Supplementary Material

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

I. Detailed Illustrations of the Proposed THU-READ

In this section, we provide more detailed descriptions of our proposed dataset [1]. We sample pair-wise representative frames from each type of action videos and present them in Fig. 1. Table I shows the list of 40 actions which appear in our dataset in detail. On one hand, our dataset is "all-about-hand", which is classified into two classes of "single-handed" and "double-handed". On the other hand, we divide our dataset into two categories: "hand-object" and "non-hand-object", according to whether "the hands interact with other objects". The name of each video has the following form:

RGB_**A**_**S**_**G**.avi or D_**A**_**S**_**G**.avi ,

where RGB and D denote the RGB modality and depth modality, **A**, **S** and **G** represent the action label, subject's name and the index of the group respectively. The standard train/test splits of cross-subject(CS) and cross-group(CG), which we adopted in this paper, have been released on the project page "http://ivg.au.tsinghua.edu.cn/dataset/THU_READ.php". In this way, others could utilize the consistent settings of the reported results in this paper.

TABLE I
A detailed list of all the actions that appear in our egocentric dataset. 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

II. Details of Implementing DSSCA and MMUDL on the THU-READ

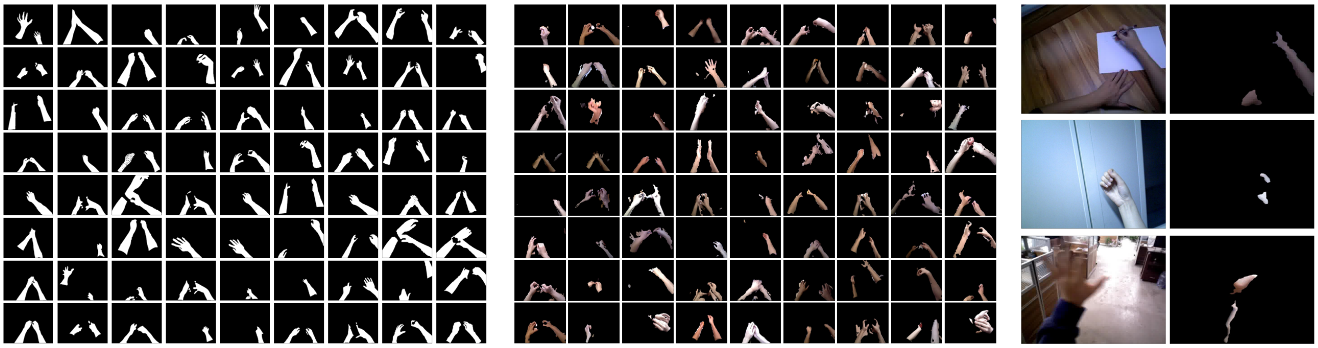
In section V.C.5 in our paper, we have conducted experiments to compare with state-of-the-art RGBD-based methods [2], [3] on our THU-READ dataset.

(1) For DSSCA [2], the authors extracted hand-crafted features around the major joints of human body, and they mainly focused on RGB features and depth features. However, the 3D coordinates of the hands are not available in our dataset, and we have extra features of the optical flows to be fused. For a fair comparison, we first extracted features on three modalities (see section IV.A.1 in the original paper for detail), and extended the multimodal learning scheme [2] for fusing three types of features. Specifically, we processed the optical flow features as the same with those on RGB and depth modalities, and modified the cost function (8) in [2] as follow:

$$\begin{aligned}\Omega^* \; = \;& \arg\min_{\Omega} \Delta(\mathbf{Y}_r,\mathbf{Y}_d)+\Delta(\mathbf{Y}_r,\mathbf{Y}_o)+\Delta(\mathbf{Y}_o,\mathbf{Y}_d)+\lambda\|\Omega\|_2\\ +\;& \zeta_r\Delta(\mathbf{X}_r,\widetilde{\mathbf{X}}_r)+\zeta_d\Delta(\mathbf{X}_d,\widetilde{\mathbf{X}}_d)+\zeta_o\Delta(\mathbf{X}_o,\widetilde{\mathbf{X}}_o)\\ +\;& \alpha_r\Psi(\mathbf{Y}_r;\rho_Y)+\alpha_d\Psi(\mathbf{Y}_d;\rho_Y)+\alpha_o\Psi(\mathbf{Y}_o;\rho_Y)\\ +\;& \beta_r\Psi(\mathbf{Z}_r;\rho_Z)+\beta_r\Psi(\mathbf{Z}_d;\rho_Z)+\beta_o\Psi(\mathbf{Z}_o;\rho_Z)\end{aligned} \tag{1}$$

Here, $\mathbf{X}_r, \mathbf{X}_d, \mathbf{X}_o$ denote the features on modalities of RGB, depth and optical flow respectively. $\mathbf{Y}$ and $\mathbf{Z}$ are shared and specific components. The meanings of other terms can be found in [2] for detail.

(2) In MMUDL [3], the author studied the problem of RGB-D person re-identification. Similar with the scheme to reimplement DSSCA in our dataset, we employed the fusion method in [3] based on extracted features on three modalities.



(a) Annotated ground truth (b) Successful examples (c) Failure examples

Fig. 2. **Visualizations of the ground truth annotation and segmentation results on THU-READ.** (a) shows some examples of the ground-truth hand annotation. (b) presents some selected successful results of the hand segmentation. (c) displays several failed examples due to the color similarity between background and hand skin (top), illumination (middle), and blur caused by the quick movement of hand (bottom).

TABLE II
EXPLORATION OF INTRODUCING MORE INTERMEDIATE LAYERS ON THE THU-READ DATASET (CS1).

	MDNN^2	MDNN^3	MDNN^4
Action Recognition Accuracy (%)	**56.25**	42.78	45.99

III. Visualizations of the Hands on the THU-READ

In this section, we present some visualization results of the ground truth annotation and segmentation results on our proposed dataset. Please see Figure 2 for details.

IV. Exploration on Introducing More Intermediate Layers

In our work, we designed two types of intermediate layers $\boldsymbol{f}(\boldsymbol{X})$ and $\boldsymbol{g}(\boldsymbol{X})$ to preserve the distinctive property for each modality and simultaneously explore their sharable information. While this method was shown to be effective, someone may cast doubt on it that the performance of the method can be further improved by introducing more linear mappings of the inputs followed by non-linearity (like $\boldsymbol{f}(\boldsymbol{X})$ and $\boldsymbol{g}(\boldsymbol{X})$). To address this issue, we conducted experiments on two variants of our method. They were formulated as MDNN^3 and MDNN^4 by adding more intermediate layers $\boldsymbol{l}_i(\boldsymbol{X}_i)$ (i = 1, 2, 3) as follow:

$$\text{MDNN}^3: \quad \boldsymbol{h}(\boldsymbol{X}) = \frac{1}{6}\sum_i [\boldsymbol{g}_i(\boldsymbol{X}_i) + \frac{1}{2}\boldsymbol{f}_i(\boldsymbol{X}_i) + \frac{1}{2}\boldsymbol{l}_i(\boldsymbol{X}_i)],$$

$$\text{MDNN}^4: \quad \boldsymbol{h}(\boldsymbol{X}) = \frac{1}{9}\sum_i [\boldsymbol{g}_i(\boldsymbol{X}_i) + \boldsymbol{f}_i(\boldsymbol{X}_i) + \boldsymbol{l}_i(\boldsymbol{X}_i)].$$

We set the dimensions of these intermediate layers $\boldsymbol{l}_i(\boldsymbol{X}_i)$ to be 512, which were the same as $\boldsymbol{f}_i(\boldsymbol{X}_i)$ and $\boldsymbol{g}_i(\boldsymbol{X}_i)$. We also kept the sum of allocated weights of $\boldsymbol{f}_i(\boldsymbol{X}_i)$, $\boldsymbol{g}_i(\boldsymbol{X}_i)$ and $\boldsymbol{l}_i(\boldsymbol{X}_i)$ to be 1 and all the other parameter settings remained unchanged.

Table II displays the comparison results on the split1 of cross-subject scenario on THU-READ dataset. We observe that introducing more linear mappings will make negative influence on the recognition performance. This is because the original $\boldsymbol{f}_i(\boldsymbol{X}_i)$ and $\boldsymbol{g}_i(\boldsymbol{X}_i)$ have their own physical meanings, i.e. the sharable components and distinctive components, and these components are regularized by our carefully designed objective function (see Page 6 eqution (5) in our original paper). However, the new added mappings (such as $\boldsymbol{l}_i(\boldsymbol{X}_i)$, i = 1, 2, 3) are lack of physical interpretation and regularizations, so they may bring some noise and cause the performance to decrease. Besides, adding more linear mappings will bring more computation cost for the entire MDNN.

V. Details of MDNN+TSN

In section V.F in our paper, we have conducted experiments by replacing the score fusion strategy in TSN model [4] with the multi-view learning method in our MDNN. In this section, we provide the details of MDNN+TSN. According to Eq(1) in [4], the TSN model for single modality was formulated as:

$$TSN(T_1, T_2, ..., T_N) = \mathcal{H}(\mathcal{G}(\mathcal{F}(T_1; \mathbf{W}), \mathcal{F}(T_2; \mathbf{W}), ..., \mathcal{F}(T_N; \mathbf{W}))). \quad (2)$$

Here we changed the K in [4] to N to avoid conflict with the K (the number of total modalities) defined in our paper. T_n, $\mathbf{W}$, $\mathcal{F}(T_n; \mathbf{W})$, $\mathcal{G}$ and $\mathcal{H}$ were corresponding to the sampled snippet, model parameter, class scores of the snippet, segmental consensus function and softmax function of the video respectively [4]. The final prediction was obtained by simply fusing the softmax scores $\mathcal{H}$ of each modality. In order to exploit the complementary information of different modalities, the new model MTN (*i.e.* MDNN+TSN) was formulated as:

$$MTN(T_1, T_2, ..., T_N) = \mathcal{H}(\mathcal{G}(\mathcal{F}'(T_1; \mathbf{W}), \mathcal{F}'(T_2; \mathbf{W}), ..., \mathcal{F}'(T_N; \mathbf{W}))), \quad (3)$$

$$\mathcal{F}'(T_n; \mathbf{W}) = S_{MDNN}(\mathbf{X}_n^{RGB}, \mathbf{X}_n^{Flow}, \mathbf{X}_n^{Depth}). \quad (4)$$

Here $\mathbf{X}_n^{RGB} = \mathbf{X}(T_n^{RGB}; \mathbf{W}^{RGB})$, $\mathbf{X}_n^{Flow} = \mathbf{X}(T_n^{Flow}; \mathbf{W}^{Flow})$ and $\mathbf{X}_n^{Depth} = \mathbf{X}(T_n^{Depth}; \mathbf{W}^{Depth})$. These three terms denoted the extracted features of fc layer from different modalities. After employing the multi-view learning method of our MDNN, we obtained the class scores $\mathcal{F}'(T_n; \mathbf{W})$ of the three modalities and fed them to the segmental consensus function $\mathcal{G}$ of the TSN model. The number of temporal segments N was set to be 3 in our experiments.

REFERENCES

[1] Tang, Y., Tian, Y., Lu, J., Feng, J., Zhou, J.: Action recognition in rgb-d egocentric videos. In: ICIP. (2017)
[2] 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
[3] 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
[4] 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