of all the methods in this paper, 3.2% over RGB-based architecture [3] and 23.8% over depth-based stream). This demonstrates the efficiency of our TCNet model in comparisons with the state-of-the-arts. Fig. 4 shows the confusion matrix of recognition results. Most of them were classified correctly except several actions like "wash hand" and "clean table". Since "wash hand" was often confused with "wash fruit" (they both performed "wash", but the objects were different) and "clean table" was sometimes misclassified to "wave hand" (hand kept moving around in these two actions). Moreover, we also employed several ways to fuse the features extracted from fc7, such as average fusion, sum fusion and concat fusion. However, this is not feasible because these features were extracted from different modalities, and the conventional fusion methods mentioned above are not efficient enough to explore their complementary information.

5. CONCLUSION AND FUTURE WORK

In this paper, we have studied the problem of action recognition in RGB-D egocentric videos. We have presented and released an RGB-D action dataset (THU-READ) with diversity and scale, which is captured from the first-person view. We have also proposed a tri-stream convolutional network to take advantage of both the RGB and depth inputs. The experiment achieves competitive performance with the state-of-the-art methods on our dataset. In the future, we will explore to share more semantic information between the RGB and depth modalities for action recognition. Moreover, it is desirable to perform more visual tasks like hand-segmentation and human-object interaction on our dataset.

6. REFERENCES

[1] Ivan Laptev, "On space-time interest points," *IJCV*, vol. 64, no. 2-3, pp. 107–123, 2005.

[2] Ivan Laptev, Marcin Marszalek, Cordelia Schmid, and Benjamin Rozenfeld, "Learning realistic human actions from movies," in *CVPR*, 2008, pp. 1–8.

[3] Karen Simonyan and Andrew Zisserman, "Two-stream convolutional networks for action recognition in videos," in *NIPS*, 2014, pp. 568–576.

[4] Heng Wang and Cordelia Schmid, "Action recognition with improved trajectories," in *ICCV*, 2013, pp. 3551–3558.

[5] Bruce Xiaohan Nie, Caiming Xiong, and Song-Chun Zhu, "Joint action recognition and pose estimation from video," in *CVPR*, 2015, pp. 1293–1301.

[6] Lahav Yeffet and Lior Wolf, "Local trinary patterns for human action recognition," in *ICCV*, 2009, pp. 492–497.

[7] Heng Wang, Alexander Klser, Cordelia Schmid, and Cheng Lin Liu, "Dense trajectories and motion boundary descriptors for action recognition," *IJCV*, vol. 103, no. 1, pp. 60–79, 2013.

[8] Shuiwang Ji, Wei Xu, Ming Yang, and Kai Yu, "3d convolutional neural networks for human action recognition," *PAMI*, vol. 35, no. 1, pp. 221–231, 2013.

[9] Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei, "Large-scale video classification with convolutional neural networks," in *CVPR*, 2014, pp. 1725–1732.

[10] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri, "Learning spatiotemporal features with 3d convolutional networks," in *ICCV*, 2015, pp. 4489–4497.

[11] "Gopro," `https://gopro.com/`.

[12] "Google glass," `https://www.google.com/glass/start/`.

[13] "Pivothead," `http://www.pivothead.com/`.

[14] Dinesh Jayaraman and Kristen Grauman, "Learning image representations tied to ego-motion," in *ICCV*, 2015, pp. 1413–1421.

[15] Omid Aghazadeh, Josephine Sullivan, and Stefan Carlsson, "Novelty detection from an ego-centric perspective," in *CVPR*, 2011, pp. 3297–3304.

[16] Yong Jae Lee, Joydeep Ghosh, and Kristen Grauman, "Discovering important people and objects for egocentric video summarization," in *CVPR*, 2012, pp. 1346–1353.

[17] Zheng Lu and Kristen Grauman, "Story-driven summarization for egocentric video," in *CVPR*, 2013, pp. 2714–2721.

[18] Alireza Fathi, Xiaofeng Ren, and James M Rehg, "Learning to recognize objects in egocentric activities," in *CVPR*, 2011, pp. 3281–3288.

[19] Xiaofeng Ren and Chunhui Gu, "Figure-ground segmentation improves handled object recognition in egocentric video.," in *CVPR*, 2010, vol. 2, p. 6.

[20] Alireza Fathi, Ali Farhadi, and James M Rehg, "Understanding egocentric activities," in *ICCV*, 2011, pp. 407–414.

[21] Hamed Pirsiavash and Deva Ramanan, "Detecting activities of daily living in first-person camera views," in *CVPR*, 2012, pp. 2847–2854.

[22] Alircza Fathi, Jessica K Hodgins, and James M Rehg, "Social interactions: A first-person perspective," in *CVPR*, 2012, pp. 1226–1233.

[23] Yin Li, Alireza Fathi, and James M Rehg, "Learning to predict gaze in egocentric video," in *ICCV*, 2013, pp. 3216–3223.

[24] Jim Gemmell, Gordon Bell, and Roger Lueder, "Mylifebits: a personal database for everything," *CACM*, vol. 49, no. 1, pp. 88–95, 2006.

[25] Dipak Surie, Thomas Pederson, Fabien Lagriffoul, Lars-Erik Janlert, and Daniel Sjölie, "Activity recognition using an egocentric perspective of everyday objects," in *UIC*, 2007, pp. 246–257.

[26] Bruno Kopp, Annett Kunkel, Herta Flor, Thomas Platz, Ulrike Rose, Karl-Heinz Mauritz, Klaus Gresser, Karen L McCulloch, and Edward Taub, "The arm motor ability test: reliability, validity, and sensitivity to change of an instrument for assessing disabilities in activities of daily living," *APMR*, vol. 78, no. 6, pp. 615–620, 1997.

[27] Sudeep Sundaram and Walterio W Mayol Cuevas, "High level activity recognition using low resolution wearable vision," in *CVPR Workshops*, 2009, pp. 25–32.

[28] Alireza Fathi, Yin Li, and James M Rehg, "Learning to recognize daily actions using gaze," in *ECCV*, 2012, pp. 314–327.

[29] "Primesense," `https://en.wikipedia.org/wiki/PrimeSense`.

[30] Grégory Rogez, James S Supancic, and Deva Ramanan, "Understanding everyday hands in action from rgb-d images," in *ICCV*, 2015, pp. 3889–3897.

[31] Mohammad Moghimi, Pablo Azagra, Luis Montesano, Ana C Murillo, and Serge Belongie, "Experiments on an rgb-d wearable vision system for egocentric activity recognition," in *CVPR Workshops*, 2014, pp. 597–603.

[32] Christoph Feichtenhofer, Axel Pinz, and Andrew Zisserman, "Convolutional two-stream network fusion for video action recognition," in *CVPR*, 2016, pp. 1933–1941.

[33] Christopher Zach, Thomas Pock, and Horst Bischof, "A duality based approach for realtime tv-l 1 optical flow," in *DAGM*, 2007, pp. 214–223.

[34] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman, "Return of the devil in the details: Delving deep into convolutional nets," in *BMVC*, 2014, pp. 1–11.

[35] Karen Simonyan and Andrew Zisserman, "Very deep convolutional networks for large-scale image recognition," in *ICLR*, 2015, pp. 1–14.

[36] A. Vedaldi and K. Lenc, "Matconvnet – convolutional neural networks for matlab," in *ACM MM*, 2015, pp. 689–692.

[37] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al., "Imagenet large scale visual recognition challenge," *IJCV*, vol. 115, no. 3, pp. 211–252, 2015.

[38] Heng Wang, Alexander Kläser, Cordelia Schmid, and Cheng-Lin Liu, "Action recognition by dense trajectories," in *CVPR*, 2011, pp. 3169–3176.

[39] Heng Wang, Muhammad Muneeb Ullah, Alexander Klser, Ivan Laptev, and Cordelia Schmid, "Evaluation of local spatio-temporal features for action recognition," in *BMVC*, 2009, pp. 1–11.

[40] Navneet Dalal, Bill Triggs, and Cordelia Schmid, "Human detection using oriented histograms of flow and appearance," in *ECCV*, 2006, pp. 428–441.

[41] Xiaojiang Peng, Limin Wang, Xingxing Wang, and Yu Qiao, "Bag of visual words and fusion methods for action recognition: Comprehensive study and good practice," *CVIU*, vol. 150, pp. 109–125, 2016.

[42] Xi Zhou, Kai Yu, Tong Zhang, and Thomas S Huang, "Image classification using super-vector coding of local image descriptors," in *ECCV*, 2010, pp. 141–154.