2023 IEEE International Conference on Energy Technologies for Future Grids
Wollongong, Australia, December 3-6, 2023

Hand Action Recognition from RGB-D Egocentric Videos in Substations Operations and Maintenance

Yiyang Yao[1,2], Xue Wang[1], Guoqing Zhou[1], Qing Wang[1]
[1] Northwestern Polytechnical University [2] State Grid Zhejiang Electric Power Co., Ltd.
yaoyiyang@mail.nwpu.edu.cn, {xwang, zhouguoqing, qwang}@nwpu.edu.cn

***Abstract*—This paper proposes a novel multimodal fusion network (MRDFNet) for egocentric hand action recognition from RGB-D videos. First, we utilize three separate streams to extract individual spatio-temporal features for different modalities, which include RGB frames, optical flow stacks, and depth frames. Particularly, for RGB and depth streams, an Attention-based Bidirectional Long Short-Term Memory network (Bi-LSTA) is used to identify regions of interest both spatially and temporally. Then, the extracted features are fed into a fusion module to obtain the integrated feature, which is finally used for egocentric hand action recognition. The fusion module is capable of learning complementary information from multiple modalities, i.e., preserving the distinctive property for each modality and meanwhile exploring the shareable property across modalities. Experimental results on both self-collected RGB-D Egocentric Manual Operation Dataset in Electrical Substations (REMOD-ES) and the THU-READ containing daily-life actions show the superiority of the proposed approach over state-of-the-art methods.**

Keywords—hand action recognition, human-object interaction, multimodal data, attention mechanism, egocentric video

I. INTRODUCTION

Electrical systems need regular maintenance to prevent system and equipment failures and ensure maximum safety and efficiency in the utilization of the facilities. The maintenance of electrical equipment is usually conducted by electricians and maintenance personnel who are knowledgeable in the maintenance process. They typically follow or use an electrical maintenance checklist. To ensure a safe work environment when working with or around electrical equipment, the implementation of electrical safety rules must always be followed by the personnel in their work area, such as use proper safety equipment, shut off power, inspect electrical equipment for damage, perform proper and reliable operations in order. Except the safety consciousness of the personnel, strengthening the management of power safety production can also ensure the stability and safety of the power system. With the development of intelligent surveillance technology, such as action recognition, timely identifying unsafe actions of the personnel and providing reminders or alerts to prompt correction could greatly contribute to the electrical systems.

Action recognition has traditionally been studied from a third-person view, for example, from a static or a handheld camera. Compared with the third-person view of security cameras which usually have a long capture range and suffer from occlusions, it is easier to capture the hand action from the first-person view when the wearers act and provide the close-ups with their visual attention. Hand action is an human activity that involves hand-object interaction, which can be described by an action-object pair. For example, when substation operators perform common operation and maintenance of equipment and protective switchgears, many manual operations are conducted, such as open an isolator, turn off a circuit breaker, measure the bus voltage, and clean dirt deposits on the bushings.

Moreover, complementary RGB-D modalities enrich raw data for action modeling, which has been proven to improve the accuracy for action recognition. However, due to the limited capture range of RGB-D cameras, multimodal action recognition methods cannot be applied to long-range surveillance. This shortcoming is greatly mitigated in first-person hand action recognition which is oriented mainly toward short-range applications, such as home healthcare, smart TV, robot imitation learning, and teleoperation. 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 exploring complementary information from multiple modalities for egocentric action recognition, 2) the inherent challenges in the egocentric scenario, such as noisy background and ego-motion of the camera, and 3) the scarcity of publicly available RGB-D egocentric hand action dataset.

In this work, we focus on the electrical substation scene, and propose a multi-stream deep learning-based method for RGB-D egocentric hand action recognition. The main contributions of our work are threefold. Firstly, we introduce the spatio-temporal attention mechanism for feature extraction to mitigate the effect of cluttered background and ego-motion of the camera. Secondly, we present a multimodal fusion sub-network to explore complementary information from multiple modalities for egocentric hand action recognition, by learning the distinctive property for each modality and the shared property across modalities. Finally, we collect a new dataset for RGB-D egocentric hand action recognition in the electrical substation scene. Extensive experimental results on the self-collected dataset and another benchmark for daily-life action recognition clearly show that our proposed method achieves superior performance compared with state-of-the-arts methods.

The remainder of this article is organized as follows. Section II introduces related studies. Our single modal feature extraction sub-network and multimodal fusion sub-network are introduced in Section III and Section IV respectively. Experiments on two first-person hand action recognition datasets are performed in Section V. Finally, the conclusion of this study is presented in Section VI.

This work was supported by the Science and Technology Project of the State Grid Corporation of China (5700-202019186A-0-0-00).

II. Related Works

A. Traditional first-person hand action recognition

Developing discriminative action representations with handcrafted features are one of the most important perspectives of studies in this field. Kitani et al. [1] computed optical flow images and extracted corresponding global action descriptors for representing first-person actions. To exploit the spatial distribution of features, Bambach et al. [2] proposed hand regions helped understanding first-person behavior, Pirsiavash et al. [3] used multiple HOG features to build corresponding spatial model, Fashi et al. [4] proposed intermediate layer action features and associated attention information.

B. Deep learning-based first-person hand action recognition

Deep learning has been overwhelmingly successful in multiple fields compared to traditional methods, such as computer vision, natural language processing, video/speech recognition. Ryoo et al. [5] proposed a new pooled feature representation with Convolutional Neural Network (CNN). Singh et al. [6] designed an end-to-end CNN to detect both head and hand motions and extract saliency regions. Ma et al. [7] proposed a gesture recognition method consisting of multiple sub-networks, including hand action segmentation sub-network, manipulated object localization sub-network, and hand action recognition sub-network. Zhang et al. [8] proposed a Long Short-Term Memory network (LSTM) for egocentric hand action recognition, which was capable of handling such "long-term dependencies" and solving the vanishing gradient problem. Considering the discriminative information in the input sequence can not be spatially localized, which is one of the shortcomings of LSTM, Sudhakaran et al. [9] further introduced spatio-temporal attention.

However, most of the above mentioned methods are proposed for single modal data. A few multimodal hand action recognition methods have been proposed. Multistream networks are the most representative approaches for fusing multimodal features. Yamazaki et al. [10] presented a framework for recognizing first-person hand-object interactions from a RGB-D sequence. Tekin et al. [11] proposed to predict frame-wise hand poses, object poses, object classes, and action classes. Li et al. [12] studies transfer learning from object recognition to hand action recognition. These methods basically adopt either early fusion or late fusion strategies. In early fusion, multimodal features are combined before they are fed into a classifier. In late fusion, the features of each modality are often associated with a classification score. All the scores are jointly used to predict an action class label.

C. Egocentric Hand Action Recognition in Substations Operations and Maintenance

Considering the accessing to large-scale labeled datasets, most existing recognition models are focused on recognizing egocentric actions in daily life, since publicly available first-person hand action datasets are mostly collected during daily living [13-15]. Little research on recognizing first-person hand action in a particular scene, such as an electrical substation, has been done[16-17]. The main reason is the scarcity of publicly available RGB-D egocentric hand action dataset in the electrical substation scenario.

III. Single Modal Feature Extraction Sub-Network

To extract spatio-temporal features for single modal input, we construct a sub-network consisting of a spatial module and a temporal module. Specifically, for the spatial module, on the basis of a skeleton CNN, such as a pretrained ResNet-34 [18] on ImageNet [19], we adopt the Class Activation Map (CAM) [20] mechanism to generate the attention map which is used to weight the output of the last convolutional layer (layer4) of ResNet-34. For the temporal module, we present a Bidirectional Long Short-Term Attention (Bi-LSTA) unit for a smooth and focused tracking of a latent representation of the video to generate the spatio-temporal features.

A. Spatial module

We present a schematic view of the spatial module in Fig. 1. For a given image $\mathbf{I}_t$ at time t, let $f_l^t(i)$ represent the activation value of convolutional unit l in the last convolutional layer at spatial location i, and ω_{lc}^t represent the weight for class c in unit l. Then the CAM for class c at time t is

$$\mathbf{M}_c^t(i) = \sum_l \omega_{lc}^t f_l^t(i), \tag{1}$$

which directly indicates the importance of activation at spatial location i leading to the classification of image $\mathbf{I}_t$ to class c. Hence the CAM can be regarded as a saliency map indicating different spatial weights for local regions leading to the classification. Normally the regions containing the wearer's hand and the manipulated object have higher weights comparing to other background regions.

We use a pretrained ResNet-34 as the base network, and add a spatial attention layer after the last convolutional layer (layer4). By multiplying the attention map calculated from the CAM using a softmax operation with the output of layer4 of ResNet-34, the attention-based spatial feature $f_{SA}^t(i)$ for $\mathbf{I}_t$ by the spatial module is,

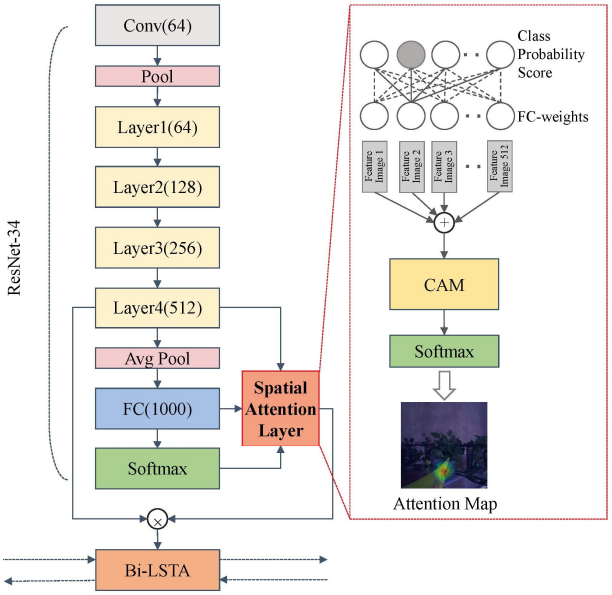

Fig. 1. Network architecture of the spatial module. The details of the spatial attention layer is showed in the right subfigure surrounded by red dashed lines. The Bi-LSTA unit is from the temporal module presented next.

$$f_{SA}^{t}(i) = f^{t}(i) \otimes \frac{e^{M_c^t(i)}}{\sum_i e^{M_c^t(i)}}, \quad (2)$$

where $f^t(i)$ is the output feature of layer4 at spatial location i, and $\otimes$ denotes the Hadamard product.

B. Temporal module

After extracting spatial features for single frame, we use a temporal module to further extract discriminative spatio-temporal patterns. Inspired by [9] and [18], we build a bidirectional LSTA architecture to focus on features from relevant spatial parts while attention being tracked smoothly across the sequence. The bidirectional strategy helps extract better spatio-temporal patterns from both directions.

The LSTM unit has two key components, the attention pooling operation that selects one out of a pool of specialized mappings to realize attention tracking and the output gating. The attention pooling ζ on spatial features $\mathbf{f}_{SA}^t$ returns a map v_a that is fed through a conventional RNN cell with memory $\mathbf{a}_t$ and output gate $\mathbf{s}_t$. Its output state $\mathbf{s}_t \odot \eta(\mathbf{a}_t)$ is added to the input v_a and softmax calibrated to obtain an attention map s. The map s is then applied to $\mathbf{f}_{SA}^t$, that is, $s \odot \mathbf{f}_{SA}^t$ is the attention filtered feature for updating memory state $\mathbf{c}_t$ using conventional LSTM recurrence. The output gate uses a filtered view of the updated memory state $v_c \odot \mathbf{c}_t$. To obtain v_c through pooling we use $s \odot \mathbf{f}_{SA}^t$ to control the bias of operator ζ, hereby coupling attention tracking with output gating. This model instantiated for hand action recognition from egocentric video in its convolutional version is,

$$v_a = \zeta(\mathbf{f}_{SA}^t, \omega_a), \quad (3)$$

$$(i_a, f_a, \mathbf{s}_t, a) = (\sigma, \sigma, \sigma, \eta)(W_a * [v_a, \mathbf{s}_{t-1} \odot \eta(\mathbf{a}_{t-1})]), \quad (4)$$

$$\mathbf{a}_t = f_a \odot \mathbf{a}_{t-1} + i_a \odot a, \quad (5)$$

$$s = \mathrm{softmax}\big(v_a + \mathbf{s}_t \odot \eta(\mathbf{a}_t)\big), \quad (6)$$

$$(i_c, f_c, c) = (\sigma, \sigma, \eta)(W_c * [s \odot \mathbf{f}_{SA}^t, \mathbf{o}_{t-1} \odot \eta(\mathbf{c}_{t-1})]), \quad (7)$$

$$\mathbf{c}_t = f_c \odot \mathbf{c}_{t-1} + i_c \odot c, \quad (8)$$

$$v_c = \zeta\big(\mathbf{c}_t, \omega_c + \omega_o \epsilon(s \odot \mathbf{f}_{SA}^t)\big), \quad (9)$$

$$\mathbf{o}_t = \sigma(W_o * [v_c \odot \mathbf{c}_t, \mathbf{o}_{t-1} \odot \eta(\mathbf{c}_{t-1})]). \quad (10)$$

Eqs. 3-6 implement our recurrent attention, Eqs. 9-10 are our coupled output gating. Bold symbols $\mathbf{a}_t$, $\mathbf{s}_t$, $\mathbf{c}_t$ and $\mathbf{o}_t$ represent recurrent variables, ω_a, ω_c, W_a, W_c and W_o are trainable parameters, σ and η are sigmoid and tanh activation functions, $*$ is convolution, $\odot$ is point-wise multiplication. The pooling ζ can be viewed as a trainable attention model for input features, and ϵ is the consistent reduction associated to ζ.

To enhance the effect of LSTA, we follow the bidirectional LSTM model [21] and build a Bi-LSTA architecture, which consists of a LSTA cell with two hidden cell states. Fig. 2 illustrates the network architecture of our proposed Bi-LSTA.

IV. Multimodal Fusion Sub-Network

In this section, we explain our multimodal fusion sub-network for egocentric hand action recognition incorporating the single modal feature extraction sub-network of Sec. III. To deal with multimodal inputs, inspired by [15], we build a three stream architecture with early fusion strategy to generate the final feature for egocentric hand action recognition: one stream for encoding spatio-temporal information from RGB frames, the second stream for encoding motion information from optical flow stacks, and the third stream for encoding spatio-temporal information from depth frames.

Let $\mathbf{X} = \{\mathbf{X}_i\}_{i=1}^K$ represent the features extracted from individual modals, where $\mathbf{X}_i$ denotes the features from modality i and K is the total number of modalities. In this work, we set K to 3. The extracted single modal features $\mathbf{X}$ is fed to the multimodal fusion sub-network to obtain the final feature $H(\mathbf{X})$ by exploring shared feature $g(\mathbf{X})$ and distinctive features $\{f_i(\mathbf{X}_i)\}_{i=1}^K$. The shared feature $g(\mathbf{X})$ can be calculated by

$$g(\mathbf{X}) = \frac{1}{K}\sum_{i=1}^{K} g_i(\mathbf{X}_i). \quad (11)$$

The relationship between $\mathbf{X}$ and $g_i(\mathbf{X}_i)$ is

$$g_i(\mathbf{X}_i) = F(\mathbf{W}_i^s \mathbf{X}_i + \mathbf{b}_i^s), i = 1,2,\ldots,K, \quad (12)$$

where F is a nonlinear function (convolutional block), $\mathbf{W}_i^s$ and $\mathbf{b}_i^s$ represent weight matrix and bias matrix respectively.

Similarly with $g_i(\mathbf{X}_i)$, the distinctive features $\{f_i(\mathbf{X}_i)\}_{i=1}^K$ are calculated following

$$f_i(\mathbf{X}_i) = F(\mathbf{W}_i^d \mathbf{X}_i + \mathbf{b}_i^d), i = 1,2,\ldots,K. \quad (13)$$

By assigning different weights to the objective function $H(\mathbf{X})$, the multimodal feature fusion can be formulated by

$$H(\mathbf{X}) = \sum_{i=1}^{K} \alpha_i f_i(\mathbf{X}_i) + \beta g(\mathbf{X}), \quad (14)$$

$$\sum_{i=1}^{K} \alpha_i + \beta = 1, \quad (15)$$

where α_i and β are hyperparameters corresponding to intermediate features. The fused feature $H(\mathbf{X})$ is then transformed to a fully connected layer, and a softmax function for hand action prediction.

Fig.3 shows a schematic diagram of the complete workflow of our proposed method. The network is named Multimodal RGB-D Flow Network, abbreviated as MRDFNet. As shown in Fig.3, three streams are built to extract features from individual modalities, which are then fed to the multimodal fusion network. Note that the streams for RGB sequences and depth sequences are based on the single modal feature extraction network of Sec. III, while the stream for optical flow stacks is based on ResNet-34 since only motion features are concerned.

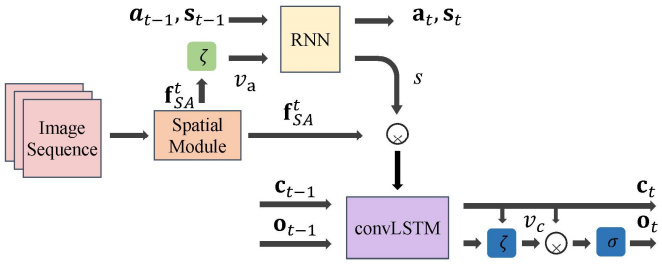

Fig. 2. Network architecture of the proposed Bi-LSTA.

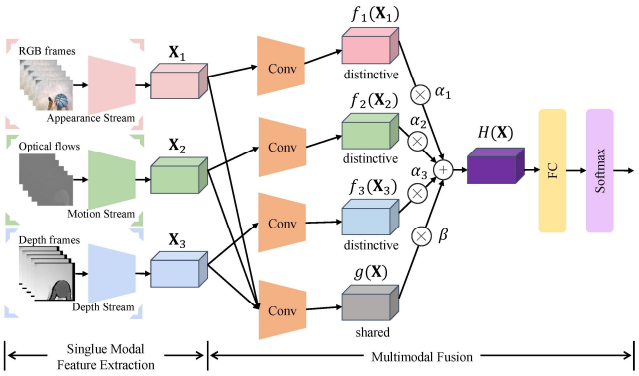

Fig. 3. Network architecture of the proposed MRDFNet. The convolutional blocks for distinctive features share the same structure (different parameters), while the convolution block for shared feature has a different structure since there exists an additional average operation over $g_i(\mathbf{X}_i)$.

V. Experiments and Results

A. Datasets

We evaluate the proposed method on the self-collected RGB-D Egocentric Manual Operation Dataset in Electrical Substations (REMOD-ES). The Kinect Azure is used for collecting registered RGB-D videos. For the purpose of acquiring egocentric action videos, we mount the sensor on a helmet and encourage the subject wearing the helmet (camera's orientation is roughly coincident with the subject's gaze) to perform manual operations with electrical devices as naturally as possible, which brings greater challenges of shifting backgrounds and various motion speeds to the task of hand action recognition. Fig. 4(a)(b) show the equipment and the condition during data collection. Finally, the dataset contains 12 hand action classes (move electric testing handcart, measure currency, measure voltage, push button, manipulate air breaker, manipulate rotary switch, manipulate earth switch, manipulate isolator, manipulate switch inside an operation cabinet, unlock, lock, lock pillar) and 124 video clips (2 modalities, 4 operators, 2 substations). Fig. 4(c) shows the visualization of partial egocentric hand actions with both RGB and depth modalities. Considering the limited size of the the dataset, we use data augmentation techniques to improve the sample diversity for network training, i.e., scale jittering and corner cropping. For cross validation, all the video clips are divided into 4 splits, of which 3 splits are sampled for training and the left one for testing. The average recognition accuracy is reported.

Moreover, a standard first person hand action recognition dataset, namely THU-READ [15], is also used to validate the generalization performance of the proposed method on the daily living scene. The dataset contains 40 hand action classes and 1920 video clips in total (8 subjects × 2 modalities × 40 classes × 3 repeated times). We follow the cross-subject (CS) setting as [15] to separate 8 subjects into 4 splits and use samples from 3 splits for training and the other for testing. The recognition accuracy on all splits and their average results are reported respectively.

B. Experimental settings

The training for single modal feature extraction network consists of two stages. In the first stage, the appearance stream

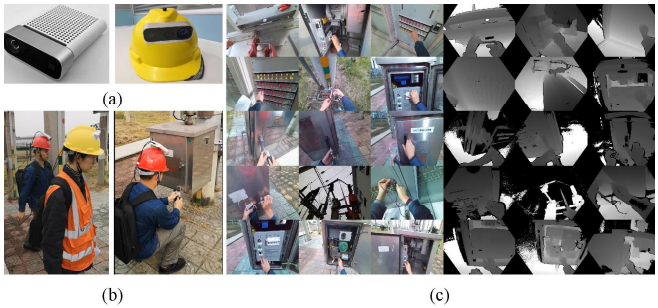

Fig. 4. Self-collected REMOD-ES. (a) The helmet mounted with a Kinect Azure sensor. (b) The condition during data collection. (c) Visualization of partial egocentric hand actions with both RGB and depth modalities.

was trained for 300 epochs with a learning rate of 10^{-3} and decayed by a factor of 0.1 after 25, 75 and 150 epochs, and the depth stream was trained for 200 epochs with a learning rate of 10^{-3} and decayed by a factor of 0.1 after 25, 75 and 150 epochs. In the second stage, the network was further trained for 150 epochs with a learning rate of 10^{-4} and decayed by a factor of 0.1 after 25 and 75 epochs. For the motion stream with ResNet-34, the network was trained for 750 epochs with an initial learning rate of 10^{-2} and decayed by a factor of 0.5 after 150, 300 and 500 epochs. We stacked five optical flow images, calculated using the TV-L1 algorithm [22], together as the input and obtain an averaged score. The multimodal fusion sub-network was trained for 50 epochs with an initial learning rate of 10^{-2} and the decay rate per step was set to 0.99.

C. Evaluation on THU-READ

1) *State-of-the-art comparison:* Our method was compared against the state-of-the-art methods on Tab. 1. We mainly compared our approach with several existing multimodal action recognition methods: ThreeStream [23], TSN [24], MDNN [15] and CAPF [25]. Following the two stream architecture, ThreeStream uses three separate CNN networks for single modal feature extraction and a fully connected layer for joint learning. TSN uses three individual TSN networks for single modal feature extraction and also a fully connected layer for joint learning. MDNN uses three individual TSN networks for single modal feature extraction and the multi-view learning for multimodal fusion, which corresponds to the MDNN+TSN model in the original work. CAPF performs a decoupling and recoupling learning as well as a cross-modal adaptive posterior fusion. From the table, we can see that the proposed method exceeds SOTA results for egocentric hand action recognition and achieves the best average accuracy (88.54%) in this protocol, which demonstrates the robustness of our method to noisy background and its strong spatio-temporal attention perception abilities.

2) *Multimodal fusion strategy analysis:* In order to verify the importance of the proposed multimodal fusion sub-network, we have conducted experiments by employing our proposed MRDFNet with four different fusion strategies. Direct fusion indicates that the output scores of each feature extraction stream are averaged and fused. Fully connected layer fusion indicates that the features of different modalities are output to the same fully connected layer. Both average fusion and weighted fusion employ the proposed multimodal fusion

approach of Sec. IV, where the former uses fixed hyperparameters ($\alpha_1 = \alpha_2 = \alpha_3 = 1/6$, $\beta = 1/2$). Tab. 1 also shows the performances of different fusion strategies. It can be seen that the proposed weighted fusion strategy outperforms all the other fusion strategies, which validates the effect of cross-modal fusion where distinctive and shared features are both explored.

Fig. 5 shows the confusion matrix obtained by the proposed MRDFNet with the weighted fusion strategy on THU-READ (split2). The horizontal axis in the confusion matrix indicates the predicted action category, and the vertical axis indicates the real action category, and a darker color for the matrix entry indicates a higher proportion. The points on the diagonal line indicate the proportion of actions which have been correctly recognized. It can be seen that most of the action categories are accurately identified by the proposed method.

3) *Parameter analysis:* There are several important parameters in this work. In Eq. 15, the parameters $\alpha_i (i = 1, 2, 3)$ and β are used to control the contributions of distinctive components and the shared information respectively. We analyzed how these parameters influenced model performance. Fig. 6 shows the performance on the CS setting by assigning different values to these parameters. We set $\alpha_1 = \alpha_2 = \alpha_3$ for simplicity. It can be seen that action recognition accuracy reaches a peak when the shared information and distinctive characteristics are simultaneously explored ($\alpha_1 = \alpha_2 = \alpha_3 = 0.23, \beta = 0.3$).

D. Evaluation on REMOD-ES

1) *State-of-the-art comparison:* Our approach is compared against the state-of-the-art methods on Tab. 2. We only compared our approach with TSN [24] and MDNN [15] on the REMOD-ES. From the table, we can see that the proposed method exceeds the SOTA results and achieves the best average accuracy (94.74%).

2) *Multimodal fusion strategy analysis:* Tab. 2 also shows the performances of four different fusion strategies. It can be seen that the proposed weighted fusion strategy outperforms all the other fusion strategies. Moreover, the proposed MRDFNet with different fusion strategies all exceeds the SOTA results.

TABLE I. COMPARISON OF THE HAND ACTION RECOGNITION ACCURACY (%) ON THU-READ

Methods	CS1	CS2	CS3	CS4	Average
ThreeStream [23]	80.98	80.65	81.67	79.84	80.79
TSN [24]	82.50	81.67	83.33	82.08	82.40
MDNN [15]	83.33	84.67	84.22	82.50	83.68
CAPF [25]	-	-	-	-	87.04
MRDFNet (Direct)	87.08	87.50	86.25	87.91	87.19
MRDFNet (FcNet)	85.00	87.08	82.91	85.42	85.10
MRDFNet (Average)	87.08	87.92	85.42	86.67	86.77
MRDFNet (Weighted)	**88.33**	**90.83**	**87.08**	**87.92**	**88.54**

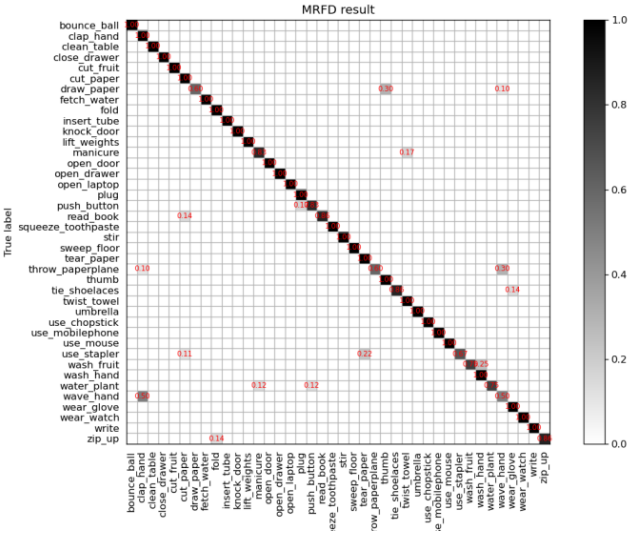

Fig. 5. Confusion matrix generated by the proposed MRDFNet with the weighted fusion strategy on THU-READ (split2).

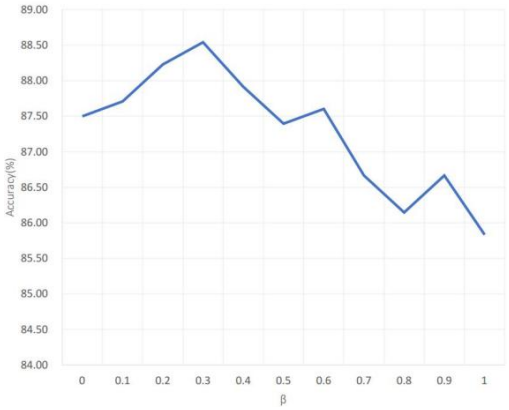

Fig. 6. Model performance influenced by different parameter values.

3) *Challenges for egocentric hand action recognition in electrical substation scenario:* We find that it is difficult to distinguish certain manual operations when they share similar appearance features and motion patterns. For example, as shown in Fig. 7, the two hand actions *manipulate_air_breaker* and *manipulate_rotary_switch* show subtle differences in terms of appearance (object and background) and motion. The mild motion further increases the difficulty for recognition.

VI. CONCLUSION

We propose a learning-based multimodal fusion network for egocentric hand action recognition. Firstly, spatial and temporal attention mechanisms are integrated for single modal feature extraction, which allows adaptively selecting important regions and key frames for action recognition. Secondly, a three-stream (appearance, motion and depth) architecture is built. Finally, a multimodal fusion sub-network is introduced to explore both distinctive features of each modality and modality-shared features. Experimental results on different scenarios demonstrate the proposed method can effectively encourage the network to extract and fuse discriminative multimodal spatio-temporal features. In the future, we will

extend our method to untrimmed video clips for both temporal localization and action recognition.

TABLE II. COMPARISON OF THE HAND ACTION RECOGNITION ACCURACY (%) ON REMOD-ES

Methods	Average
TSN [24]	73.68
MDNN [15]	78.95
MRDFNet (Direct)	84.21
MRDFNet (FcNet)	78.95
MRDFNet (Average)	84.21
MRDFNet (Weighted)	**94.74**

Fig. 7. Challenges of different actions with similar appearance and motion patterns. Left: *manipulate_air_breaker*. Right: *manipulate_rotary_switch*.

ACKNOWLEDGMENT

The authors would like to thank all the participants for dataset collection, especially Weilun Pang for data preparation.

REFERENCES

[1] K. M. Kitani, T. Okabe, Y. Sato, and A. Sugimoto, "Fast unsupervised ego-action learning for first-person sports videos," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), pp. 3241-3248, 2011.

[2] S. Bambach, S. Lee, D. J. Crandall, and C. Yu, "Lending a hand: Detecting hands and recognizing activities in complex egocentric interactions," in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), pp. 1949-1957, 2015.

[3] H. Pirsiavash and D. Ramanan, "Detecting activities of daily living in first-person camera views," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), pp. 2847-2854, 2012.

[4] A. Fathi, Y. Li, and J. M. Rehg, "Learning to recognize daily actions using gaze," in Proc. Euro. Conf. Comput. Vis. (ECCV), pp. 314-327, 2012.

[5] M. S. Ryoo, B. Rothrock, and L. Matthies, "Pooled motion features for first-person videos," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), pp. 896-904, 2015.

[6] S. Singh, C. Arora, and C. V. Jawahar, "First person action recognition using deep learned descriptors," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), pp. 2620-2628, 2016.

[7] M. Ma, H. Fan, and K. M. Kitani, "Going deeper into first-person activity recognition," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), pp. 1894-1903, 2016.

[8] Y. Zhang, S. Sun, L. Lei, H. Liu, and H. Xie, "STAC: Spatial-temporal attention on compensation information for activity recognition in fpv," Sensor, vol. 21, no. 4, pp. 1106, 2021.

[9] S. Sudhakaran, S. Escalera and O. Lanz, "LSTA: Long short-term attention for egocentric action recognition," in Proc. IEEE/CVF Conf. Comput. Visi. Pattern Recognit. (CVPR), pp. 9946-9955, 2019.

[10] W. Yamazaki, M. Ding, J. Takamatsu, and T. Ogasawara, "Hand pose estimation and motion recognition using egocentric RGB-D video," in Proc. IEEE Int. Conf. Robot. Biomimetics (ROBIO), pp. 147-152, 2017.

[11] B. Tekin, F. Bogo, and M. Pollefeys, "H+O: Unified egocentric recognition of 3D hand-object poses and interactions," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), pp. 4506-4515, 2019.

[12] R. Li, H. Wang, Z. Liu, N. Cheng, and H. Xie, "First-person hand action recognition using multimodal data", IEEE Trans. Cog. Dev. Syst., vol. 14, no. 4, pp. 1449-1464, 2022.

[13] A. Fathi, X, Ren, and J. M. Rehg, "Learning to recognize objects in egocentric activities", in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), pp. 3281-3288, 2011.

[14] G. Garcia-Hernando, S. Yuan, S. Baek, and T.-K. Kim, "First-person hand action benchmark with RGB-D videos and 3D hand pose annotations," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), pp. 409-419, 2018.

[15] Y. Tang, Z. Wang, J. Lu, J. Feng, and J. Zhou, "Multi-stream deep neural networks for RGB-D egocentric action recognition," IEEE Trans. Circuits Syst. Video Technol., vol. 29, no. 10, pp. 3001-3015, 2019.

[16] R. Yao, S. Jin, Y. Fei and L. Fang, "Behavior recognition of substation maintenance personnel based on deep learning," in Proc. IEEE Conf. Tele., Opt. Comput. Sci. (TOCS), pp. 218-221, 2022.

[17] C. Zhu, G. Liu, L. Liu, F. Gao, and Q. Gao, "Kinect-based substation operation and maintenance action recognition method," in Proc. Int. Conf. Artif. Intell. Inf. Syst., pp. 7001-7007, 2021.

[18] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.(CVPR), pp: 770-778, 2016.

[19] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, et al., "ImageNet large scale visual recognition challenge," Int. J. Comput. Vis. (IJCV), vol. 115, no. 12, pp: 211-252, 2015.

[20] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), pp: 2921-2929, 2016.

[21] A. Hanson, K. PNVR, S. Krishnagopal, and L. Davis, "Bidirectional convolutional lstm for the detection of violence in videos," in Proc. Euro. Conf. Comput. Vis. (ECCV) Workshops, pp: 280-295, 2018.

[22] C. Zach, T. Pock, and H. Bischof, "A duality based approach for realtime tv-l1 optical flow," in Proc. Germ. Association Pattern Recognit. (DAGM), pp: 214-223, 2007.

[23] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," in Proc. Adv. Neural Inf. Proces. Syst., vol. 1, pp: 568-576, 2014.

[24] L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang, et al., "Temporal segment networks: towards good practices for deep action recognition," in Proc. Euro. Conf. Comput. Vis. (ECCV), pp: 20-36, 2016.

[25] B. Zhou, P. Wang, J. Wan, Y. Liang, F. Wang, D. Zhang, et al., "Decoupling and recoupling spatiotemporal representation for rgb-d-based motion recognition," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), pp. 20122-20131, 2022.