Multi-stream Deep Neural Networks for RGB-D Egocentric Action Recognition

Yansong Tang, Zian Wang, Jiwen Lu, *Senior Member, IEEE*, Jianjiang Feng, *Member, IEEE*, and Jie Zhou, *Senior Member, IEEE*

***Abstract*—In this paper, we investigate the problem of RGB-D egocentric action recognition. Unlike conventional human action videos which are passively recorded by static cameras, egocentric videos are self-generated from wearable sensors, which are more flexible and provide the close-ups with the visual attention of the wearers when they act. Moreover, RGB-D videos contain the spatial appearance and temporal information in the RGB modality, and reflect the 3D structure of the scenes in the depth modality. To adequately learn the nonlinear structure of heterogeneous representations from different modalities and exploit their complementary characteristics, we develop a multi-stream deep neural networks (MDNN) method, which aims to preserve the distinctive property for each modality and simultaneously explore their sharable information in a unified deep architecture. Specifically, we deploy a Cauchy estimator to maximize the correlations of the sharable components, and enforce the orthogonality constraints on the distinctive components to guarantee their high independencies. Since the egocentric action recognition is usually sensitive to hand poses, we extend our MDNN by integrating with the hand cues to enhance the recognition accuracy. Extensive experimental results on a newly collected dataset and two additional benchmarks are presented to demonstrate the effectiveness of our proposed methods for RGB-D egocentric action recognition.**

***Index Terms*—Egocentric action recognition, RGB-D videos, multi-view learning, deep learning.**

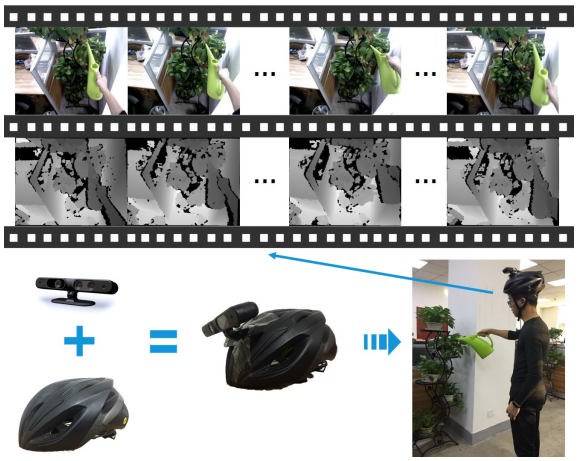


Fig. 1. **Illustration of collecting RGB-D egocentric action videos.** We mounted an RGB-D sensor on a helmet to make an egocentric equipment. The wearer was looking at his hand, plants and the bottle while performing the "*water plant*" action, which was recorded by the egocentric camera on his head. We present the captured RGB video frames at the top row, while their counterparts of depth modality are shown at the second row.

I. Introduction

With the development of wearable cameras such as GoPro and Google Glass, the number of egocentric videos is growing dramatically in recent years. Different from conventional human action videos which are passively recorded by static cameras, these videos are self-generated when the wearers act and provide the close-ups with their visual attention. Therefore, increasing works have been proposed to analyze the egocentric videos from different aspects [1], *e.g.,* frame sampling [2], video summarization [3], visual recognition [4]–[6], person re-identification [7] and gaze analysis [8]. Among these problems, action recognition in egocentric videos presents significant importance for some real-world applications, *e.g.,* healthcare [9], life logging [10], virtual reality [11] and tele-rehabilitation [6].

The authors are with the Department of Automation, Tsinghua University, State Key Lab of Intelligent Technologies and Systems, and Beijing National Research Center for Information Science and Technology (BNRist), Beijing, 100084, China. E-mail: tys15@mails.tsinghua.edu.cn; wza15@mails.tsinghua.edu.cn; lujiwen@tsinghua.edu.cn; jfeng@tsinghua.edu.cn; jzhou@tsinghua.edu.cn. (Corresponding author: Jiwen Lu.)

In recent years, efforts on the egocentric action recognition [5], [6], [12] have been devoted by employing conventional human action recognition techniques [13], [14] and extra semantic cues (*e.g.,* hand pose and gaze) in the egocentric videos. Generally speaking, these works are mainly based on RGB videos, which contain the spatial appearance and temporal information. As stated in [15], [16], the primary limitations of RGB videos are the absence of 3D information and the sensitivity to illumination variations, while an exclusive depth modality is capable of covering these shortages. In fact, RGB-D action recognition [16]–[19] has been widely studied in the literature. However, limited research has been conducted to investigate this problem in the egocentric paradigm. This is mainly due to three reasons: 1) the difficulty of jointly utilizing different modalities and exploiting their complementary information for egocentric action recognition, 2) the inherent challenges in the egocentric scenario, such as the cluttered background and large movements of the wearable camera, and 3) the scarcity of the publicly available RGB-D egocentric dataset for action recognition.

In this work, we propose a multi-stream deep neural networks (MDNN) method for RGB-D egocentric action recognition, which aims to adequately learn the non-linear structure

of heterogeneous representations from different modalities and exploit their complementary characteristics. Specifically, we utilize three deep convolutional neural networks to learn the spatial appearance, temporal information and geometric property for the RGB frames, optical flows and depth frames (*i.e.,* frames extracted from the video on depth modality) accordingly. Moreover, our MDNN enforces two criteria on learning features: 1) preserving the distinctive characteristics for each modality via orthogonality constraints, and 2) maximizing the correlations across different modalities by deploying the Cauchy estimator [20]. To further improve the recognition performance, we extend our MDNN by fusing with the hand cues in the egocentric videos at classification score level. In order to demonstrate the effectiveness of our proposed methods, we present a new RGB-D egocentric dataset called THU-READ, which contains 340K video frames of 40 different daily-life actions and 200M-pixel hand annotation. Extensive experimental results on the THU-READ and two additional benchmarks clearly show that our model achieves superior performance in comparison with the state-of-the-arts for RGB-D egocentric action recognition.

Our main contributions are summarized as follow:

1) We have developed a multi-stream deep neural networks (MDNN) method for RGB-D egocentric action recognition. Our MDNN exploits the shared properties and distinctive characteristics for different modalities simultaneously, which are more adequate to learn the complementary properties of the spatial appearance, temporal information and geometric structure of the scene.
2) We have collected a new dataset for RGB-D egocentric action recognition, which is currently the largest dataset for this emerging task. Extensive experimental results on the proposed dataset and the other two benchmarks have clearly shown that our MDNN achieved superior performance compared with the state-of-the-arts. Moreover, we have included the experiments on parameter analysis and t-SNE visualization [21] to show the importance of the sharable and distinctive characteristics of different modalities.
3) We have provided over 200M-pixel hand annotation in our dataset. Besides, we have strengthened our MDNN by incorporating with hand cues in the egocentric videos and conducted experiments to demonstrate its effectiveness.

The remainder of this paper is organized as follows: Section II briefly reviews some related work. Section III introduces the collected RGB-D egocentric action dataset. Section IV describes the proposed MDNN for egocentric RGB-D action recognition in details. Section V reports experimental presents and analysis, and Section VI concludes the paper. It is to be noted that a preliminary version of this work was initially presented in [22][1].

[1] Partial results in this paper have been published in our previous conference paper [22]. As an extension, our MDNN with a new object function can better exploit the complementary information of different modalities than the score fusion approach in [22]. Moreover, we have conducted experiments on the other two datasets and provided more in-depth analysis on the experimental results. Besides, we have presented two standard evaluation protocols for our proposed dataset and extend our MDNN by integrating with the hand cues to enhance the recognition accuracy.

II. Related Work

In this section, we briefly review three related topics: 1) conventional action recognition, 2) multi-view learning, 3) egocentric action analysis, and 4) optical flow.

A. Conventional Action Recognition

Conventional action recognition [13], [14], [23]–[31] aims to classify the action category for a given RGB video. Over the past two decades, extensive works have been proposed to develop discriminative features to capture the appearance and motion information of an action. These works can be roughly divided into two categories: *hand-crafted features* and *deeply-learned features*. The *hand-crafted features* generally describe local visual patterns based on the spatio-temporal interest points or trajectories, such as Space-Time Interest Points (STIP) [23], Histogram of Gradient and Histogram of Optical Flow (HOG/HOF) [24], Histogram of Motion Boundary (MBH) [32], dense trajectory (DT) [33] and improved dense trajectory (IDT) [13]. However, these methods require strong prior knowledge by hand and may lack power to model the non-linear relationship between the high dimensional videos and action labels. To overcome these limitations, the *deeply-learned features* have been proposed for action recognition by designing various deep architectures, *e.g.,* 2D ConvNet [14], [34], 3D ConvNet [35], [36], recurrent neural network [37], etc. Recently, Simonyan and Zisserman developed the two-stream ConvNets [14], where the spatial stream captures the appearance from the static RGB frames, and the temporal stream learns the dynamics from the optical flows. This architecture achieves high performance for action recognition, however, it has not fully exploited the sharable and individual information between the two streams. To address this issue, we simultaneously explore their distinctive characteristics and sharable properties. Besides, we capture the 3D structures of the scenes from an extra depth modality and utilize the hand cues in the egocentric videos to enhance the performance of action recognition.

B. Multi-view Learning

Recent years have witnessed that multi-view learning has attracted growing attention in the research fields of machine learning and computer vision [49]–[59], which aims to jointly learn an optimal combination on multi-view (multi-modal) representations. For example, Yang *et al.* [60] proposed a Multi-feature Learning via Hierarchical Regression (MLHR) algorithm, which effectively mined the information in multiple features of both labelled and unlabelled data for multimedia analysis. Dong *et al.* [61] investigated the problem of few-example object detection by a self-paced strategy. During the learning process, they employed a multi-modal learning method to embed multiple detection models to avoid local minimums. Over the past few decades, a typical series of methods such as canonical correlation analysis (CCA) [62], [63] and its variants [50]–[52] are developed to explore a shared latent subspace across different views to maximize their correlation. However, these approaches may ignore the

TABLE I
PUBLICLY RELEVANT EGOCENTRIC ACTION/ACTIVITY DATASETS

Dataset	Subjects	Clips	Camera	Mount	Frames	Classes	Year	Video	Hand Annotation
CMU-MMAC [38]	39	175	RGB	Head	–	29	2009	✓	✗
UEC EgoAction [39]	1	2	RGB	Head	–	37	2011	✓	✗
GTEA [5]	4	28	RGB	Head	31,253	71	2011	✓	pixel-level annotation
Disney World [40]	18	113	RGB	Head	–	6	2012	✓	✗
GTEA gaze [12]	14	17	RGB	Head	52,260	40	2012	✓	✗
GTEA gaze+ [12]	5	30	RGB	Head	–	44	2012	✓	✗
UCI ADL [6]	20	20	RGB	Chest	93,293	18	2012	✓	✗
JPL-Interaction [41]	1	62	RGB	Head	–	7	2013	✓	✗
HUJI EgoSeg [42]	3	44	RGB	Head	–	–	2014	✓	✗
WCVS [43]	4	918	RGB-D	Head	–	2/4/18*	2014	✓	✗
GUN-71 [44]	8	–	RGB-D	Chest	12,000	71	2015	✗	contact point and force annotation
MILADL [45]	20	122	RGB	Head/Wrist	–	23	2016	✓	✗
EgoHands [46]	4	48	RGB	Head	130,000	4	2016	✓	pixel-level annotation
Standford-ECM [47]	10	113	RGB	Chest	–	24	2017	✓	✗
FHAD [48]	6	1,175	RGB-D	Shoulder	105,459	45	2018	✓	hand pose annotation
THU-READ (Ours)	**8**	**1,920**	**RGB-D**	**Head**	**343,626**	**40**	**2017**	✓	pixel-level annotation

* WCVS dataset has 3-level categories, which will be described in Section V.B in detail.

view-specific characteristics, which encode some discriminative information of each view (*e.g.,* color or texture property of an RGB image) [64]. To address this issue, researchers attempted to simultaneously exploit the sharable and individual components of different views to strengthen their models, which achieved competitive performance on various RGB-D vision tasks, *e.g.*, object recognition [65], person re-identification [66], etc. Similarly, recent advances on RGB-D action recognition [15], [16], [19], [67], [68] have also focused on mining the complementary information between the RGB and depth modalities to model actions in spatio-temporal domain. For example, Hu *et al.* [16] extracted heterogeneous hand-crafted features from the RGB-D sequences and explored their shared-specific properties through iterative optimization. Shahroudy *et al.* [15] fed multi-view features into a deep auto-encoder network to factorize their shared-specific components. Different from these works, we develop a unified deep architecture and directly leverage the RGB pixels, optical flows and depth frames as the network inputs. Further more, unlike recent RGBD-based methods [15], [66] which mainly adopted l_2 norm to measure the distance between different components, we deploy a smooth Cauchy estimator [20], which is robust to the outliers, to maximize the correlations across different modalities. Moreover, we apply orthogonality constraints to guarantee the high independency for each modality, so that its distinctive characteristics can be well preserved.

C. Egocentric Action Analysis

With the development of wearable cameras such as GoPro and Google Glass, more and more datasets have been proposed in the egocentric research field to evaluate different approaches. Table I summarizes the comparison between some publicly relevant egocentric action/activity datasets and our proposed dataset. While the existing datasets present various challenges for action recognition to some extent, they still have some limitations in different aspects, *e.g.,* scale, annotation or data modality. In contrast, our THU-READ has the advantages of: 1) larger scale with much more video clips and video frames, 2) pixel-level hand annotation, which provide visual hints for action recognition and hand-object interaction, and 3) videos in RGB and depth modalities, which simultaneously present the appearance information and 3D structure of the scenes during the action process.

Egocentric video analysis is a rapidly growing field and has been explored from different aspects, including action detection, action anticipation, action classification and many others [69]–[72]. Action detection, which aims to recognize actions with temporal segmentation, is an emerging and challenging topic in recent years [73], [74]. For example, Damen *et al.* [75] proposed a large-scale egocentric benchmark in the kitchen environments, enabling the egocentric community to apply and develop various data-driven learning approaches for this task. Huang *et al.* [76] presented a temporal action proposals (TAPs) method to localize generic actions in egocentric videos. For action anticipation, Bertasius *et al.* [77] proposed a model consisting of a future convolutional neural network and a goal verifier network, to generate the basketball motion sequence from a single first-person image. Rhinehart *et al.* [78] explored the tasks of the first-person trajectory forecasting and human activities prediction based on the online learning theory and inverse reinforcement learning method. For action classification, Possas *et al.* [79] developed a reinforcement learning framework, which minimized the energy cost of wearable sensor while keeping the competitive accuracy levels of egocentric activity recognition. Recently, researchers have explored some particular semantic cues (*e.g.,* hands, gaze, etc.) in the egocentric scenarios to study this fundamental problem [5], [6], [12], [43], [44], [80]–[82]. For example, Fathi *et al.* [5] proposed a hierarchical inference architecture to explore the consistent relationship of object,

hand and action. Singh *et al.* [81] extended the two-stream architecture [14] with an EgoConvNet, which utilized the hand mask, head motion and saliency map for action recognition. For the RGB-D egocentric action recognition, Moghimi *et al.* [43] analyzed the performance of different image-based representations and leveraged depth information to help skin segmentation. However, they ignored the temporal consistency in the videos and left room for making use of depth information. Garcia-Hernando *et al.* [48] collected a first-person hand action benchmark with RGB-D videos, but they did not explore how to jointly utilize these two modalities to enhance the recognition performance. In comparison, we adequately exploit the complementary properties of static appearance, temporal consistency and depth information in the egocentric videos, while the hand cues are further leveraged to boost the recognition accuracy.

D. Optical Flow

Over the past decades, various methods have been proposed to estimate optical flow since the pioneering work presented by Horn and Schunck [83]. For example, Zach *et al.* [84] employed a total variation L1 norm regularization (TVL1) method to compute optical flow in real-time. Brox *et al.* developed FlowNet and FlowNet 2.0 architectures [85], [86] by deep convolutional networks, which achieved promising performance for optical flow estimation. More recently, Hui *et al.* [87] further adopted technologies of cascaded flow inference and flow regularization, and presented a more efficient network architecture named LiteFlowNet.

As for video analysis, numbers of tasks (*e.g.,* action recognition [14], video-based facial landmark detection [88] and object segmentation [89]) have benefited from optical flow. For example, Dong *et al.* [88] presented an unsupervised supervision-by-registration approach for facial landmark detection on both images and videos, which leveraged optical flow to guarantee the coherency of detection results between two adjacent frames. In this work, we aim to capture the motion information in egocentric videos by optical flow. We employed optical flow based on TVL1 algorithm [84], which achieves a good trade-off between accuracy and efficiency.

III. RGB-D Egocentric Action Dataset

In this section, we will describe our proposed RGB-D Egocentric Action Dataset (THU-READ). The motivation of collecting this dataset is to simultaneously record egocentric videos in RGB and depth modalities, which represent the spatial appearance, temporal information and 3D structure of the scenes. To our best knowledge, this is currently the largest RGB-D video-based dataset for egocentric action recognition.

A. Data Collection

We collected our RGB-D egocentric action dataset by a Primesense Carmine camera, which is a RGB-D sensor released by Primesense[2]. This sensor has the capability of simultaneously recording videos in RGB and depth modalities at 30 fps. Resolutions of these two modalities are both 640×480. Fig. 1 shows the equipment and method of data collection. We mounted the RGB-D sensor on a helmet, which was placed on the subject's head. The device is about 1.23 lbs totally, this weight is light and would not bring burden when the subject performed the action. For the purpose of acquiring egocentric action videos, we kept the camera 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 video frames ranging from 0.3m to 5m effectively in practice, covering the space where the subjects performed the actions from the first-person view during the process of data collection. 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, heights ranging from 1.62 m to 1.85 m) to repeat performing the action of each class for the same N times (here we chose N to be 3). Finally, we obtained 1920 video clips, where

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

B. Data Preprocessing and Annotation

Since the sensor is sensitive to illumination, 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 total. 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.

As suggested in previous works [5], [81], [82], hands provide important visual cues for understanding human action and human-object interaction in egocentric videos. However, pixel-level handmask annotation is limited [80] and valuable. In order to make use of this semantic information, we provide over 200 million labeled pixels annotation of our database. We have already released our THU-READ and the hand annotation at http://ivg.au.tsinghua.edu.cn/dataset/THU_READ.php.

IV. Proposed Approach

In this section, we describe the proposed multi-stream deep neural networks (MDNN) and its extension with the hand cues (MDNN + hand) in details.

A. MDNN

In order to learn discriminative and robust features for RGB-D egocentric actions recognition, the basic idea of our MDNN is to preserve the distinctive characteristics for each modality and simultaneously explore their sharable information. Fig. 2 shows the pipeline of our proposed method, which mainly consists of three steps: 1) feature extraction, 2) multi-view learning, and 3) classification. We mainly describe the first two steps as follow.

[2]https://en.wikipedia.org/wiki/PrimeSense

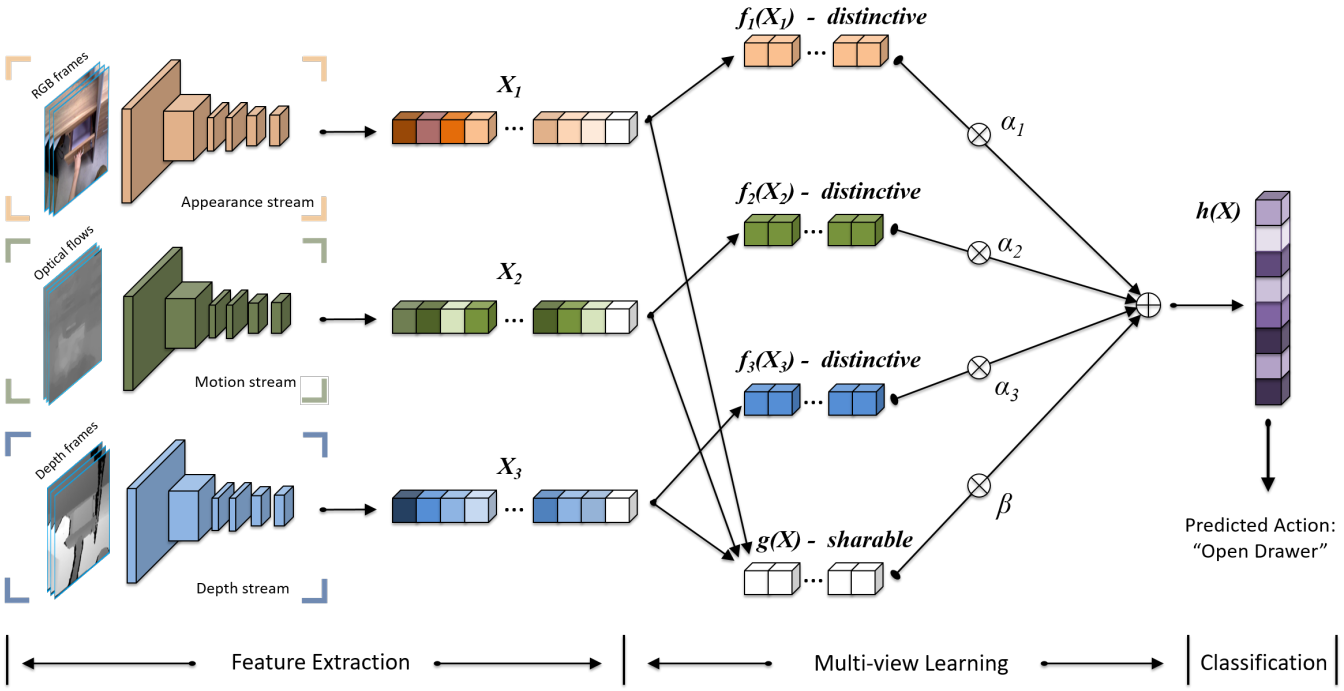


Fig. 2. **An overview of the proposed muti-stream deep convolutional neural networks (MDNN).** The inputs to our MDNN consist of the RGB frames, optical flows and depth frames extracted from the RGB-D egocentric videos. At the *feature extraction* stage, we utilize three deep neural networks to learn the spatial appearance, temporal information and geometric property, and extract the features $\boldsymbol{X}_1$, $\boldsymbol{X}_2$ and $\boldsymbol{X}_3$ for each modality accordingly. During the *multi-view learning* process, we carefully disentangle the learned features into four components as $\boldsymbol{f}_1(\boldsymbol{X}_1), \boldsymbol{f}_2(\boldsymbol{X}_2), \boldsymbol{f}_3(\boldsymbol{X}_3)$ and $\boldsymbol{g}(\boldsymbol{X})$, where $\boldsymbol{f}_1(\boldsymbol{X}_1), \boldsymbol{f}_2(\boldsymbol{X}_2), \boldsymbol{f}_3(\boldsymbol{X}_3)$ preserve the distinctive characteristics for each modality and $\boldsymbol{g}(\boldsymbol{X})$ contains their sharable information. We combine the four components by allocating different weights $\{\alpha_1, \alpha_2, \alpha_3, \beta\}$. Having obtained the fused feature $\boldsymbol{h}(\boldsymbol{X})$, we pass it to the final *classification* layer to predict the action label (best viewed in the color pdf file).

1) Feature Extraction: Unlike recent approaches for RGB-D action recognition [15], [16], which extracted hand-crafted features around the major joints of human body, our MDNN directly takes RGB frames, optical flows and depth frames as the network inputs, because the 3D skeleton information is not available in the egocentric videos. As shown in Fig. 2, we utilize three streams, namely appearance stream, motion stream and depth stream to capture the static appearance, temporal information and 3D structure of a pair-wise RGB-D egocentric video accordingly. To be specific, we first decompose the RGB video into spatial and temporal components as RGB images and optical flows, and simultaneously extract depth frames from the depth video. We employ the TVL1 algorithm [84] to calculate optical flow between two adjacent frames of the RGB videos (except the last frame of a video) because of its balance between accuracy and efficiency. We follow [90] to discretize the optical flow into the interval from 0 to 255 by a linear transformation. The inputs of our MDNN are frames on the three modalities at the same time-stamp, which are assigned with the action label corresponding to the original video. We deploy three deep convolutional neural networks to learn the deep features of the network inputs (please see section V.A for details), then we develop a multi-view learning method to fuse the features and obtain the class score at the end of the network. Finally, we average the scores of all the frames to predict the action category of the video. In this way, we are capable to leverage the information of the entire video adequately and ensure the temporal consistency of the intermediate features extracted from different modalities.

2) Multi-view Learning: Since features extracted from different streams exhibit their own property, it is lack of physical meaning to combine them directly with some typical fusion methods (*e.g.,* concatenation or element-wise sum). To address this issue, we aim to transfer the heterogeneous features into some new spaces, which bridge the modality gap so that they can be compared. As features in different modalities reflect the properties of a specific action from different aspects, they are neither fully independent nor fully correlated. Therefore, the spaces we seek should contain the sharable information and distinctive characteristics of different modalities.

More formally, we denote the features extracted from each single network as $\boldsymbol{X} = \{\boldsymbol{X}_i\}_{i=1}^{K}$, where $\boldsymbol{X}_i$ represents the features in the ith modality and K is the total number of modalities. We define the fusion function as: $\boldsymbol{X} \rightarrow \boldsymbol{h}(\mathbf{X})$,

which combines the input features $\boldsymbol{X}$ into the output feature $\boldsymbol{h}(\boldsymbol{X})$. In order to adequately explore the sharable information and distinctive characteristics of different modalities, we introduce two types of intermediate features $\boldsymbol{g}(\boldsymbol{X})$ and $\{\boldsymbol{f}_i(\boldsymbol{X}_i)\}_{i=1}^{K}$. On one hand, $\boldsymbol{g}(\boldsymbol{X})$ contains the sharable information of different modalities:

$$\boldsymbol{g}(\boldsymbol{X}) = \frac{1}{K}\sum_{i=1}^{K}\boldsymbol{g}_i(\boldsymbol{X}_i), \tag{1}$$

where $\{\boldsymbol{g}_i(\boldsymbol{X}_i)\}_{i=1}^{K}$ are the sharable components. To better

model the non-linear relationship between the features of $\boldsymbol{X}_i$ and $\boldsymbol{g}_i(\boldsymbol{X}_i)$, we perform linear mapping followed by a non-linear activated function:

$$\boldsymbol{g}_i(\boldsymbol{X}_i) = \sigma(\mathbf{W}_i^s \boldsymbol{X}_i + \mathbf{b}_i^s), \quad i = 1, 2, ..., K, \tag{2}$$

where $\sigma(.)$ denotes the non-linear function. We employ the tanh function in our implementation. $\mathbf{W}_i^s$ and $\mathbf{b}_i^s$ are the corresponding weighted matrix and bias term.

On the other hand, $\boldsymbol{f}_i(\boldsymbol{X}_i)$ retains the discriminative characteristics that exclusively exists in each modality. Similar to $\boldsymbol{g}_i(\boldsymbol{X}_i)$, we obtain $\boldsymbol{f}_i(\boldsymbol{X}_i)$ by another non-linear transformation:

$$\boldsymbol{f}_i(\boldsymbol{X}_i) = \sigma(\mathbf{W}_i^d \boldsymbol{X}_i + \mathbf{b}_i^d), \quad i = 1, 2, ..., K, \tag{3}$$

Note that the intermediate features $\boldsymbol{g}(\boldsymbol{X})$ and $\{\boldsymbol{f}_i(\boldsymbol{X}_i)\}_{i=1}^K$ may not be equally important for action recognition. Thus, we integrate them to obtain the target feature $\boldsymbol{h}(\boldsymbol{X})$ by allocating different weights as follow:

$$\begin{aligned} \boldsymbol{h}(\boldsymbol{X}) \;=\; & \sum_{i=1}^{K} \alpha_i \boldsymbol{f}_i(\boldsymbol{X}_i) + \beta \boldsymbol{g}(\boldsymbol{X}), \\ \text{subject to} \quad & \sum_{i=1}^{K} \alpha_i + \beta = 1, \\ & 0 \leqslant \alpha_1, \alpha_2, ..., \alpha_K, \beta \leqslant 1, \end{aligned} \tag{4}$$

where $\{\alpha_i\}_{i=1}^K$ and β are the soft assignment weights corresponding to the intermediate features $\{\boldsymbol{f}_i(\boldsymbol{X}_i)\}_{i=1}^K$ and $\boldsymbol{g}(\boldsymbol{X})$ respectively. Having obtained the fused feature $\boldsymbol{h}(\boldsymbol{X})$, we feed it into a fully-connected layer followed by a softmax function to predict the action label. We will conduct experiments to analyze the hyper-parameters $\{\alpha_i\}_{i=1}^K$ and β in later experiments.

3) Objective Function: We formulate our MDNN as the following optimization problem:

$$\begin{aligned} \min_{\boldsymbol{\theta}} \boldsymbol{J} \;=\; & \boldsymbol{J}_{cls} + \lambda_1 \boldsymbol{J}_s + \lambda_2 \boldsymbol{J}_d \\ =\; & -\sum_{l=1}^{L} \mathbb{1}(y = l) log(s_l) \\ +\; & \lambda_1 \sum_{1 \leq i < j \leq K} \Phi_s(\boldsymbol{g}_i(\boldsymbol{X}_i), \boldsymbol{g}_j(\boldsymbol{X}_j)) \\ +\; & \lambda_2 [\sum_{1 \leq i < j \leq K} \Phi_d(\boldsymbol{f}_i(\boldsymbol{X}_i), \boldsymbol{f}_j(\boldsymbol{X}_j)) \\ +\; & \sum_{i=1}^{K} \Phi_d(\boldsymbol{f}_i(\boldsymbol{X}_i), \boldsymbol{g}_i(\boldsymbol{X}_i))]. \end{aligned} \tag{5}$$

Here $\boldsymbol{\theta}$ denotes the parameters of the entire network, λ_1 and λ_2 are two hyper-parameters to balance the effects of different terms to make a good trade-off. The physical interpretations of the $\boldsymbol{J}_{cls}$, $\boldsymbol{J}_s$ and $\boldsymbol{J}_d$ are respectively explained as below.

The first term $\boldsymbol{J}_{cls}$ represents classification loss for action recognition. We calculate the categorical cross-entropy loss, where $\mathbb{1}$ is the indicator function which equals 1 when the prediction $y = l$ is true and 0 otherwise. Here y and l denote the predicted label and ground-truth label, L is the number of the total action categories and the softmax output s_l represents the corresponding class probability.

The second term $\boldsymbol{J}_s$ aims to exploit the sharable informations of different modalities. To achieve this goal, we employ the Cauchy estimator [20] $\rho_0(x) = log(1+(x/a)^2)$ to measure the correlations between the sharable components $\{\boldsymbol{g}_i(\boldsymbol{X}_i)\}_{i=1}^K$ and minimize it during the optimization. Here, a is a hyper-parameter and $\boldsymbol{J}_s$ is derived as follow:

$$\Phi_s(\boldsymbol{g}_i(\boldsymbol{X}_i), \boldsymbol{g}_j(\boldsymbol{X}_j)) = \log(1 + \frac{\|\boldsymbol{g}_i(\boldsymbol{X}_i) - \boldsymbol{g}_j(\boldsymbol{X}_j)\|^2}{a^2}). \tag{6}$$

Typically, it is more straightforward to calculate the L_1 or L_2 distance to estimate the correlations between the sharable components $\{\boldsymbol{g}_i(\boldsymbol{X}_i)\}_{i=1}^K$. However, the illumination variation and camera movement usually cause some outliers in the egocentric videos and neither L_1 nor L_2 distance is robust to the outliers [49]. To further illustrate this, we consider the influence function $\Psi(x)$ of an estimator $\rho(x)$, which is mathematically defined as $\Psi(x) = \partial\rho(x)/\partial x$. For the absolute value estimator (*i.e.,* L_1 distance) $\rho_1(x) = |x|$, its influence function has no cut-off, while the least-squares estimator (*i.e.,* L_2 distance) $\rho_2(x) = x^2/2$ has the influence function $\Psi_2(x) = x$, which increases linearly with x. In comparison, the influenced function of Cauchy estimator is $\Psi_0(x) = 2x/(a^2 + x^2)$, which has the upper bound $1/a$ for $x > 0$ and more smooth, therefore it is more robust to the outliers [49]. In practice, we set the hyper-parameter a to be 1. We also conduct the experiments to demonstrate the advantage of Cauchy estimator compared with the L_1 and L_2 distance in the later section.

The third term $\boldsymbol{J}_d$ in equation (5) attempts to preserve the distinctive characteristics for each modality. Towards this goal, we enforce the orthogonality constraints on $\{\boldsymbol{g}_i(\boldsymbol{X}_i)\}_{i=1}^K$ and $\{\boldsymbol{f}_i(\boldsymbol{X}_i)\}_{i=1}^K$ as follow:

$$\begin{aligned} \Phi_d(\boldsymbol{f}_i(\boldsymbol{X}_i), \boldsymbol{f}_j(\boldsymbol{X}_j)) \;&=\; |\boldsymbol{f}_i(\boldsymbol{X}_i) \odot \boldsymbol{f}_j(\boldsymbol{X}_j)|, \\ \Phi_d(\boldsymbol{f}_i(\boldsymbol{X}_i), \boldsymbol{g}_i(\boldsymbol{X}_i)) \;&=\; |\boldsymbol{f}_i(\boldsymbol{X}_i) \odot \boldsymbol{g}_i(\boldsymbol{X}_i)|, \end{aligned} \tag{7}$$

where $\odot$ is the element-wise Hadamard product. Through the orthogonality constraints, the distinctive components $\{\boldsymbol{f}_i(\boldsymbol{X}_i)\}_{i=1}^K$ are enforced to be independent to each other. Also, $\boldsymbol{f}_i(\boldsymbol{X}_i)$ is regularized to be irrelevant to its corresponding sharable component $\boldsymbol{g}_i(\boldsymbol{X}_i)$. Therefore we are able to guarantee the specifics for different modalities by minimizing $\boldsymbol{J}_d$.

To optimize (5), we employ the standard back-propagation method for learning all the parameters $\boldsymbol{\theta}$ of our MDNN. The gradient $\partial \boldsymbol{J}/\partial \boldsymbol{\theta}$ is calculated by the deep learning toolbox [91] automatically. We summarize the pipeline of our MDNN in **Algorithm 1**.

B. MDNN+hand

While the proposed MDNN learns to fuse the data in different modalities for action recognition, it directly leverages the *global* information of the video frames and optical flows as the network inputs. In this part, we aim to further exploit some *local* information which contains important semantic hints in the egocentric videos to improve the performance of action recognition. More specifically, we focus on the hand cues, because 1) human often pay attention to their hands when they act, so that the hands usually appear in the egocentric videos,

Algorithm 1: MDNN

Input: Training videos: $\{\mathbf{V}_{RGB}, \mathbf{V}_{depth}\}$, lable l, Parameters: Γ (iterative number), η (learning rate) and ϵ (convergence error).
Output: the weights of MDNN $\boldsymbol{\theta}$.
// Data-preprocessing:
Extract $\mathbf{V}_{RGB}$ to RGB frames and optical flows.
Extract $\mathbf{V}_{depth}$ to depth frames.
// MDNN training:
Initialize $\boldsymbol{\theta}$.
Perform forward propagation.
Calculate the initial $\boldsymbol{J}_0$ by (5).
for $t \leftarrow$ *1, 2, ...,* Γ **do**
  // Update $\boldsymbol{\theta}$ *by back propagation:*
  $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} - \eta\, \partial \boldsymbol{J} / \partial \boldsymbol{\theta}$
  Perform forward propagation.
  Compute the objective function $\boldsymbol{J}_t$ using (5).
  If $|\boldsymbol{J}_t - \boldsymbol{J}_{t-1}| < \epsilon$, go to **Return**.
end
Return: The network parameters $\boldsymbol{\theta}$ of MDNN.

and 2) hands are the principal parts of egocentric actions and the recognition accuracy is usually sensitive to hand poses. In order to leverage the hand hints, we extent our MDNN to MDNN+hand by integrating with a carefully designed hand-module.

Fig. 3 shows an overview of our MDNN+hand. Concretely, the hand-module takes the RGB frames as input. Firstly, it segments hands in these video frames and outputs the binary handmasks $\mathbf{M}_s$ by a fully convolutional networks (FCN) model [92]. Since there are at most two hands appearing in the cameras, we only retain the largest two connected components in these binary images. Sequentially, we utilize the segmentation masks $\mathbf{M}_s$ to black out the cluttered background of the corresponding RGB frames. The obtained results $\mathbf{B}$ are fed into another CNN architecture. We calculate the sum of per-pixel two-class softmax loss [92] for the FCN and the categorical cross-entropy loss for the CNN. In order to integrate the hand-module with the MDNN, we combine their softmax scores through a weighted fusion strategy:

$$\mathbf{s} = \gamma \mathbf{s}_h + \mathbf{s}_M, \tag{8}$$

where γ is a balanced hyper-parameter and the index of the max element in $\mathbf{s}$ indicates the final action label.

V. Experiments and Analysis

In this section, we conducted experiments on three egocentric RGB+D datasets for action recognition, including the proposed THU-READ, Wearable Computer Vision Systems dataset (WCVS) [43] and Grasp Understanding dataset (GUN71) [44]. The experimental results and analysis are described in details as follows.

A. Implementation Details

We implemented our MDNN model on the Keras toolbox [91] and trained the networks with 2 Nvidia Tesla K80 GPUs. For feature extraction, we adopted VGG-16 network[3] [93] pretrained on ImageNet [94]. Then we removed its last prediction layer, which was actually a fully-connected layer activated by a softmax function. We fed the input frames into the network and obtained the features $\{\boldsymbol{X}_i\}_{i=1}^K$ of the fc7 layers. We set the dimensions of the intermediate features $\{\boldsymbol{f}_i(\boldsymbol{X}_i)\}_{i=1}^K$ and $\boldsymbol{g}(\boldsymbol{X})$ to be 512. The weighted matrices $\mathbf{W}$ in multi-view learning stage were initialized by glorot uniform method [95]:

$$\mathbf{W}^{(m)} \sim \mathrm{U}[-\frac{\sqrt{6}}{\sqrt{p^{(m)}+p^{(m+1)}}}, \frac{\sqrt{6}}{\sqrt{p^{(m)}+p^{(m+1)}}}], \tag{9}$$

where $p^{(m)}$ was the size of the mth layer and the bias terms $\mathbf{b}$ were initialized to be zeros. We normalized $\boldsymbol{J}_s$ and $\boldsymbol{J}_d$ in all experiments and used $\lambda_1 = \lambda_2 = 1$. We employed stochastic gradient descent (SGD) method to optimize the parameters of whole network and assigned the value of initial learning rate and batchsize to 0.001 and 4, respectively.

For the hand-module, we applied FCN-8s model [92] to segment hands in RGB video frames. We adapted its last layer from a 21-dimensional output to a 2-dimensional output, as the the hand segmentation problem could be considered as a pixelwise binary classification problem between hand and background. The parameters of the FCN-8s model were initialized as those of the trained VGG-16 network [93] where the FCN-8s and VGG-16 net structures matched, and as random values for the rest. We finetuned the model on 1,391 pixel-wise labeled images, including 652 images from our THU-READ and 743 images from the EDSH dataset [96]. To fuse the hand-module with our MDNN, we empirically set the hyper-parameter γ in equation (8) to be 0.13, 0.39 and 0.15 on the THU-READ, WCVS and GUN-71 dataset, as hand cues played different importance in different datasets respectively.

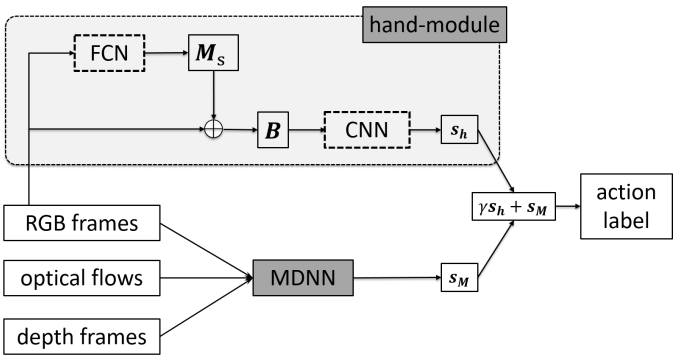


Fig. 3. **Flowchart of the MDNN+hand**. We segment the hands in the egocentric videos by the FCN model and utilize the handmask $\mathbf{M}_s$ to black out the non-hand regions. We feed the obtained results $\mathbf{B}$ into a CNN model and combine the classification scores of hand-module and MDNN by weighted fusion to predict the final action label.

[3]Configurations of the VGG16 Network: [block1_conv1, block1_conv2, block1_pool, block2_conv1, block2_conv2, block2_pool, block3_conv1, block3_conv2, block3_conv3, block3_pool, block4_conv1, block4_conv2, block4_conv3, block4_pool, block5_conv1, block5_conv2, block5_conv3, block5_pool, flatten, fc6, fc7, prediction], the ReLU activation function is not shown for brevity.

B. Datasets and Experimental Setup

1) THU-READ: We have described our THU-READ in Section III. In our previous work [22], we randomly sampled about 30% video clips for model training and used the other clips for testing. In order to provide a standard experimental setting, we formally defined two evaluation protocols in this work, including cross-group (CG) and cross-subject (CS) settings. In the CG setting, considering each subject performed each action for 3 times, we divided the videos samples into 3 groups, where a group was for training and the rest for testing. In the CS setting, we separated the 8 subjects into 4 splits and used samples from 3 splits for training and the other for testing. We calculated the recognition accuracy on all groups and splits, and their average results on the two settings respectively.

2) WCVS: Wearable Computer Vision Systems (WCVS) dataset [43] provides RGB-D egocentric video samples performed by 4 different subjects in 2 different scenarios. This dataset was organized into 3 levels of granularity. Level I contained two classes as manipulation and non-manipulation actions. At level II, the manipulation actions were divided into 4 categories and the non-manipulation actions were splited into other 6 classes. At the level III, the manipulation actions of level II were divided into finer categories, but the frequency of them was too low to train a classifier [43]. Therefore, we evaluated our methods on the manipulation actions of the level II [43]. Although there were only 4 action classes in this setting, the dataset presented great challenges due to the large intra-class variations caused by multiple users and scenarios. As suggested in [43], we conducted experiments on different methods with the leave-one-subject-out cross-validation scheme.

3) GUN-71: Grasp Understanding (GUN-71) dataset [44] is a challenging image-based dataset for egocentric hand-action understanding. We chose this image-based database due to the scarcity of the video-based egocentric RGB-D datasets for action recognition. The GUN-71 included roughly 12000 labeled RGB-D images captured from a chest-mounted RGB-D camera, which covered 71 everyday grasps. This dataset was performed by 8 different subjects. Following [44], we adopted the leave-one-out cross-validation protocol to evaluate our method and reported the average accuracy as the final result.

4) Compared methods: We mainly compared our approaches with two types of existing methods: hand-crafted features based methods and deep learning based methods. For the hand-crafted features based methods, we evaluated IDT [13] features on the first two video-based datasets, due to its better performance compared with the other exsiting hand-crafted spatio-temporal descriptors [33], [98]. We obtained the HOG [97], HOF [24] and MBH [32] features, which were extracted based on the trajectories of IDT, as well as their combination on both RGB and depth videos. Then, we employed the higher-dimensional encodings methods [99], with gmmSize set to 256 to generate good performance. We tested several encoding methods and finally chose the Fisher Vector (FV) [100] due to its higher performance than other encoding algorithms in [99]. For deep learning based methods, we presented the recognition performance on each single stream trained separately with the cross-entropy loss function. We also tested three multi-view learning methods, including score fusion, feature fusion and deep canonical correlation analysis (DCCA) method [51]. Here, score fusion denotes averaging the softmax classification scores of each single stream [22] and feature fusion represents averaging the multi-modal features of the last fully connected layers. For the GUN-71 dataset, we reported the performance of HOG descriptors, and evaluated deep learning methods based on the appearance and depth streams only, as there was no temporal information in this image-based dataset to be utilized.

C. Evaluation on the THU-READ

1) Results of Existing Methods: We first conducted experiments on various methods to build a benchmark for our proposed THU-READ. Table II tabulates the comparisons of classification accuracy for action recognition. For the hand-crafted feature based methods, the RGB and depth modalities both contribute important information for egocentric action recognition. The combined features [13] on RGB videos achieve more promising performances than those on depth videos, *e.g.*, they are respectively 6.36% and 4.06% higher on the CS and CG settings. Besides, we notice that the accuracy of different methods on CG setting are much higher than those on CS setting, this phenomenon is also found on the other models. It may be the reason that the actions performed by the same actor have higher similarity than those performed by different actors. For the deep learning based approaches, we first evaluated each single stream. From Table II we see that, the appearance stream achieves the value of 41.90% (CS) and 83.14% (CG), which is best of three the single streams. The performances of the motion stream are 37.19% (CS) and 72.81% (CG), while the depth stream achieves 34.06% (CS) and 78.38% (CG) respectively. This is similar to the results on hand-crafted feature based methods, which indicates the virtual importance of the RGB modality. We also adopted three multi-view learning methods to combine the three streams, including score fusion, feature fusion and DCCA method [51] described in Section V.B. We discover that combining three streams is capable to improve the performance of each single stream, which demonstrates their complementary property. Moreover, DCCA achieves better performances than the other two fusion methods, because it explicitly maximized the correlation of three modalities, so that their complementary properties are adequately leveraged.

2) Results of MDNN and MDNN+hand: We conducted experiments on our proposed methods. As described in the caption of Table II, MDNN^1 and MDNN^2 denotes adopting different hyper-parameters $\alpha_i (i = 1, 2, 3)$ and β. From the results we observe that, our proposed MDNN achieves promising performance compared with the existing approaches. Moreover, the MDNN+hand improves the performance in the most cases, which demonstrates the advantage of introducing the hand cues for helping egocentric action recognition. However, in some situations, the accuracy improvements are not significant when integrating with hand cues. This is mainly

TABLE II
COMPARISONS OF THE ACTION RECOGNITION ACCURACY (%) ON THE PROPOSED THU-READ DATASET. THE MDNN^1 AND MDNN^2 DENOTE ADOPTING DIFFERENT HYPER-PARAMETERS: MDNN^1 FOR $\alpha_i = \beta = 1/4(i = 1, 2, 3)$, MDNN^2 FOR $\alpha_i = 1/6(i = 1, 2, 3), \beta = 1/2$. THE HAND-CRAFTED FEATURES ARE BASED ON IDT [13].

Methods	**CS1**	**CS2**	**CS3**	**CS4**	**Average**	**CG1**	**CG2**	**CG3**	**Average**
HOG-RGB [97]	38.75	42.08	32.50	43.75	39.93	82.81	84.69	81.56	83.02
HOF-RGB [24]	37.08	44.58	45.00	45.00	46.27	73.75	76.09	73.28	74.37
MBH-RGB [32]	52.92	57.92	52.08	57.50	55.11	77.19	81.41	80.00	79.53
Combine-RGB [13]	53.33	56.25	52.50	62.50	56.15	86.09	88.75	87.03	87.29
HOG-Depth [97]	49.58	45.83	44.17	43.75	45.83	80.47	86.41	81.25	82.71
HOF-Depth [24]	42.92	44.17	46.25	42.50	43.96	73.91	78.59	70.94	84.48
MBH-Depth [32]	47.92	43.75	41.25	39.58	43.13	72.97	78.91	72.97	74.95
Combine-Depth [13]	51.25	49.17	50.00	48.75	49.79	82.19	86.09	81.41	83.23
Appearance Stream [93]	35.42	44.17	43.44	44.58	41.90	82.66	86.88	80.47	83.14
Depth Stream [93]	29.58	40.42	34.17	32.08	34.06	77.03	82.34	75.78	78.38
Motion Stream [93]	37.50	40.42	38.33	32.50	37.19	70.31	77.50	70.63	72.81
Depth Stream + Motion Stream	45.00	45.83	45.83	47.50	46.04	82.34	87.34	83.28	84.32
Appearance Stream + Motion Stream	50.83	59.58	49.17	60.83	55.10	89.53	90.00	87.34	88.96
Appearance Stream + Depth Stream	45.83	58.33	50.42	56.25	52.71	85.16	89.06	87.03	87.08
Score Fusion [22]	38.75	51.67	48.75	47.50	46.67	84.38	88.12	83.19	85.47
Feature Fusion	49.17	56.67	54.17	57.92	54.48	90.00	89.38	87.50	88.96
DCCA [51]	56.67	60.00	51.67	57.92	56.57	90.31	91.87	89.53	90.59
MMUDL [66]	53.33	61.25	52.50	62.08	57.29	90.00	92.03	88.91	90.31
DSSCA [15]	55.42	62.08	52.50	65.83	58.96	88.91	90.31	86.25	88.49
MDNN^1	62.92	63.33	57.92	63.75	61.98	90.00	91.72	89.84	90.52
MDNN^2	56.25	63.75	57.08	65.42	60.63	92.19	92.34	90.16	91.56
MDNN^1 + hand	**64.58**	63.33	**60.00**	64.58	**62.92**	90.78	92.03	90.00	90.94
MDNN^2 + hand	57.50	**67.50**	57.08	**67.92**	62.50	**92.19**	**92.66**	**90.31**	**91.72**

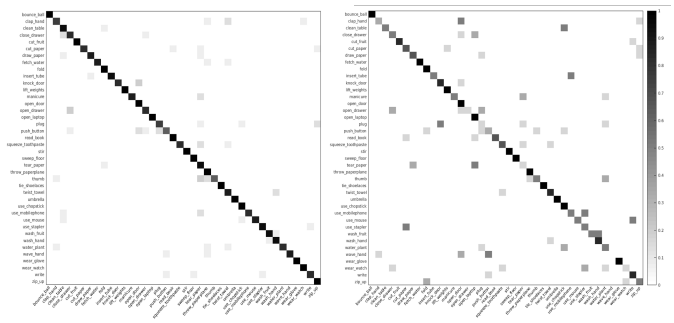


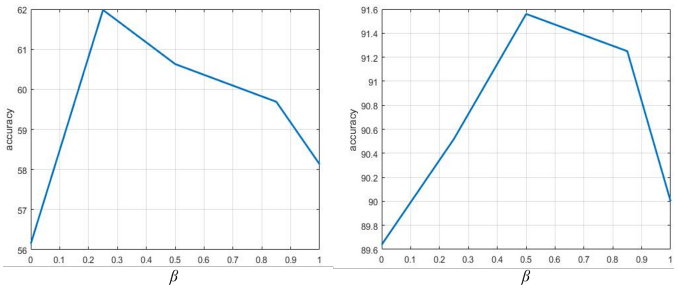
Fig. 4. **The confusion matrices of the MDNN^2 + hand method on THU-READ.** The result of group3 on CG setting is shown on the left and that of split4 on CS setting is presented on the right. The ground truth and the predicted labels are displayed on the vertical and horizontal axis respectively (best viewed in the pdf file).

because the failed segmentation may make bad influence on the recognition tasks. In the supplementary material, we further show some typical examples of the ground-truth hand annotation and the segmentation results, including some successful examples and some failed samples for visualization. In the future, it is an interesting work to employ some high-performance segmentation methods, like Mask R-CNN [101], to obtain hand masks with higher quality and further improve the recognition accuracy.

Fig. 4 shows the recognition confusion matrix obtained by MDNN^2+hand on group3 (CG) and split4 (CS). For group3 (CG), most actions are classified correctly except several actions like "push button" and "thumb". More specifically, "push button" is sometimes confused with the action "plug" (hand both interacted with the plug board for these two actions) and "thumb" is sometimes misclassified to "tear paper" (the action backgrounds are sometimes similar in our dataset). Compared with group3 (CG), the action recognition results on split4 (CS) are relatively low and several actions are often misclassified, such as "open drawer", "wave hand" and so on. This is mainly because the actions performed by different subjects have larger intra-class variance, which also shows the challenge of our proposed dataset on the CS setting.

3) Ablation Studies: In order to verify the importance of each single modality, we have conducted the ablation experiments. When we employ our proposed MDNN to fuse two streams, we empirically set $\beta = 1/2$ and $\alpha_1 = \alpha_2 = 1/4$ and the reason is explained in the later *"Parameter Analysis"* section. As shown in Table II, once we combine two streams together, the accuracy consistently increases based on the single streams. And finally the MDNN, which utilizes data on three modalities, achieves the higher results than employing two streams. This phenomenon has demonstrated the importance of each modality for action recognition, and shown the effectiveness of our method to explore the complementary information of different modalities.

4) Comparison with Existing Multi-view Learning Methods: As shown in Table II, the proposed MDNN outperforms the methods of score fusion and feature fusion. Compared with DCCA method [51], our MDNN obtains comparable results on the CG setting and is superior on the CS setting. This is because DCCA only utilizes the sharable information of different modalities, while the MDNN further preserves their distinctive characteristics. Thus, more information is retained



(a) Cross-subject Accuracy (b) Cross-group Accuracy

Fig. 5. The curves corresponding to Table III. For both Cross-Subject setting (left) and Cross-Group setting (right), the action recognition accuracy reaches peak when the sharable information and distinctive characteristics are simultaneously explored.

TABLE III
THE RELATIONSHIP BETWEEN DIFFERENT PARAMETERS AND THE ACTION RECOGNITION ACCURACY (%) ON THE THU-READ. MDNN$_{softmax}$ DENOTES EMPLOYING THE SOFTMAX LAYER TO CONSTRAIN $\sum_{i=1}^{3}\alpha_i+\beta=1$.

$\alpha_i(i=1,2,3)$	β	Cross-Subject	Cross-Group
1/3	0	56.15	89.64
1/4	1/4	**61.98**	90.52
1/6	1/2	60.63	**91.56**
1/20	17/20	59.69	91.25
0	1	58.13	90.00
MDNN$_{softmax}$		56.56	89.84

to enhance the recognition performance.

In recent years, numbers of RGBD-based methods have been proposed to exploit the sharable and individual components of different modalities, which mainly adopted l_2 norm to measure the distance between different components. Different from these works, we utilized the Cauchy estimator [20] to measure the correlations between the sharable components, and employed the orthogonality constraint to preserver the distinctive characteristics of different modalities. We have conducted experiments to compare with state-of-the-art RGBD-based methods DSSCA [15] and MMUDL [66] [4]. From Table II observer that, our MDNN have generally outperforms the state-of-the-art RGBD-based methods on the THU-READ dataset, which demonstrates the robustness of our fusion scheme.

5) Parameter Analysis: There are several important parameters in this work. In equation (4), the parameter β denotes the importance of the sharable information, while the $\alpha_i(i=1,2,3)$ represents the contributions of each distinctive component. We conducted several experiments to analyze how these parameters influenced the performance of the recognition. Table III and Fig. 5 show the performances on the CG and CS settings by allocating different values to the parameters, which satisfy $\Sigma_{i=1}^{3}\alpha_i+\beta=1$ and $0\leqslant\alpha_1,\alpha_2,\alpha_3,\beta\leqslant 1$. For simplicity, we set $\alpha_1=\alpha_2=\alpha_3$. We notice that, when β is close to 0, the accuracy is relatively low since the sharable component of the three modalities plays less significance role.

[4]See the supplementary materiel for the implementation details.

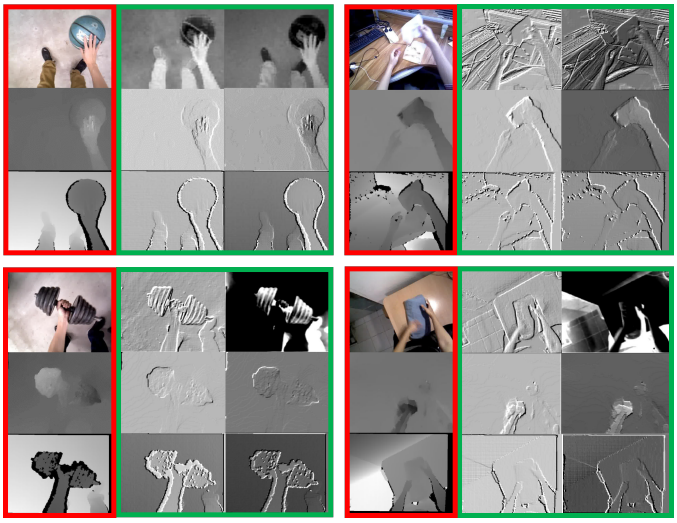


Fig. 6. **Visualization of feature maps extracted from each single stream.** We display the results of four groups. For each group, the input video frames are shown on the left, while their Conv1 feature maps are presented on the right. The results on the top, middle and bottom are corresponding to the modalities of RGB, optical flow and depth, respectively.

When β increases, the recognition accuracy will reach the peak. However, when β approaches to 1, the performance will go down due to the vanishing of the distinctive components. This phenomenon demonstrates the importance of both sharable information and distinctive characteristics.

Moreover, we have further conducted experiments by using a softmax layer to constrain $\sum_{i=1}^{3}\alpha_i+\beta=1$. [5] As shown in Table III, we observe that using the softmax layer performs worse than the hand-designed strategies MDNN^1 ($\alpha_i=\beta=1/4, i=1,2,3$) and MDNN^2 ($\alpha_i=1/6,\beta=1/2,i=1,2,3$). It may be the reason that, the fc layers before the softmax layer brought more parameters to train, thus more training data are needed to obtain the better results.

In the later experiments, we chose the parameter settings as $\beta=1/2$, and $\alpha_i=\alpha_j(i\neq j)$ (the parameter settings for MDNN^2). This is because in the Table II, on the cross-subject scenario, MDNN^2 performs better than MDNN^1 on CS2 and CS4, and the results are opposite when it comes to CS1 and CS3. On the cross-group scenario, the results of MDNN^2 are consistently higher than those of MDNN^1.

6) Visualization: We also performed some feature visualization to qualitatively evaluate our proposed method. We extracted the feature maps from the Conv1 layer of each single stream. Fig. 6 shows the visualization results, from which we observe that each stream is capable to capture some distinctive characteristics of its corresponding modality. To be specific, while the RGB features mainly carry the color and textural information, the feature maps of optical flows encode some motion patterns, especially the movement of hand and its interacted object. Moreover, the depth features reflect the 3D structures of the scene according to the distances of different

[5] We first sent the intermediate features $\boldsymbol{f}_1(\boldsymbol{X}_1)$, $\boldsymbol{f}_2(\boldsymbol{X}_2)$, $\boldsymbol{f}_3(\boldsymbol{X}_3)$ and $\boldsymbol{g}(\boldsymbol{X})$ into four fc layers as: $s_i=tanh(\boldsymbol{W}_i\boldsymbol{f}_i(\boldsymbol{X}_i)+\mathbf{b}_i), i=1,2,3, s_4=tanh(\boldsymbol{W}_4\boldsymbol{g}(\boldsymbol{X})+\mathbf{b}_4)$. Then we fed $s_i(i=1,2,3,4)$ into a softmax layer to obtain $\alpha_i(i=1,2,3)$ and β as: $[\alpha_1,\alpha_2,\alpha_3,\beta]=softmax([s_1,s_2,s_3,s_4])$.

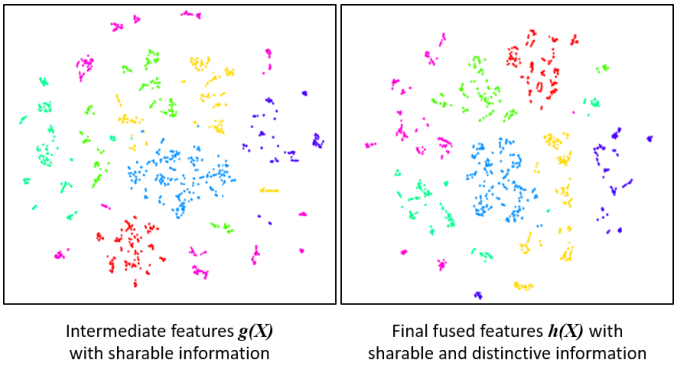


Fig. 7. The t-SNE visualizations [21] of 7 classes randomly selected from the THU-READ. The left figure shows the intermediate features $\boldsymbol{g}(\boldsymbol{X})$ with sharable information, while the right figure presents the final fused features $\boldsymbol{h}(\boldsymbol{X})$ learned by the MDNN (best viewed in the color pdf file).

objects to the wearable camera.

Also, we employed t-SNE methods [21] to compare the intermediate features $\boldsymbol{g}(\boldsymbol{X})$ with the sharable information and the final fused features $\boldsymbol{h}(\boldsymbol{X})$. Fig. 7 presents the visualization results of 7 classes randomly selected from the THU-READ. Compared with the intermediate features, the final fused features perform better to enlarge the margin among the inter-class samples and reduce the intra-class distance, which demonstrates the discriminative power of our MDNN.

D. Evaluation on the WCVS Dataset

1) Results of Existing Methods: Table IV tabulates our experimental results and some exsiting results reported in [43]. For the hand-crafted based features, the IDT features achieves the highest recognition accuracy, *e.g.,* 59.26% on RGB modality and 52.79% on depth modality. For the deep learning methods, the appearance stream performs better than the CNN-RGB and CNN(MULTIWINDOW)-RGB [43]. This is mainly because the appearance stream was implemented by the VGG-16 network [93], which has deeper architecture to model the non-linear information for action recognition than the DeCAF model [102] used in [43]. For the three multi-view learning methods, we find that the feature fusion method and DCCA [51] method improve the recognition accuracy of the appearance stream (*i.e.* the best single stream) by 1.80% and 0.09% respectively, while the score fusion approach degrades the performance by 1.04%. This is due to the reason that score fusion is lack of physical meaning and may cause the negative effect sometimes.

2) Results of MDNN and MDNN+hand: We evaluated our proposed methods on the WCVS dataset with the same parameter setting of MDNN[2] in the THU-READ as $\alpha_1 = \alpha_2 = \alpha_3 = 1/6$ and $\beta = 1/2$. Table IV shows that, our MDNN model attains the performance of 65.67% and MDNN+hand brings 1.37% improvement, which achieves favourable performance compared with other methods for action recognition. This demonstrates the advantages of our fusion strategy.

3) Analysis on Different Estimators: To investigate the robustness of the Cauchy estimator in equation (6), we further conducted experiments on other two estimators. Table V tabulates the comparison results, which shows that the Cauchy estimator achieves the recognition accuracy of 65.67%, surpassing the L_1 and L_2 distance estimators by 2.07% and 2.79% respectively. This indicates that the Cauchy estimator is more robust to the outliers than the other two estimators for modelling the correlation of different modalities.

TABLE IV
COMPARISONS OF THE ACTION RECOGNITION ACCURACY (%) ON THE WCVS DATASET.

Methods	Cross-Subject Accuracy
SKIN_HIST-RGB-D [43]	27.00
GIST-RGB [43]	35.00
SIFT-RGB [43]	44.00
HOG-RGB [97]	52.14
HOF-RGB [24]	48.50
MBH-RGB [32]	54.36
Combine-RGB [13]	59.26
HOG-Depth [97]	50.61
HOF-Depth [24]	41.25
MBH-Depth [32]	44.46
Combine-Depth [13]	52.79
CNN-RGB [43]	52.00
CNN(MULTIWINDOW)-RGB [43]	57.00
Appearance Stream [93]	60.36
Depth Stream [93]	58.47
Motion Stream [93]	37.34
Score Fusion [22]	59.32
Feature Fusion	62.16
DCCA [51]	60.45
MDNN	**65.67**
MDNN + hand	**67.04**

TABLE V
COMPARISION OF ACTION RECOGNITION ACCURACY (%) ON DIFFERENT ESTIMATORS ON THE WCVS DATASET.

Estimators	Cross-Subject Accuracy
L_1 Distance Estimator	63.60
L_2 Distance Estimator	62.88
Cauchy Estimator	**65.67**

E. Evaluation on the GUN-71 Dataset

1) Results and Analysis: We also conducted experiments on the GUN-71 dataset. As GUN-71 is an image-based dataset, the MDNN only consists of the appearance stream and depth stream. Similar to the two previous datasets, we set $\beta = 1/2$ and $\alpha_1 = \alpha_2 = 1/4$. As shown in the Table VI, our MDNN and MDNN+hand achieve the recognition results of and 33.89% and 34.04% respectively, consistently outperforming the other existing methods. Fig. 8 displays the visualization of the best results in [44] and our proposed approach, which presents the significant improvement over the previous work.

2) Comparison with Existing Multi-view Learning Methods: We tested the three fusion approaches on this dataset. The recognition results of score fusion, feature fusion and DCCA are 26.24, 29.36 and 32.56 respectively. The accuracies of the first two methods are relatively poor, this is mainly because the performance of the depth stream (15.60%) is much lower

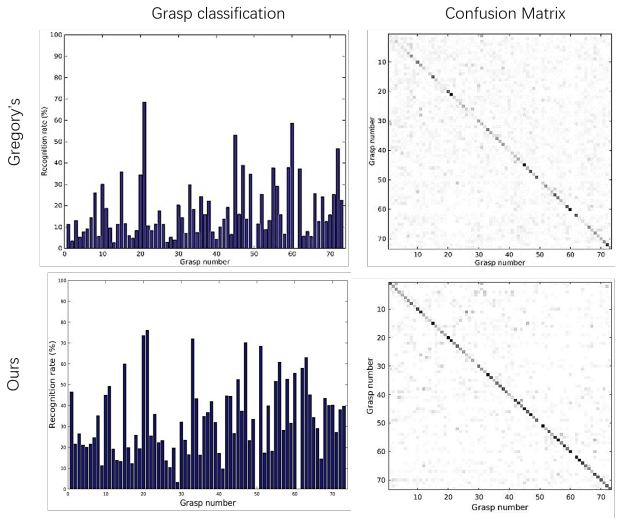


Fig. 8. Visualizations of best results reported in [44] (upper row) and our MDNN+hand approach (lower row). The recognition accuracies of the 71 hand actions are shown on the left, while the corresponding confusion matrices are presented on the right.

TABLE VI
COMPARISONS OF THE ACTION RECOGNITION ACCURACY (%) ON THE GUN-71 DATASET.

Methods	**Cross-Subject Accuracy**
Deep-RGB [44]	11.31
Best from [44]	17.97
Appearance Stream [103]	26.00
Depth Stream [103]	15.60
Score Fusion [22]	26.24
Feature Fusion	29.36
DCCA [51]	32.56
MDNN	**33.89**
MDNN+hand	**34.04**

than that of the appearance stream (26.00%). Since the score fusion method roughly averages the classification probabilities of the two streams and the DCCA method aims to maximize the correlation between the two modalities, they may retain the negative property of the depth stream and harm the fusion results. In comparison, our proposed MDNN further preserves the distinctive properties for different modalities, which can be regarded as a compensation when the sharable information does not perform well. Therefore, the MDNN is capable to bring 1.33% improvement over the feature fusion method and avoid the problem mentioned above.

F. Analysis

In previous sections, we have shown the effectiveness of the fusion strategy of our MDNN compared with other multi-view learning methods. In this section, we have conducted experiments on more state of the art methods, including TSN [90], SeDyn [104] and EgoTDD [81]. These methods were originally designed for the color videos, in order to evaluate them on the depth modality, we processed the depth frames in the same way as the RGB frames. Table VII displays the experimental results on the cross-subject settings

TABLE VII
COMPARISONS OF ACTION RECOGNITION ACCURACY (%) WITH THE STATE-OF-THE-ARTS ON THE CROSS-SUBJECT SETTING OF THU-READ AND WCVS DATASET.

Method	THU-READ	WCVS
EgoTDD-RGB [81]	62.81	57.02
EgoTDD-Depth [81]	54.58	55.04
SeDyn-RGB [104]	66.67	58.42
SeDyn-Depth [104]	47.50	55.06
TSN-RGB [90]	73.85	66.02
TSN-Flow [90]	62.39	59.48
TSN-Depth [90]	65.00	59.32
TSN-Flow+RGB (score fusion) [90]	78.23	67.05
TSN-Flow+RGB+Depth(score fusion) [90]	81.67	70.09
MDNN + TSN	**83.54**	**71.83**

of THU-READ and WCVS dataset. We observe that the TSN-RGB [90] achieves 73.85% and 66.02% recognition accuracies on the THU-READ and WCVS dataset, which performs best among the methods based on single modality. In particular, this model combined a sparse temporal sampling strategy and video-level supervision to model the long-range temporal structure. It also utilized several engineering techniques (*e.g.*, partial batch normalization and data augmentation) to generate good performance. It is remarkable that these techniques should be orthogonal to our method. This is because our MDNN, which is equipped with simple baseline models for the single streams, mainly focuses on exploring the complementary information of different modalities and fusing them more effectively. To demonstrate this, we further conducted experiments by replacing the score fusion strategy in TSN model with the multi-view learning method in our MDNN (see supplementary material for details). As shown in Table VII, MDNN+TSN achieves the accuracy of 83.54% and 71.83% on the THU-READ and WCVS dataset, outperforming all the single modalities and the score fusion strategy of TSN. This corroborates the advantages of our proposed method, and further illustrates that it can be easily adopted during the fusion strategy of the other state-of-the-arts.

VI. CONCLUSION AND FUTURE WORK

In this paper, we have investigated the problem of RGB-D egocentric action recognition. We have collected a dataset called THU-READ with over 340K video frames and 200M-pixel hand annotation, which is the currently largest RGB-D egocentric action dataset. In order to adequately exploit the complementary information of the static appearance, temporal information and depth property, we have proposed a multi-stream deep neural networks (MDNN) method, which aims to simultaneously mine the sharable information and distinctive characteristic of different modalities. Moreover, we have extended our MDNN by integrating with the hand cues to further enhance the recognition accuracy. The experiments achieved superior performance in comparison with the state-of-the-art methods on our proposed THU-READ and additional two datasets. In the future, we will extend our THU-READ to a larger dataset with more participants and more challenges.

Moreover, it is interesting to explore the self-adapted methods to decide the importance of each modality and mine the semantic cues more adequately for human-object interaction task on our dataset.

ACKNOWLEDGEMENT

This work was supported in part by the National Key Research and Development Program of China under Grant 2017YFA0700802, in part by the National Natural Science Foundation of China under Grant 61822603, Grant U1713214, Grant 61672306, Grant 61572271, and Grant 61527808, in part by the National 1000 Young Talents Plan Program, and in part by the Shenzhen Fundamental Research Fund (Subject Arrangement) under Grant JCYJ20170412170602564. The authors would like to thank all the participants for dataset collection, Dr. Hao Liu, Dr. Tianfu Wu and Yang Liu for valuable discussions, Yu Zheng and Yi Tian for data preparation and conducting partial experiments in this paper.

REFERENCES

[1] Betancourt, A., Morerio, P., Regazzoni, C.S., Rauterberg, M.: The evolution of first person vision methods: A survey. IEEE Transactions on Circuits and Systems for Video Technology **25**(5) (2015) 744–760

[2] Halperin, T., Poleg, Y., Arora, C., Peleg, S.: Egosampling: Wide view hyperlapse from egocentric videos. IEEE Transactions on Circuits and Systems for Video Technology **28**(5) (2018) 1248–1259

[3] Lu, Z., Grauman, K.: Story-driven summarization for egocentric video. In: CVPR. (2013) 2714–2721

[4] Fathi, A., Ren, X., Rehg, J.M.: Learning to recognize objects in egocentric activities. In: CVPR. (2011) 3281–3288

[5] Fathi, A., Farhadi, A., Rehg, J.M.: Understanding egocentric activities. In: ICCV. (2011) 407–414

[6] Pirsiavash, H., Ramanan, D.: Detecting activities of daily living in first-person camera views. In: CVPR. (2012) 2847–2854

[7] Chakraborty, A., Mandal, B., Yuan, J.: Person reidentification using multiple egocentric views. IEEE Transactions on Circuits and Systems for Video Technology (2017) 484–498

[8] Li, Y., Fathi, A., Rehg, J.M.: Learning to predict gaze in egocentric video. In: ICCV. (2013) 3216–3223

[9] Bernal, E.A., Yang, X., Li, Q., Kumar, J., Madhvanath, S., Ramesh, P., Bala, R.: Deep temporal multimodal fusion for medical procedure monitoring using wearable sensors. IEEE Transactions on Multimedia **20**(1) (2018) 107–118

[10] Gemmell, J., Bell, G., Lueder, R.: Mylifebits: a personal database for everything. Communications of the ACM **49**(1) (2006) 88–95

[11] Surie, D., Pederson, T., Lagriffoul, F., Janlert, L.E., Sjölie, D.: Activity recognition using an egocentric perspective of everyday objects. In: UIC. (2007) 246–257

[12] Fathi, A., Li, Y., Rehg, J.M.: Learning to recognize daily actions using gaze. In: ECCV. (2012) 314–327

[13] Wang, H., Schmid, C.: Action recognition with improved trajectories. In: ICCV. (2013) 3551–3558

[14] S, K., Zisserman, A.: Two-stream convolutional networks for action recognition in videos. In: NIPS. (2014) 568–576

[15] Shahroudy, A., Ng, T.T., Gong, Y., Wang, G.: Deep multimodal feature analysis for action recognition in rgb+d videos. IEEE Transactions on Pattern Analysis and Machine Intelligence **40**(5) (2018) 1045–1058

[16] Hu, J., Zheng, W., Lai, J., Zhang, J.: Jointly learning heterogeneous features for RGB-D activity recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence **39**(11) (2017) 2186–2200

[17] Zhang, H., Parker, L.E.: Code4d: Color-depth local spatio-temporal features for human activity recognition from RGB-D videos. IEEE Transactions on Circuits and Systems for Video Technology **26**(3) (2016) 541–555

[18] Liu, M., Liu, H., Chen, C.: Robust 3d action recognition through sampling local appearances and global distributions. IEEE Transactions on Multimedia **20**(8) (2018) 1932–1947

[19] Neverova, N., Wolf, C., Taylor, G., Nebout, F.: Moddrop: adaptive multi-modal gesture recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence **38**(8) (2014) 1692–1706

[20] Mizera, I., MÃijller, C.H.: Breakdown points of cauchy regression-scale estimators. Statistics and Probability Letters **57**(1) (2002) 79–89

[21] Maaten, L.V.D., Hinton, G.: Visualizing data using t-sne. Journal of Machine Learning Research **9**(2605) (2008) 2579–2605

[22] Tang, Y., Tian, Y., Lu, J., Feng, J., Zhou, J.: Action recognition in rgb-d egocentric videos. In: ICIP. (2017)

[23] Laptev, I.: On space-time interest points. International Journal of Computer Vision **64**(2-3) (2005) 107–123

[24] Laptev, I., Marszalek, M., Schmid, C., Rozenfeld, B.: Learning realistic human actions from movies. In: CVPR. (2008) 1–8

[25] Li, X., Chuah, M.C.: SBGAR: semantics based group activity recognition. In: ICCV. (2017) 2895–2904

[26] Li, X., Chuah, M.C.: ReHAR: robust and efficient human activity recognition. In: WACV. (2018) 362–371

[27] Xu, K., Jiang, X., Sun, T.: Two-stream dictionary learning architecture for action recognition. IEEE Transactions on Circuits and Systems for Video Technology **27**(3) (2017) 567–576

[28] Wang, P., Cao, Y., Shen, C., Liu, L., Shen, H.T.: Temporal pyramid pooling-based convolutional neural network for action recognition. IEEE Transactions on Circuits and Systems for Video Technology **27**(12) (2017) 2613–2622

[29] Feng, Q., Zhou, Y.: Kernel regularized data uncertainty for action recognition. IEEE Transactions on Circuits and Systems for Video Technology **27**(3) (2017) 577–588

[30] Tang, Y., Tian, Y., Lu, J., Li, P., Zhou, J.: Deep progressive reinforcement learning for skeleton-based action recognition. In: CVPR. (2018)

[31] Tang, Y., Wang, Z., Li, P., Lu, J., Yang, M., Zhou, J.: Mining semantics-preserving attention for group activity recognition. In: ACM MM. (2018)

[32] Dalal, N., Triggs, B., Schmid, C.: Human detection using oriented histograms of flow and appearance. In: ECCV. (2006) 428–441

[33] Wang, H., Kläser, A., Schmid, C., Liu, C.L.: Action recognition by dense trajectories. In: CVPR. (2011) 3169–3176

[34] Karpathy, A., Toderici, G., Shetty, S., Leung, T., Sukthankar, R., Fei-Fei, L.: Large-scale video classification with convolutional neural networks. In: CVPR. (2014) 1725–1732

[35] Ji, S., Xu, W., Yang, M., Yu, K.: 3d convolutional neural networks for human action recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence **35**(1) (2013) 221–231

[36] Tran, D., Bourdev, L., Fergus, R., Torresani, L., Paluri, M.: Learning spatiotemporal features with 3d convolutional networks. In: ICCV. (2015) 4489–4497

[37] Donahue, J., Hendricks, L.A., Rohrbach, M., Venugopalan, S., Guadarrama, S., Saenko, K., Darrell, T.: Long-term recurrent convolutional networks for visual recognition and description. IEEE Transactions on Pattern Analysis and Machine Intelligence **39**(4) (2014) 677–691

[38] Spriggs, E.H., De La Torre, F., Hebert, M.: Temporal segmentation and activity classification from first-person sensing. In: CVPR Workshops. (2009) 17–24

[39] Kitani, K.M., Okabe, T., Sato, Y., Sugimoto, A.: Fast unsupervised ego-action learning for first-person sports videos. In: CVPR. (2011) 3241–3248

[40] Fathi, A., Hodgins, J.K., Rehg, J.M.: Social interactions: A first-person perspective. In: CVPR. (2012) 1226–1233

[41] Ryoo, M.S., Matthies, L.: First-person activity recognition: What are they doing to me? In: CVPR. (2013) 2730–2737

[42] Poleg, Y., Arora, C., Peleg, S.: Temporal segmentation of egocentric videos. In: CVPR. (2014) 2537–2544

[43] Moghimi, M., Azagra, P., Montesano, L., Murillo, A.C., Belongie, S.: Experiments on an rgb-d wearable vision system for egocentric activity recognition. In: CVPR Workshops. (2014) 597–603

[44] Rogez, G., Supancic, J.S., Ramanan, D.: Understanding everyday hands in action from rgb-d images. In: ICCV. (2015) 3889–3897

[45] Ohnishi, K., Kanehira, A., Kanezaki, A., Harada, T.: Recognizing activities of daily living with a wrist-mounted camera. In: CVPR. (2016) 3103–3111

[46] Bambach, S., Lee, S., Crandall, D.J., Yu, C.: Lending a hand: Detecting hands and recognizing activities in complex egocentric interactions. In: ICCV. (2015) 1949–1957

[47] Katsuyuki, N., Serena, Y., Alexandre, A., Li, F.F.: Jointly learning energy expenditures and activities using egocentric multimodal signals. In: ICCV. (2017)

[48] Garcia-Hernando, G., Yuan, S., Baek, S., Kim, T.: First-person hand action benchmark with RGB-D videos and 3d hand pose annotations. In: CVPR. (2018)

[49] Xu, C., Tao, D., Xu, C.: Multi-view intact space learning. IEEE Transactions on Pattern Analysis and Machine Intelligence **37**(12) (2015) 2531–2544

[50] Lai, P.L., Fyfe, C.: Kernel and nonlinear canonical correlation analysis. International Journal of Neural Systems **10**(5) (2000) 365–377

[51] Andrew, G., Arora, R., Bilmes, J., Livescu, K.: Deep canonical correlation analysis. In: ICML. (2013) 1247–1255

[52] Wang, W., Arora, R., Livescu, K., Bilmes, J.: On deep multi-view representation learning. In: ICML. (2015) 1083–1092

[53] Han, Y., Wu, F., Tao, D., Shao, J., Zhuang, Y., Jiang, J.: Sparse unsupervised dimensionality reduction for multiple view data. IEEE Transactions on Circuits and Systems for Video Technology **22**(10) (2012) 1485–1496

[54] Wei, X., Jiang, Y., Ngo, C.: Concept-driven multi-modality fusion for video search. IEEE Transactions on Circuits and Systems for Video Technology **21**(1) (2011) 62–73

[55] Pala, F., Satta, R., Fumera, G., Roli, F.: Multimodal person reidentification using RGB-D cameras. IEEE Transactions on Circuits and Systems for Video Technology **26**(4) (2016) 788–799

[56] Zhang, Z., Zhang, W., Liu, J., Tang, X.: Multiview facial landmark localization in rgb-d images via hierarchical regression with binary patterns. IEEE Transactions on Circuits and Systems for Video Technology **24**(9) (2014) 1475–1485

[57] Liu, H., Lu, J., Feng, J., Zhou, J.: Two-stream transformer networks for video-based face alignment. IEEE Transactions on Pattern Analysis and Machine Intelligence (2017) doi: 10.1109/TPAMI.2017.2737538

[58] Lu, J., Liong, V.E., Zhou, J.: Simultaneous local binary feature learning and encoding for homogeneous and heterogeneous face recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence **40**(8) (2018) 1979–1993

[59] Lu, J., Wang, G., Moulin, P.: Localized multifeature metric learning for image-set-based face recognition. IEEE Transactions on Circuits and Systems for Video Technology **26**(3) (2016) 529–540

[60] Yang, Y., Song, J., Huang, Z., Ma, Z., Sebe, N., Hauptmann, A.G.: Multi-feature fusion via hierarchical regression for multimedia analysis. IEEE Transactions on Multimedia **15**(3) (2013) 572–581

[61] Dong, X., Zheng, L., Ma, F., Yang, Y., Meng, D.: Few-example object detection with model communication. IEEE Transactions on Pattern Analysis and Machine Intelligence (2018) doi: 10.1109/TPAMI.2018.2844853

[62] Kakade, S.M., Foster, D.P.: Multi-view regression via canonical correlation analysis. In: COLT. (2007) 82–96

[63] Chaudhuri, K., Kakade, S.M., Livescu, K., Sridharan, K.: Multi-view clustering via canonical correlation analysis. In: ICML. (2009) 129–136

[64] Hu, J., Lu, J., Tan, Y.P.: Sharable and individual multi-view metric learning. IEEE Transactions on Pattern Analysis and Machine Intelligence **40**(9) (2018) 2281–2288

[65] Wang, A., Lu, J., Cai, J., Cham, T., Wang, G.: Large-margin multimodal deep learning for RGB-D object recognition. IEEE Transactions on Multimedia **17**(11) (2015) 1887–1898

[66] Ren, L., Lu, J., Feng, J., Zhou, J.: Multi-modal uniform deep learning for RGB-D person re-identification. Pattern Recognition **72** (2017) 446–457

[67] Liu, L., Shao, L.: Learning discriminative representations from RGB-D video data. In: IJCAI. (2013) 1493–1500

[68] Kong, Y., Fu, Y.: Bilinear heterogeneous information machine for RGB-D action recognition. In: CVPR. (2015) 1054–1062

[69] Furnari, A., Farinella, G.M., Battiato, S.: Temporal segmentation of egocentric videos to highlight personal locations of interest. In: ECCV Workshop. (2016) 474–489

[70] Furnari, A., Battiato, S., Farinella, G.M.: Personal-location-based temporal segmentation of egocentric videos for lifelogging applications. Journal of Visual Communication and Image Representation **52** (2018) 1–12

[71] Furnari, A., Farinella, G.M., Battiato, S.: Recognizing personal locations from egocentric videos. IEEE Transactions on Human-Machine Systems **47**(1) (2017) 6–18

[72] Furnari, A., Battiato, S., Farinella, G.M.: On the exploitation of hidden markov models to improve location-based temporal segmentation of egocentric videos. In: The Workshop on Wearable Multimedia. (2017) 1–4

[73] Fan, H., Chang, X., Cheng, D., Yang, Y., Xu, D., Hauptmann, A.G.: Complex event detection by identifying reliable shots from untrimmed videos. In: ICCV. (2017) 736–744

[74] Furnari, A., Battiato, S., Farinella, G.M.: How shall we evaluate egocentric action recognition? In: ICCVW. (2017) 2373–2382

[75] Damen, D., Doughty, H., Farinella, G.M., Fidler, S., Furnari, A., Kazakos, E., Moltisanti, D., Munro, J., Perrett, T., Price, W., Wray, M.: Scaling egocentric vision: The epic-kitchens dataset. In: ECCV. (2018)

[76] Huang, S., Wang, W., He, S., Lau, R.W.H.: Egocentric temporal action proposals. IEEE Transactions on Image Processing **27**(2) (2018) 764–777

[77] Bertasius, G., Chan, A., Shi, J.: Egocentric basketball motion planning from a single first-person image. In: CVPR. (2018)

[78] Rhinehart, N., Kitani, K.M.: First-person activity forecasting with online inverse reinforcement learning. In: ICCV. (2017) 3716–3725

[79] Possas, R., Pinto Caceres, S., Ramos, F.: Egocentric activity recognition on a budget. In: CVPR. (2018)

[80] Zhou, Y., Ni, B., Hong, R., Yang, X., Tian, Q.: Cascaded interactional targeting network for egocentric video analysis. In: CVPR. (2016) 1904–1913

[81] Singh, S., Arora, C., Jawahar, C.: First person action recognition using deep learned descriptors. In: CVPR. (2016) 2620–2628

[82] Ma, M., Fan, H., Kitani, K.M.: Going deeper into first-person activity recognition. In: CVPR. (2016) 1894–1903

[83] Horn, B.K.P., Schunck, B.G.: Determining optical flow. Artificial Intelligence **17**(1-3) (1981) 185–203

[84] Zach, C., Pock, T., Bischof, H.: A duality based approach for realtime tv-l1 optical flow. In: DAGM. (2007) 214–223

[85] Dosovitskiy, A., Fischer, P., Ilg, E., Häusser, P., Hazirbas, C., Golkov, V., van der Smagt, P., Cremers, D., Brox, T.: Flownet: Learning optical flow with convolutional networks. In: ICCV. (2015) 2758–2766

[86] Ilg, E., Mayer, N., Saikia, T., Keuper, M., Dosovitskiy, A., Brox, T.: Flownet 2.0: Evolution of optical flow estimation with deep networks. In: CVPR. (2017) 1647–1655

[87] Hui, T.W., Tang, X., Change Loy, C.: Liteflownet: A lightweight convolutional neural network for optical flow estimation. In: CVPR. (2018)

[88] Dong, X., Yu, S.I., Weng, X., Wei, S.E., Yang, Y., Sheikh, Y.: Supervision-by-registration: An unsupervised approach to improve the precision of facial landmark detectors. In: CVPR. (2018)

[89] Jain, S.D., Xiong, B., Grauman, K.: Fusionseg: Learning to combine motion and appearance for fully automatic segmentation of generic objects in videos. In: CVPR. (2017) 2117–2126

[90] Wang, L., Xiong, Y., Wang, Z., Qiao, Y., Lin, D., Tang, X., Gool, L.V.: Temporal segment networks: Towards good practices for deep action recognition. In: ECCV. (2016) 20–36

[91] Chollet, F., et al.: Keras. https://github.com/keras-team/keras (2015)

[92] Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: CVPR. (2015) 3431–3440

[93] Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: ICLR. (2015) 1–14

[94] Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A., Li, F.: Imagenet large scale visual recognition challenge. International Journal of Computer Vision **115**(3) (2015) 211–252

[95] Glorot, X., Bengio, Y.: Understanding the difficulty of training deep feedforward neural networks. In: AISTATS. (2010) 249–256

[96] Li, C., Kitani, K.M.: Pixel-level hand detection in ego-centric videos. In: CVPR. (2013) 3570–3577

[97] Wang, H., Ullah, M.M., Kläser, A., Laptev, I., Schmid, C.: Evaluation of local spatio-temporal features for action recognition. In: BMVC. (2009) 1–11

[98] Yeffet, L., Wolf, L.: Local trinary patterns for human action recognition. In: ICCV. (2009) 492–497

[99] Peng, X., Wang, L., Wang, X., Qiao, Y.: Bag of visual words and fusion methods for action recognition: Comprehensive study and good practice. Computer Vision and Image Understanding **150** (2016) 109–125

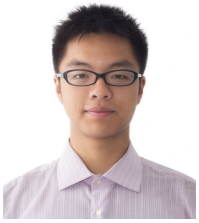
[100] Perronnin, F., Mensink, T.: Improving the fisher kernel for large-scale image classification. In: ECCV. (2010) 143–156

[101] He, K., Gkioxari, G., Dollár, P., Girshick, R.B.: Mask R-CNN. In: ICCV. (2017) 2980–2988

[102] Donahue, J., Jia, Y., Vinyals, O., Hoffman, J., Zhang, N., Tzeng, E., Darrell, T.: Decaf: A deep convolutional activation feature for generic visual recognition. In: ICML. (2014) 647–655

[103] He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR. (2016) 770–778

[104] Zaki, H.F.M., Shafait, F., Mian, A.S.: Modeling sub-event dynamics in first-person action recognition. In: CVPR. (2017) 1619–1628



Yansong Tang received the B.S. degrees in 2015 from the Department of Automation, Tsinghua University, Beijing, China, where he is currently pursuing the Ph.D. degree with the Department of Automation. His research lies in computer vision, especially multi-modal action recognition and egocentric vision analytics.

Zian Wang is an undergraduate student currently pursuing the B.S. degree from the Department of Automation, Tsinghua University, Beijing, China. His research interests include computer vision and machine learning.

Jiwen Lu (M'11-SM'15) received the B.Eng. degree in mechanical engineering and the M.Eng. degree in electrical engineering from the Xi'an University of Technology, Xi'an, China, in 2003 and 2006, respectively, and the Ph.D. degree in electrical engineering from Nanyang Technological University, Singapore, in 2012. He is currently an Associate Professor with the Department of Automation, Tsinghua University, Beijing, China. His current research interests include computer vision, pattern recognition, and machine learning. He has authored/co-authored over 200 scientific papers in these areas, where 60 of them are IEEE Transactions papers (including 11 T-PAMI papers) and 40 of them are CVPR/ICCV/ECCV/NIPS papers. He serves an Associate Editor for several international journals including the IEEE Transactions on Circuits and Systems for Video Technology, the IEEE Transactions on Biometrics, Behavior, and Identity Science, and Pattern Recognition. He is a member of the Multimedia Signal Processing Technical Committee and the Information Forensics and Security Technical Committee of the IEEE Signal Processing Society, and a member of the Multimedia Systems and Applications Technical Committee of the IEEE Circuits and Systems Society. He also served as Workshop Chair/Special Session Chair/Area Chair for more than 20 international conferences such as ICIP, ICPR, ICME, ACCV and WACV. He was a recipient of the National 1000 Young Talents Program of China in 2015, and the National Science Fund of China for Excellent Young Scholars in 2018, respectively. He is a senior member of the IEEE.

Jianjiang Feng is an associate professor in the Department of Automation at Tsinghua University, Beijing. He received the B.S. and Ph.D. degrees from the School of Telecommunication Engineering, Beijing University of Posts and Telecommunications, China, in 2000 and 2007. From 2008 to 2009, he was a Post Doctoral researcher in the PRIP lab at Michigan State University. He is an Associate Editor of Image and Vision Computing. His research interests include fingerprint recognition and computer vision.

Jie Zhou (M'01-SM'04) received the BS and MS degrees both from the Department of Mathematics, Nankai University, Tianjin, China, in 1990 and 1992, respectively, and the PhD degree from the Institute of Pattern Recognition and Artificial Intelligence, Huazhong University of Science and Technology (HUST), Wuhan, China, in 1995. From then to 1997, he served as a postdoctoral fellow in the Department of Automation, Tsinghua University, Beijing, China. Since 2003, he has been a full professor in the Department of Automation, Tsinghua University. His research interests include computer vision, pattern recognition, and image processing. In recent years, he has authored more than 100 papers in peer-reviewed journals and conferences. Among them, more than 30 papers have been published in top journals and conferences such as the IEEE Transactions on Pattern Analysis and Machine Intelligence, IEEE Transactions on Image Processing, and CVPR. He is an associate editor for the IEEE Transactions on Pattern Analysis and Machine Intelligence and two other journals. He received the National Outstanding Youth Foundation of China Award. He is a senior member of the IEEE.