MBR-MDA: Multi-person Behavior Recognition Method Based on Multi Descriptors Aggregations

Yang Luo, Rongheng Lin
State Key Laboratory of Networking and Switching Technology
School of Compute Science(National Pilot Software Engineering School)
Beijing University of Posts and Telecommunications
Beijing, China
{luoyang, rhlin}@bupt.edu.cn

***Abstract*—Multi-person behavior recognition is an important task in intelligent video surveillance. In this paper, we propose a multi-person behavior recognition method based on multi descriptors aggregations (MBR-MDA) for real-time surveillance scenarios. Our method employs multi-object tracking to obtain consecutive frames of each person, and uses a 2D convolutional network with temporal shift module (TSM) for behavior recognition. To address the limitation of 2D convolutional network in capturing global temporal features, we introduce a plug-and-play module called MDA that can be integrated into the 2D convolutional network. By applying data augmentation and embedding the MDA3D module, our method achieves a 4.8% improvement over TSM baseline on the HMDB51 dataset, with only a minimal speed loss of 0.3ms. We evaluate our method on several public datasets and demonstrate that embedding MDA into other methods can also enhance their performance.**

***Index Terms*—intelligent monitoring, behavior recognition, multi descriptors aggregations.**

I. Introduction

Traditional video surveillance methods need regular video supervision, which requires a large workforce and resources. In recent years, intelligent surveillance systems have also flourished with the development of deep learning. Intelligent surveillance systems perform object and behavior recognition using computer vision technology. When abnormal conditions are detected, the system can alert the personnel to handle it.

Despite advancements in deep learning, intelligent surveillance systems still face numerous challenges, including occlusion, illumination variation, pose variation, and background clutter. Multi-person behavior recognition tasks can be particularly challenging, as they involve multiple persons interacting with each other or objects in a scene, making it difficult to accurately and efficiently capture the temporal and spatial features of their behaviors.

In general, there are 3 methods for multi-person behavior recognition. The first method is based on object detection, which identifies the location and behavioral categories of people using methods such as YOLO [1]. This approach has a fast recognition speed, but the lack of temporal information in a single frame. The second method obtains the person location through object detection and extracts key points using a human skeleton model such as STGCN [2]. This method is slow in acquiring human key points, and is very sensitive to the problem of occlusion. The third method obtains each person continuous images by multi-object tracking, and then recognize human behavior based on frame sequences, which is balanced in speed and accuracy, and can also handle the occlusion situation. Therefore, we choose the third method as the overall technical route.

To address the issue of delayed response time in multi-person behavior recognition systems, this paper proposes using multi-object tracking and behavior recognition based on 2D convolution. To enhance the performance of 2D convolution in behavior recognition, this paper proposes MDA which produces a fusion of features from multi-pooling methods to achieve a better feature representation for image sequence. For time-series feature fusion, a single pooling method compromises the accuracy of information description. This paper employs GemPooling——a generalizable pooling method that aggregates various pooling methods to obtain a more precise description. Based on the MDA method, we propose Multi Descriptors Aggregation by 2D Pooling(MDA2D) and Multi Descriptors Aggregation by 3D Pooling(MDA3D) as methods for enhancing feature representation in this study.

The main contributions of our paper are as follows:

1. The proposed MBR-MDA method employs multi-object tracking and behavior recognition based on 2D convolution to effectively capture temporal and spatial features of multi-person behavior, making it highly suitable for multi-person behavior recognition..

2. The proposed MDA method achieves this by effectively fusing features from multiple pooling methods to provide a better feature representation for image sequences. This not only enhances accuracy but also maintains calculation speed, making it highly effective for real-time behavior recognition applications.

II. Related Work

A. Object Detection

To recognize behaviors in surveillance scenes, it is necessary to locate people through object detection. Object detection are mainly categorized as two-stage method such as Faster-RCNN [3], and one-stage method such as YOLO [1].

Research in this paper is supported by the Key Research and Development Program of Jiangxi Province(20212BBE51002)
DOI reference number: 10.18293/SEKE23-169

One-stage method as YOLO algorithm has been pursuing more optimized speed and relatively practical accuracy in development. Yolov4 [4] proposed the structure of FPN+PAN to achieve a better fusion of the low-level and high-level features. Yolov5 proposed focus operation at the image input and adaptive padding method which achieved better latency and accuracy.

B. Multi-object Tracking

Multi-object tracking associates the trajectories of multiple objects in a video sequence and assigns a unique ID, thus enabling the acquisition of consecutive multi-frame images of a single object. Multi-object tracking first requires the object location through object detection and then matches objects and trajectories based on kinematic features or appearance features.

Bewley et al. proposed the SORT [5] to achieve simple real-time trajectory tracking using the Kalman filter algorithm and the Hungarian algorithm. Wojke et al. introduced the ReID network for extracting the appearance features of objects based on the SORT algorithm. They proposed DeepSort [6] to achieve more accurate tracking through appearance features. Wang et al. proposed the JRE [7] by embedding ReID module into the Yolov3 [8] to achieve multi-object tracking by using only one network. Zhang et al. proposed anchor-free method to realize fair multi-object tracking FairMOT [9], which has better accuracy.

C. Human Behavior Recognition

After obtaining continuous video sequences of each target by object tracking, the video sequences are fed into a behavior recognition network. The following three network structures based on RNN, 3D convolution, and 2D convolution are mainly used for behavior recognition.

Based on RNN for temporal modeling of features, the temporal information is modeled by the recurrent neural network. Shikhar Sharma et al. proposed AttentionLSTM [10] to process continuous video frames with LSTM units and also introduced Attention mechanism in the model to pay more attention to the regions of action changes to improve the accuracy. However the RNN is slow and unsuitable for recognition tasks in surveillance videos.

Based on 3D convolutional network to extract the features of video sequences, the temporal information is modeled by 3D convolution operation. In 2015, Tran D et al. proposed a general network C3D [11] for video recognition, which takes the conventional 2D convolution and pooling operations and puts them into 3D space to effectively obtain the temporal information. Christoph et al. proposed SlowFast [12] which adopted a two-branch network to obtain behavioral features. Slow branch used to model spatial information, and fast branch used to model temporal information.

Based on 2D convolutional networks to extract the overall features of video sequences, such as Wang et al.proposed TSN [13] in 2016, and Zhou et al. proposed TSM [14] based on 2D time series displacement in 2018. TSM captures video motion information by completing the information exchange between frames through temporal displacement.

In contrast to 3D convolution, 2D convolution is weak in modeling temporal information, but it has a fast inference speed. This paper aims to improve the accuracy of 2D convolution while preserving the advantages of 2D convolution calculation.

III. Our Method

A. Behavior Recognition Algorithm Framework

The multi-person behavior recognition algorithm framework based on multi-object tracking and behavior recognition is shown in Fig.1, which mainly includes three parts: object detection, multi-object tracking, and behavior recognition.

Fig. 1. Multi-person Behavior Recognition Method Based on Multi Descriptors Aggregations.

Firstly, YOLOv5s is used to obtain the person's location. The features of different scales are used to calculate the bounding box and category of the object. Finally, the NMS(Non-Maximum Suppression) method filters and gets the object detection results.

The object detection results are fed into the DeepSort multi-object tracking algorithm, which uses Kalman filtering to provide the optimal estimate of the object's location in the next frame. The ReID network extracts features of the objects, and the Hungarian algorithm is then used to assign object IDs. This allows us to obtain a continuous video sequence of a single object.

Since the size of the bounding box detected by the target changes, the size of the video frame obtained by target tracking is not fixed. To achieve a more stable effect, we use the fixed center and surrounding padding method to align the size. This allows us to obtain a stable video input into the behavior recognition network.

Finally, the continuous frames are input to the behavior recognition network for recognition, and the result is outputted. Fig. 1 shows the behavior recognition effect in an equipment hall, which can recognize the action states of people standing, looking at each other, walking, lying down, holding objects, and more.

We conducted latency statistics for the multi-person recognition framework using 24 frames of 1920×1080 surveillance video input. The three steps of object detection, multi-object tracking, and behavior recognition took 637ms, 835ms, and 1082ms, respectively. Behavior recognition was the most time-consuming step, and recognition time significantly increased when more people appeared in the scene. To improve accuracy and reduce behavior recognition time, we propose an improved 2D convolution-based behavior recognition method in the next section.

B. Multi Descriptors Aggregation——MDA

Compared to 3D CNNs, 2D CNNs have the advantages of fewer parameters and faster speed. However, due to information loss in the ability of 2D convolution to acquire temporal features, it is less accurate than 3D convolutional networks for this task. Therefore, we designed a behavior recognition network based on 2D convolution and proposed a multi-pooling feature description aggregation operation based on TSM (Temporal Shift Module). MDA compensates for the loss of 2D convolution in acquiring temporal information, resulting in improved accuracy for behavior recognition.

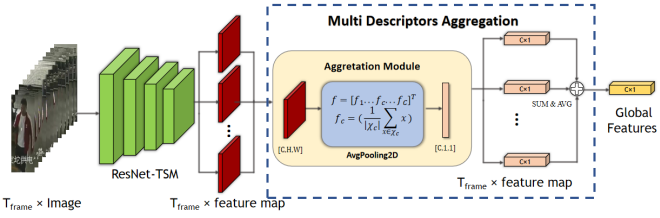

Fig. 2. Video feature extraction framework TSM-ResNet50.

*1) **Network Structure**:* Fig.2 shows the behavior recognition method based on the TSM(Temporal Shift Module). Firstly, adding the temporal shift module to the backbone ResNet50 [15], obtaining the time series' interactive information, and enhancing the expression of 2D convolution to obtain the T-frame level feature map. In the case of ResNet50 [15] each feature map is usually 2048×7×7 size, and 8 frames are extracted from each video sequence. After obtaining the frame-level feature maps, the feature maps need to be aggregated by dimensionality reduction. Then through the aggregation module, the 1-dim features of each frame will be obtained by the aggregation module, usually directly in the form of AvgPooling2D, then summing and averaging the N frames of 1-dim features, and finally obtaining the 1-dim global feature representation at the video level.

*2) **Multi Descriptors Aggregation by 2D Pooling——MDA2D**:* Usually, in order to enhance the global feature representation of video sequences, different weights can be given in the spatial dimension and channel dimension by adding attention modules, such as the NonLocal [16] method. However, NonLocal method will introduce more parameter computation and increase the GPU occupation during model training and inference. We introduce the Multi Descriptors Aggregation by 2D Pooling (MDA2D) to replace the original AvgPooling module in the video feature extraction framework in Fig.2. It can obtain better feature representation for each frame in the video sequence, enable end-to-end training, and improve video sequence behavior recognition.

The pooling operation in images can be defined as follows, given image I, the output of the backbone network is a tensor C × H × W. C is the number of channels and H × W is the resolution of the feature map. The pooling operation can be defined as a function of the calculation of the feature surface X_c for a single channel H × W of the feature map. After

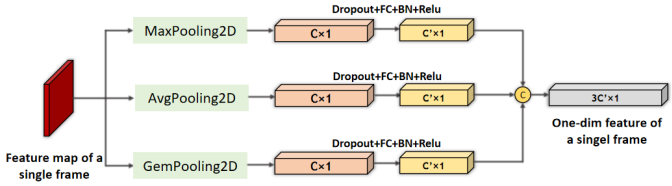

Fig. 3. Multi Descriptors Aggregation by 2D Pooling Framework.

pooling, the features change to a one-dimensional tensor f of dimension C. The definition of the pooling operation can be generalized and defined as the following equation.

$$f = [f_1 ... f_c ... f_C]^T, f_c = (\frac{1}{|X_c|} \sum_{x \in X_c} x^{p_c})^{\frac{1}{p_c}} \tag{1}$$

From the above Eq.1, the average pooling layer $p_c = 1$, while the maximum pooling layer $p_c \to \infty$. For the Gem-Pooling layer with generalization, pc takes the value of (1, ∞), expressing the intermediate state between AvgPooling and MaxPooling. Depending on the dataset, different pooling approaches may produce inconsistent differences in effects, e.g., average pooling can obtain information about larger regions in the image. In contrast, Max pooling can focus more on information about focused regions in the image.

As shown in Fig.3, three branches of pooling are adopted for global feature extraction of video sequences, and the output of the features from the pooling layer are downsampled by Dropout, fully connected layer, Batch Normalization, and ReLU in turn. Finally, the three branches' features are fused by stacking, and the final output is a 1-dim of single-frame feature expression. Assume that each pooling branch outputs features as $v_{(b_i)}$, b_i proxy for a certain pooling branch, which is selected from MaxPooling, AvgPooling, and GemPooling. And the input frame-level features are v_{ori_2d}, with dimensions (C, H, W), has the following expression relation defined by Eq.2.

$$\begin{aligned} v^{b_i} &= Relu(BN(D(W^{b_i}) \cdot v_{ori_2d})), \\ b_i &\in \{Max2D, Avg2D, Gem2D\} \end{aligned} \tag{2}$$

W^{b_i} is the weight parameter of the fully connected layer. $D(\cdot)$ donate dropout is a common way to mitigate network overfitting in networks, dropping neuron parameters with a certain probability. Finally, the nonlinear expression of the features is boosted by ReLU. As shown in Eq.3 and 4 below, the final 1-dim feature expression of the t frame is obtained by the aggregation operation of each pooled branch feature V^{S_t}, and further obtains the global feature expression by summation averaging V_g where $Concat$ denotes the feature stacking operation in the channel dimension.

$$V^{S_t} = Concat(\nu^{b_1}, \nu^{b_2}, \phi^{(b_3)}) \tag{3}$$

$$V_g = \frac{1}{T} \sum_{t}^{T} V^{S_t} \tag{4}$$

*3) **Multi Descriptors Aggregation by 3D Pooling——MDA3D**:* From the single-frame feature aggregation descriptor by 2D Pooling above, we can obtain an aggregated feature for each video sequence frame. And then obtain the global features of the video sequence by final sum & avg operation, which is still not elegant and intuitive enough. The features between T-frames only interact with each other once by final sum & Avg in Eq.4. Therefore, we propose Multi Descriptors Aggregation by 3D Pooling (MDA3D), which is an improvement on MDA2D. The global feature representation of video sequences is further improved by enhancing the feature aggregation capability of the aggregation module through various 3D pooling methods.

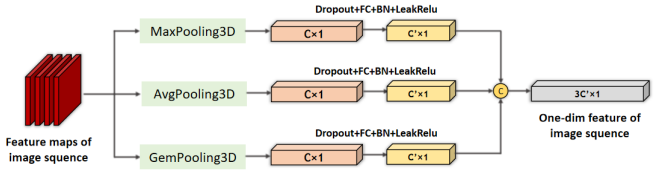

Fig. 4. Multi Descriptors Aggregation by 3D Pooling framework.

As shown in Fig.4, the MDA3D aggregate feature for the input N-frame feature map $vori_3d$ with dimensions (T, C, H, W). Unlike the MDA2D, MDA3D uses 3DPooling to pool the 3D feature map directly, and LeakyReLU is used instead of ReLU. Based on Eq.2, MDA3D method can be obtained by modifying the input, pooling methods, and the activation function as shown in Eq.5 below.

$$\begin{aligned} v^{b_i} &= Relu(BN(D(W^{b_i}) \cdot v_{ori_3d})), \\ b_i &\in \{Max3D, Avg3D, Gem3D\} \end{aligned} \tag{5}$$

MDA3D's input dimension is different from MDA2D. Through the 3D pooling operation, which contains four times of information fusion between T-frames while the 2D Pooling contains only once, the temporal information can be aggregated more effectively through the 3D pooling operation. Since the 3D Pooling already aggregates the time dimension, there is no need to sum and average the time dimension by Eq.4. The 1-dim global representation of the video sequence can be obtained by directly performing the stacking operation of Eq.6.

$$V^{S_t} = Concat(\nu^{b_1}, \nu^{b_2}, \phi^{(b_3)}) \tag{6}$$

IV. Experiments

A. DataSets

We evaluate improved behavior recognition algorithm on benchmark datasets——HMDB51 [17] and UCF101 [18]. Finally, the algorithm will be applied to a equipment hall monitoring scenario.

The HMDB51 dataset is a video sequence classification data released by Brown University in 2011. HMDB51 contains 6849 segments of samples, divided into 51 categories, and each category contains at least 101 segments of samples. The number of the training set in the experiment is 3570 video sequences, and the number of validation set is 1530 video sequences.

UCF101 is a medium-sized action video dataset collected from YouTube by the University of Central Florida Computer Vision Research Center. UCF101 contains 13,320 video clips with 101 subdivided behavioral action categories. Each of these categories is divided into 25 subsets, containing 4 to 7 groups of actions. The number of the training set is 9537 video sequences, and the number of the validation set is 3783 video sequences.

B. Training Details

The behavior recognition network is implemented using the PyTorch 1.8.0. Training and testing on a Linux server with 2 NVIDIA RTX 3090 GPUs.

The video sequences of the training set are averaged into 8 segments, and 1 frame is extracted from each segment to produce 8 frames of video sequences as the network input. Meanwhile, MultiScaleCrop, RandmoFlip, and MixUp are used for data enhancement.

To achieve a better performance in real-world scenarios, we first pre-train the TSM-Resnet50 model on the Kinetics400 dataset, and then load the backbone parameters to finetune on HMDB51 and UCF101 datasets.

For training, batchsize is set to 32 for 30 epochs. The SGD optimizer is used for gradient descent, and the learning rate is updated in a momentum manner. The initial learning rate is set to 0.005 and decremented using the cosine approach. Cross entropy is used as the loss function in classification, and label smoothing is also used to regularize the classifier.

For prediction, we also tested the inference speed of the network by setting batchsize to 1. The average prediction time of model inference was obtained by 300 times inferences after 10 times GPU pre-inferences.

C. Main Results

*1) **MDA2D Results**:* Firstly, the original TSM method was used as the experimental baseline to obtain 0.7051 and 0.9477 Top1-ACC on HMDB51 and UCF101 datasets. After that, the Top1-ACC accuracy of model was improved by adding tricks such as data enhancement, label smoothing, etc., which improved 2.7% and 1.03% on HMDB51 and UCF101, respectively. That indicated these tricks on image classification can also effectively improve video classification model results.

TABLE I
MDA2D ablation experiments on HMDB51.

Models	Params	FLOPs	Latency	Top1-ACC
TSM(baseline)	23.6M	32.9G	8.7ms	0.7051
TSM+Tricks				0.7322
TSM+Tricks+NL	31.0M	49.4G	14.8ms	0.7370
TSM+Tricks+MDA2D	**26.7M**	**32.9G**	**8.6ms**	**0.7448**

We use the NonLocal attention mechanism module as an experimental control for MDA2D and MDA3D. NonLocal

TABLE II
MDA2D ABLATION EXPERIMENTS ON UCF101.

Models	Params	FLOPs	Latency	Top1-ACC
TSM(baseline)	23.7M	32.9G	8.7ms	0.9477
TSM+Tricks				0.9580
TSM+Tricks+NL	31.1M	49.4G	14.8ms	0.9602
TSM+Tricks+MDA2D	**26.8M**	**32.9G**	**8.6ms**	**0.9601**

enhances the feature representation of the TSM model by inlining the attention weights, which can effectively impose spatiotemporal distance constraints in video sequences. Adding the NonLocal module further improves 0.48% and 0.22% on HMDB51 and UCF101, respectively.

The MDA2D proposed in this paper further improves 1.26% on the HMDB51 dataset and 0.21% on the UCF101 dataset. The improvement on HMDB51 exceeds that of NonLocal, and the improvement on UCF101 is almost the same. Meanwhile, MDA2D has fewer parameters and Flops, which also brings faster inference speed. The inference speed by embedding the MDA2D module is basically the same as the baseline but significantly improves accuracy.

*2) **MDA3D Results**:* The improved MDA3D module shows in TableIII and TableIV below. This module can effectively fuse the global features from multiple pooling feature perspectives and achieves the best Top1-ACC on both HMDB51 and UCF101 datasets, further improving by 2.11% and 0.45%, respectively, exceeding the fusion results with NonLocal. The accuracy of MDA3D is higher than that of MDA2D, which also verifies that the 3D pooling aggregation structure has better global sequence characterization ability.

TABLE III
MDA3D ABLATION EXPERIMENTS ON HMDB51.

Models	Params	FLOPs	Latency	ACC@1
TSM(baseline)	23.6M	32.9G	8.7ms	0.7051
TSM+Tricks				0.7322
TSM+Tricks+NL	30.9M	49.4G	14.8ms	0.7370
TSM+Tricks+MDA3D	**26.7M**	**32.9G**	**9.0ms**	**0.7533**
TSM+Tricks+MDA3D+NL	34.1M	49.4G	15.1ms	0.7389

TABLE IV
MDA3D ABLATION EXPERIMENTS ON UCF101.

Models	Params	FLOPs	Latency	ACC@1
TSM(baseline)	23.7M	32.9G	8.7ms	0.9477
TSM+Tricks				0.9580
TSM+Tricks+NL	40.0M	49.4G	14.8ms	0.9602
TSM+Tricks+MDA3D	**26.7M**	**32.9G**	**9.0ms**	**0.9625**
TSM+Tricks+MDA3D+NL	34.2M	49.4G	15.1ms	0.9588

Regarding model complexity, the MDA3D method keeps the same as MDA2D, and only the inference time increases by 0.4ms, maintaining a better real-time performance. Also, MDA3D outperforms the NonLocal method in terms of parameters and inference time in the case of obtaining Top1-ACC.

When embedding both the NonLocal and MDA3D modules in the model simultaneously, the accuracy is lower than when embedding the MDA3D module alone. The network produces some degradation in performance metrics. One possible reason for this is that NonLocal is embedded in multiple stages in the backbone network, resulting in more attentional computations. The goal of fusing information between different time series in the backbone network through the attentional mechanism is consistent with the fusion aspect of different global pooling information, but adding too many parameters can lead to overfitting the model quickly.

*3) **Generality of MDA**:* We integrated the MDA3D module into SlowFast, a behavior recognition network based on 3D convolution. The network structure of SlowFast was set as follows: the number of frames in the Slow path was 2, the frame extraction step was 32, the Fast path sampling was 8×16 frames, and the number of Fast path channels was 1/8 of that in the Slow path. As shown in Table V below, tests of SlowFast embedded in MDA3D on the UCF101 dataset still showed a 0.6% improvement. While the improvement was not as obvious as in the 2D convolutional network, it demonstrates the generality and effectiveness of MDA.

TABLE V
MDA3D EMBEDDING IN TSM & SLOWFAST.

Models	Params	FLOPs	Latency	Top1-ACC
TSM(baseline)	23.7M	32.9G	8.7ms	0.9477
TSM+Tricks+MDA3D	**26.7M**	**32.9G**	**9.0ms**	**0.9625 (+0.0148)**
SlowFast(baseline)	33.8M	74.5G	23.4ms	0.9235
SlowFast+MDA3D	**36.8M**	**74.5G**	**23.9ms**	**0.9295 (+0.0030)**

*4) **Feature Space**:* Using the t-SNE dimension reduction method in the sklearn package, we obtained a feature visualization map of the model. The MDA3D method generates 1536-dimensional features, while the other comparison algorithm models generate 2048-dimensional features. As shown in Fig. 5, we can see that the features generated by the MDA3D method have a better feature representation effect than the benchmark model. The inter-class distance is larger, and the intra-class distance is smaller. Moreover, for some abnormal points like the characteristic effect of running action category 32, the MDA3D method outperforms other methods without mixing categories 42 and 8.

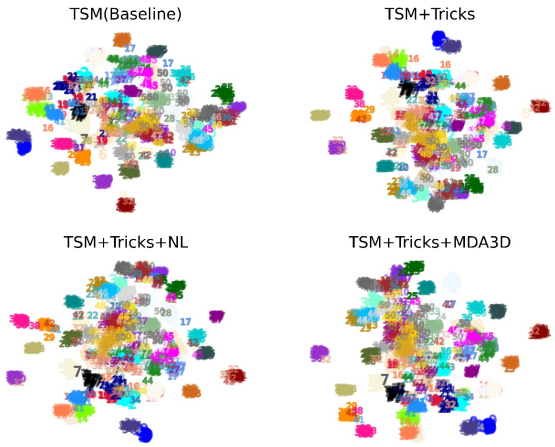

Fig. 5. t-SNE Visual rendering of feature space.

To demonstrate that the features generated by the MDA3D model have a better representation, we further classified and identified the features using various machine learning classifiers. TableVI shows that all classifiers, including MLP, SVM, KNN, LR, and RF random forest classifiers. On the hmdb51 dataset, the MDA3D feature classification results achieved the best Top1 ACC, with an effective improvement of more than 3% compared to the benchmark. This proves that MDA3D can extract better semantic features.

TABLE VI
MDA EXPERIMENTS ON OUR DATASET.

	TSM+Tricks	**TSM+Tricks+NL**	**TSM+Tricks+MDA3D**
MLP	0.729	0.733	**0.745**
SVM	0.733	0.716	**0.733**
KNN	0.716	0.733	**0.744**
LR	0.733	0.744	**0.744**
RF	0.748	0.739	**0.749**

5) ***Case Study****:* We built a dataset under the electric equipment hall scenario and tested the MDA method on the dataset.

TABLE VII
MDA EXPERIMENTS ON OUR DATASET.

Models	**ACC@1**
TSM	0.7615
TSM+Tricks	0.7824
TSM+Tricks+MDA2D	0.7947
TSM+Tricks+MDA3D	**0.8035**

The results are shown in the TableVII above: MDA3D method adopted in our dataset has also significantly improved (+4.2%) compared with the original TSM baseline.

Fig. 6. Mulit-person Behavior Recognition based on TSM-MDA3D.

We use a multi-person behavior recognition framework based on MDA3D to recognize behavior. The left picture shows a sitting posture that is occluded, making it difficult to extract the skeleton points of the entire body and accurately recognize the action. The right picture shows standing and walking behaviors that are challenging to distinguish by single-frame detection. The MBR-MDA method effectively recognizes these cases and improves accuracy using MDA while maintaining inference speed.

V. SUMMARY

In this paper, we propose a multi-person behavior recognition method based on multi descriptors aggregations (MBR-MDA) for real-time surveillance scenarios.

To address the limited ability of 2D convolution to model temporal features in video sequences, we propose Multi Descriptors Aggregation by 2D Pooling and 3D Pooling.

Our method is evaluated on HMDB51 and UCF101 datasets, and we demonstrate that MDA3D can effectively generate semantics features. With a slight increase in parameters(+3.1M) and inference time(+0.3ms), our method improves classification Top1-ACC accuracy by 4.8% compared to the baseline. Compared to the Non-local method, our method has fewer parameters, shorter inference time, and higher accuracy, making it suitable for real-time behavior recognition models.

REFERENCES

[1] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 779–788, 2016.

[2] S. Yan, Y. Xiong, and D. Lin, "Spatial temporal graph convolutional networks for skeleton-based action recognition," in *Thirty-second AAAI conference on artificial intelligence*, 2018.

[3] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *Advances in neural information processing systems*, vol. 28, 2015.

[4] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, "Yolov4: Optimal speed and accuracy of object detection," *arXiv preprint arXiv:2004.10934*, 2020.

[5] A. Bewley, Z. Ge, L. Ott, F. Ramos, and B. Upcroft, "Simple online and realtime tracking," in *2016 IEEE international conference on image processing (ICIP)*, pp. 3464–3468, IEEE, 2016.

[6] N. Wojke, A. Bewley, and D. Paulus, "Simple online and realtime tracking with a deep association metric," in *2017 IEEE international conference on image processing (ICIP)*, pp. 3645–3649, IEEE, 2017.

[7] Z. Wang, L. Zheng, Y. Liu, Y. Li, and S. Wang, "Towards real-time multi-object tracking," in *European Conference on Computer Vision*, pp. 107–122, Springer, 2020.

[8] J. Redmon and A. Farhadi, "Yolov3: An incremental improvement," *arXiv preprint arXiv:1804.02767*, 2018.

[9] Y. Zhang, C. Wang, X. Wang, W. Zeng, and W. Liu, "Fairmot: On the fairness of detection and re-identification in multiple object tracking," *International Journal of Computer Vision*, vol. 129, no. 11, pp. 3069–3087, 2021.

[10] S. Sharma, R. Kiros, and R. Salakhutdinov, "Action recognition using visual attention," *arXiv preprint arXiv:1511.04119*, 2015.

[11] D. Tran, L. D. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "C3d: generic features for video analysis," *CoRR, abs/1412.0767*, vol. 2, no. 7, p. 8, 2014.

[12] C. Feichtenhofer, H. Fan, J. Malik, and K. He, "Slowfast networks for video recognition," in *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 6202–6211, 2019.

[13] L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang, and L. V. Gool, "Temporal segment networks: Towards good practices for deep action recognition," in *European conference on computer vision*, pp. 20–36, Springer, 2016.

[14] J. Lin, C. Gan, and S. Han, "Tsm: Temporal shift module for efficient video understanding," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 7083–7093, 2019.

[15] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

[16] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7794–7803, 2018.

[17] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre, "Hmdb: a large video database for human motion recognition," in *2011 International conference on computer vision*, pp. 2556–2563, IEEE, 2011.

[18] K. Soomro, A. R. Zamir, and M. Shah, "Ucf101: A dataset of 101 human actions classes from videos in the wild," *arXiv preprint arXiv:1212.0402*, 2012.