BUPT-MCPRL at TRECVID 2021

Instance search

Yinan Song, Wenhao Yang, Zhicheng Zhao, Yanyun Zhao, Fei Su

Multimedia Communication and Pattern Recognition Labs,
Beijing Key Laboratory of Network System and Network Culture,
Beijing University of Posts and Telecommunications, Beijing 100876, China
{zhaozc, zyy, sufei}@bupt.edu.cn

We describe BUPT-MCPRL Instance Search (INS) system and evaluation results at TRECVID 2021. Specifically, a novel hierarchical multi-task INS retrieval framework is proposed. Firstly, a multi-level action recognition framework and a face matching scheme are introduced to obtain initial action and person retrieval scores separately. In particular, a novel graph-based human-object interaction (HOI) detection model, named interaction-centric graph parsing network (iCGPN), is proposed to recognize interactions between human and objects. Secondly, an improved query extension strategy is adopted to re-rank the initial person retrieval results. Thirdly, more elaborate action features are extracted to recognize complicated actions.

Activities in Extended Videos

Yunhao Du, Junfeng Wan, Binyu Zhang, Zhihang Tong,
Yanyun Zhao, Fei Su, Zhicheng Zhao

Beijing University of Posts and Telecommunications
Beijing Key Laboratory of Network System and Network Culture, China
{zyy, sufei, zhaozc}@bupt.edu.cn

We describe BUPT-MCPRL ActEV system and evaluation results at TRECVID 2021.

1. Training data: VIRAT, MEVA, COCO(pretraining), Kinetics400(pretraining).
2. Approach: (26542: 2 Detectors + 1 Subtractor + 5 Classifiers).
3. Difference: None.
4. Contribution: We believe our 5 classifiers for 5 groups of activities and targeted strategies benefit the performance much.
5. Conclusion: The activity detection task in surveillance videos still needs a comprehensive and complex system.

BUPT-MCPRL at TRECVID 2021: INS*

Yinan Song, Wenhao Yang, Zhicheng Zhao, Yanyun Zhao, Fei Su

Multimedia Communication and Pattern Recognition Labs,
Beijing Key Laboratory of Network System and Network Culture,
Beijing University of Posts and Telecommunications, Beijing 100876, China
{zhaozc, zyy, sufei}@bupt.edu.cn

Abstract

In this paper, we describe BUPT-MCPRL Instance Search (INS) system and evaluation results at TRECVID 2021[1]. Specifically, a novel hierarchical multi-task INS retrieval framework is proposed. Firstly, a multi-level action recognition framework and a face matching scheme are introduced to obtain initial action and person retrieval scores separately. In particular, a novel graph-based human-object interaction (HOI) detection model, named interaction-centric graph parsing network (iCGPN), is proposed to recognize interactions between human and objects. Secondly, an improved query extension strategy is adopted to re-rank the initial person retrieval results. Thirdly, more elaborate action features are extracted to recognize complicated actions. We submit four runs for automatic INS, and a brief description is as follows:

- **F_M_A_B_BUPT_MCPRL.21_1:** Results on 20 main tasks
- **F_P_A_B_BUPT_MCPRL.21_2:** Results on 20 progress tasks
- **F_M_A_B_BUPT_MCPRL.21_3:** Results updated by interval expansion on 20 main tasks
- **F_P_A_B_BUPT_MCPRL.21_4:** Results updated by interval expansion on 20 progress tasks

The final results are summarized in Table 1.

Table 1. Results for each run

Run ID	mAP
F_M_A_B_BUPT_MCPRL.21_1	32.6
F_P_A_B_BUPT_MCPRL.21_2	21.7
F_M_A_B_BUPT_MCPRL.21_3	32.8
F_P_A_B_BUPT_MCPRL.21_4	21.8

1. Method

To balance the efficiency and effectiveness on such a large dataset, we extract video key frames with a sample rate of 5 fps for every shot. To retrieve specific person doing specific action, we consider how to achieve satisfying results in person retrieval and action recognition respectively and fuse them in an appropriate manner.

1.1 Architecture of Instance Search

This year, we propose a novel hierarchical multi-task retrieval framework, as shown in Figure 1. The task is parsed into two main subtasks, that is, person retrieval and action retrieval. More details will be given in the following sections.

*This work is supported by Chinese National Natural Science Foundation (62076033, U1931202), and the National Key R&D Program of China (2020YFB2104604).

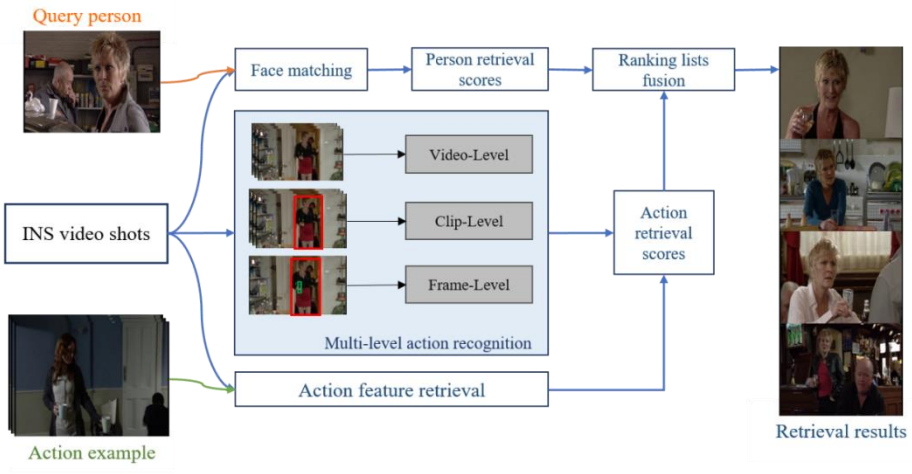

Figure 1. Architecture of Instance Search

1.2 Persons retrieval

The target of person retrieval is to find these video shots that contain the specific persons. This task is solved based on the facial feature representation of query person images. In order to improve the retrieval result, we also introduce a strategy of selecting query images.

1.2.1 Face matching

The core part of the person retrieval task is face matching. As shown in Figure 2, we firstly adopt RetinaFace[2] to detect faces in all shot key frames and query person examples, and then extract face landmarks using PFLD[3] for face alignment. In order to alleviate the problem of undetected faces when persons are away from the camera lens, we introduce DeepSORT[4] to track the persons. Then, FaceNet[5] is used to obtain facial feature representation for cosine similarity matching. We set a threshold for cosine distance to filter the shots of a target person. Besides, in order to improve our retrieval accuracy, we apply top-N α-weighted query extension(αQE)[6] strategy to get more query images. Finally we get an initial person ranklist for our instance search. Figure 3 shows our results of face matching.

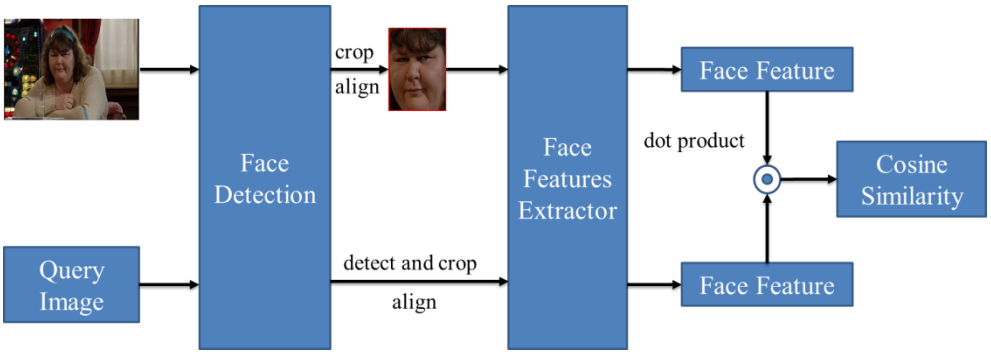

Figure 2. Architecture of Instance Search

1.2.2 Query Selection

Due to angle offset and occlusion etc, some query person examples are helpless for correctly matching according to visualization results. In addition, low brightness and blur will also affect the performance of face matching. Hence, we propose a query selection strategy to remove low relevant

queries, and use these matching results with high similarity to extend query examples. In

Figure 3. Results of face matching. Column 1-5 denote the first, the 1000-th, the 3000-th, the 5000-th, and the 10,000-th retrieval results respectively

this way, we can get better matching results.

Specifically, as shown in Figure 4, our goal is to remove the query image annotated in blue. Firstly, we extract the facial features of four query images, then calculate the cosine similarity in pairs, and display the results in the form of heat map (the middle one). We can see that the similarities between the fourth query and the other three ones are low, so a threshold is set to filter it. In this way, we effectively remove "bad" queries that may get wrong face matching results.

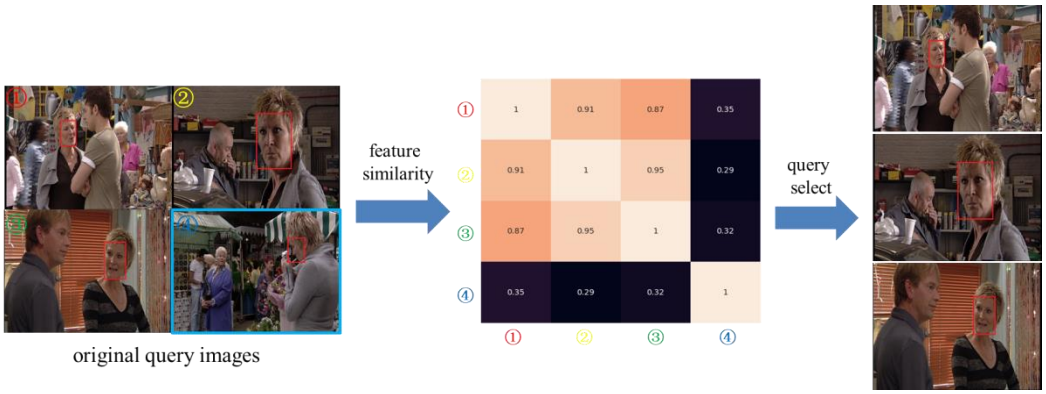

Figure 4. Illustration of the proposed query selection strategy

1.3 Action Retrieval

For action retrieval, we propose a hierarchical multi-level action recognition framework, including frame-level, clip-level and video-level, to enhance the performance of action recognition. In addition, elaborate action feature retrieval methods are adopted to improve the recognition accuracy of those actions which are rarely involved in action recognition datasets.

1.3.1 Video-level action recognition

Video-level action recognition aims to recognize actions from videos and we adopt SlowFast model pretrained on Kinetics-400[7] dataset to roughly judge whether the action occurs in video shots of INS database. Then we take the shot scores as the scores of all key frames in the shot.

1.3.2 Clip-level action recognition

The goal of clip-level action recognition is to localize persons in key-frames, and meanwhile, recognize actions from video clips. Compared with the video-level methods, it can obtain action scores of specific persons at key-frames level. We firstly train the SlowFast model on AVA-Kinetics[8] dataset. Then we use Cascade R-CNN[9] pretrained on COCO dataset to locate the persons in key-frames of INS video shots. With the pretrained model above, the action scores of each detected person in key-frames are obtained.

1.3.3 Frame-level HOI detection

It aims to recognize actions from a single frame, which contain obvious human-object interactions (HOI), such as 'sit on couch' and 'holding glass'. A HOI detection model, named iCGPN, which trained on HICO-DET[10] dataset is adopted. As shown in Figure 5, the node feature is represented by visual features and spatial information between human and objects. Then a graph convolutional network is applied to update node features according to the connectivity matrix and predict the final HOI labels. In addition, HOI detection consists of two main steps: object detection and HOI prediction. And the performance of HOI detection model relies on the object detection model to a great extent. Thus, we conduct a new object detection dataset to train Cascade R-CNN. With the trained Cascade R-CNN and iCGPN, we can get the interaction scores for each human-object pair.

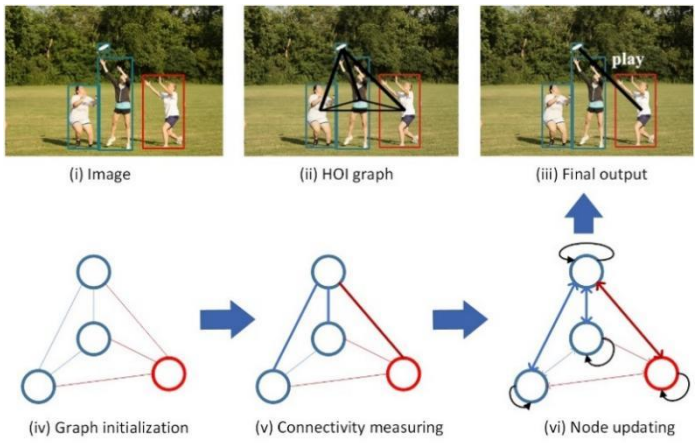

Figure 5. Illustration of the proposed interaction-centric graph parsing networks. Aiming to recognize the interaction between the right human (red box) and frisbee, we treat red node as central human node and blue nodes as object nodes.

1.3.4 Action feature retrieval

Some actions related to doors, such as "open door enter", are rarely involved in existing external datasets of action recognition. Furthermore, we cannot get prediction scores by the multi-level action recognition framework. Thus, we adopt action feature retrieval methods to handle these actions. The SlowFast[11] model pretrained on AVA-Kinetics is adopted to extract action features of key frames in query action examples and INS shots. Then we calculate the cosine similarities among them and set maximal similarity as the shot key-frames similarity. Figure 6 gives several visualization retrieval results. The results show that the action recognition scores fusion strategy can effectively exclude lots of wrong results.

Figure 6. Two groups of visualization video retrieval results. a) and (b) show Max is sitting on couch and (c) and (d) show Bradley is holding glass.

1.3.5 Emotion-related action retrieval

Considering that some actions such as shouting, laughing and crying are strongly correlated with facial expressions, lightweight Convolutional Neural Network (CNN) pretrained on CK+ [12] and FERPLUS[13] is applied to recognize these emotion related actions. Data augmentation is also performed, including horizontal flip and random cropping etc.

1.4 Ranking lists fusion

To obtain the better performance, we apply late fusion in the post-processing stage. For the ith key frame, the fusion score is calculated as the following equation.

$$f_i = \beta_1 p_i + \beta_2 a_i \tag{1}$$

$$s.t.\ \beta_1 + \beta_2 = 1 \tag{2}$$

Where $f_i,\ p_i,\ a_i$ respectively denote to the final scores, person retrieval scores and action retrieval scores, and $\beta_1,\ \beta_2$ are weight hyperparameters. Considering the action retrieval is more difficult than person retrieval, in experiments, we set $\beta_1 \leq \beta_2$.

Finally, according to the final scores of all shot key-frames, we take the maximal score of the key-frames in each shot as the shot scores to generate the shot ranking lists for all person-action pairs of INS task.

2. Conclusion

This year, we have effectively improved the overall performance of instance retrieval. We propose a novel multi-level INS framework, where specific persons and actions retrieval are accomplished and HOI is introduced to improve action recognition, and the results are fused by two ranking schemes. First, person retrieval scores are obtained by weighted ranking lists. Second, action retrieval scores are computed based on the proposed multi-level action recognition framework and action feature retrieval methods. Finally, action and person retrieval scores are merged to obtain the final shot ranking lists. In the future, We will focus on end-to-end trainable HOI models, and the integration of text and audio information.

Acknowledgment

In this paper, BBC video snapshots are used for non-commercial individual research and private study use only. BBC content included courtesy of the BBC.

References

[1] George Awad, Asad A. Butt, Keith Curtis, Jonathan Fiscus, Afzal Godil, Yooyoung Lee, Andrew Delgado, Jesse Zhang, Eliot Godard, Baptiste Chocot, Lukas Diduch, Jeffrey Liu, Yvette Graham, Gareth J. F. Jones, Georges Quénot. Evaluating Multiple Video Understanding and Retrieval Tasks at TRECVID 2021.

[2] Jiankang Deng, Jia Guo, Yuxiang Zhou, Jinke Yu, Irene Kotsia, and Stefanos Zafeiriou. Retinaface: Singlestage dense face localisation in the wild. arXiv preprint arXiv:1905.00641, 2019.

[3] Xiaojie Guo, Siyuan Li, Jinke Yu, Jiawan Zhang, Jiayi Ma, Lin Ma, Wei Liu, and Haibin Ling. Pfld: A practical facial landmark detector. arXiv preprint arXiv:1902.10859, 2019.

[4] Nicolai Wojke, Alex Bewley, and Dietrich Paulus. Simple online and realtime tracking with a deep association metric. In 2017 IEEE international conference on image processing (ICIP), pages 3645–3649. IEEE, 2017.

[5] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 815–823, 2015.

[6] Filip Radenovi´c, Giorgos Tolias, and Ondˇrej Chum. Finetuning cnn image retrieval with no human annotation. IEEE transactions on pattern analysis and machine intelligence, 41(7):1655–1668, 2018.

[7] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. arXiv preprint arXiv:1705.06950, 2017.

[8] Ang Li, Meghana Thotakuri, David A Ross, Joˇao Carreira, Alexander Vostrikov, and Andrew Zisserman. The ava-kinetics localized human actions video dataset. arXiv preprint arXiv:2005.00214, 2020.

[9] Zhaowei Cai and Nuno Vasconcelos. Cascade r-cnn: Delving into high quality object detection. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 6154–6162, 2018.

[10] Yu-Wei Chao, Yunfan Liu, Xieyang Liu, Huayi Zeng, and Jia Deng. Learning to detect human-object interactions. In 2018 ieee winter conference on applications of computer vision (wacv), pages 381–389. IEEE, 2018.

[11] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In Proceedings of the IEEE/CVF international conference on computer vision, pages 6202–6211, 2019.

[12] Patrick Lucey, Jeffrey F Cohn, Takeo Kanade, Jason Saragih, Zara Ambadar, and Iain Matthews. The extended cohnkanade dataset (ck+): A complete dataset for action unit and emotion-specified expression. In 2010 ieee computer society conference on computer vision and pattern recognitionworkshops, pages 94–101. IEEE, 2010.

[13] Barsoum E , Zhang C , Ferrer C C , et al. Training Deep Networks for Facial Expression Recognition with Crowd-Sourced Label Distribution. ACM International Conference on Multimodal Interaction, ICMI 2016.

BUPT-MCPRL at TRECVID 2021 ActEV: 215AD*

Yunhao Du, Junfeng Wan, Binyu Zhang, Zhihang Tong, Yanyun Zhao, Fei Su, Zhicheng Zhao

Beijing University of Posts and Telecommunications
Beijing Key Laboratory of Network System and Network Culture, China
{zyy, sufei, zhaozc}@bupt.edu.cn

Abstract

1. Training data: VIRAT, MEVA, COCO(pretraining), Kinetics400(pretraining).
2. Approach: (26542: 2 Detectors + 1 Subtractor + 5 Classifiers).
3. Difference: None.
4. Contribution: We believe our 5 classifiers for 5 groups of activities and targeted strategies benefit the performance much.
5. Conclusion: The activity detection task in surveillance videos still needs a comprehensive and complex system.

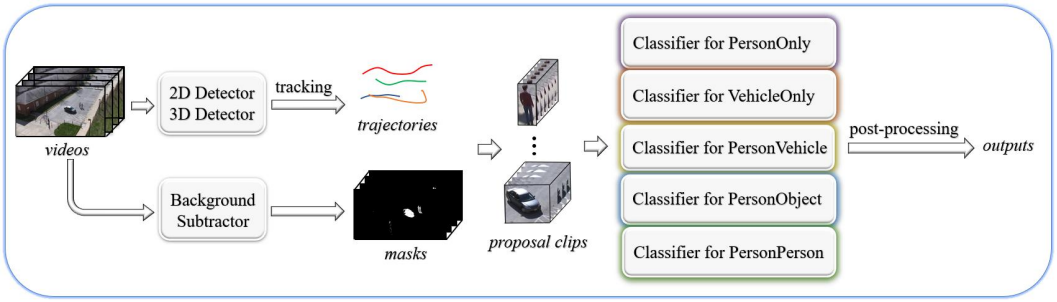

Figure 1. The framework of our activity detector, named 215AD.

1. Method

First of all, 35 categories of activities in VIRAT are divided into 5 groups according to their characteristics:

1) Person-only: *person crouches, person sits, person stands, person gestures, person walks, person runs, person uses tool, person carries object, person talks on phone, person texts on phone;*

2) Vehicle-only: *vehicle makes u turn, vehicle turns left, vehicle turns right, vehicle starts, vehicle stops, vehicle moves;*

3) Person-Vehicle: *person opens facility or vehicle door, person closes facility or vehicle door, person enters facility or vehicle, person exits facility or vehicle, person opens trunk, person closes trunk, person loads vehicle, person unloads vehicle, vehicle drops off person, vehicle picks up person;*

4) Person-Object: *person rides bicycle, person interacts object, person carries heavy object, person pickups object, person sets down object, person pulls object, person pushs object;*

5) Person-Person: *person person interaction, person talks to person.*

Based on that, we propose a comprehensive framework for activity detection in surveillance videos, named 215AD. As shown in figure 1, it consists of **2** detectors, **1** background subtractor and **5** classifiers. Finally, a post-processing strategy is applied similar with [9] to generate

*This work is supported by Chinese National Natural Science Foundation (62076033, U1931202), and the National Key R&D Program of China (2020YFB2104604).

activities detection results.

Figure 2 shows the difference of our 2D and 3D detectors. For all person and vehicle objects, we adopt 2D Cascade R-CNN [2] for frame-level bounding boxes, and DeepSORT [3] for tracking. For some complex activities, we directly generate 3D proposals using 3D Cascade R-CNN [4] and IOUTracker [5].

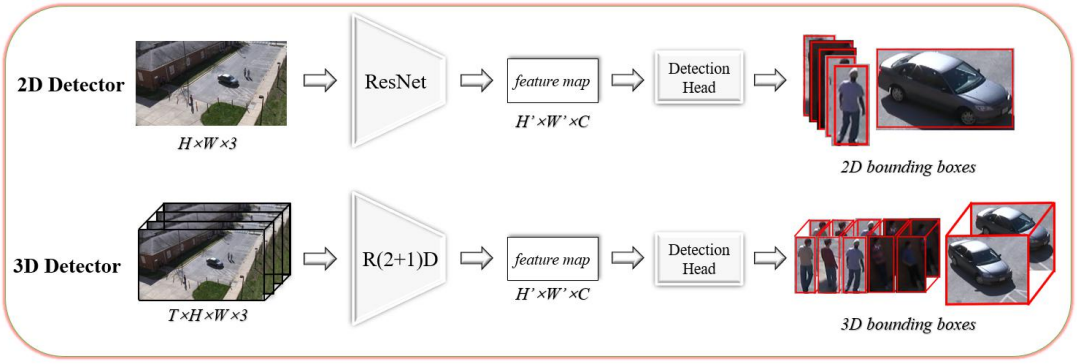

Figure 2. Our 2D detector and 3D detector.

As shown in figure 3, MOG2 [6] is utilized to generate foreground masks for each frame. Then a median filter with a 3 × 3 kernel is applied for denoising. We combine the results of the detectors and subtractor to get activity proposals for the classifiers.

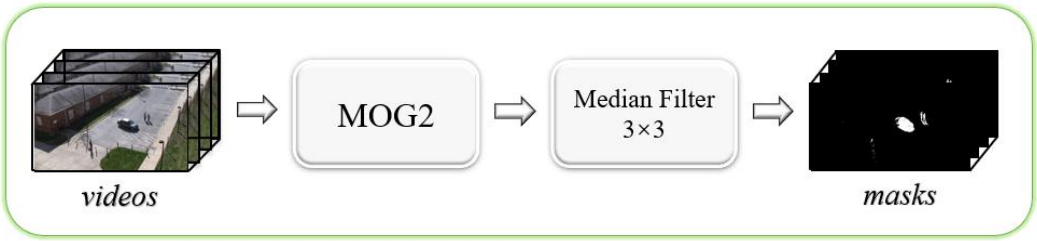

Figure 3. Our background subtractor.

As for classification, we use MViT [7] or X3D [8] as baseline, and design targeted strategies for 5 groups of activities. For person-only activities, we propose the part-attention mechanism (figure 4). By combining the global scores $\vec{s}_1$ and part scores $\vec{s}_2, \vec{s}_3$, we could force the classifiers not only capture global information, but also focus on detailed features. This could benefit activities which are related to postures and gestures, e.g., "person crouches", "person_talks_on_phone" and "person gestures".

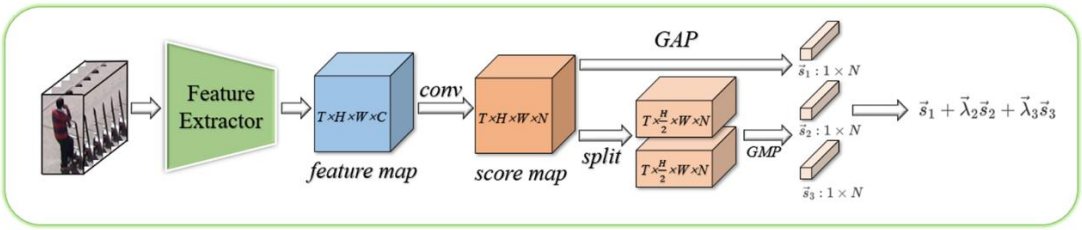

Figure 4. Our person-only classifier.

For vehicle-only activities (figure 5), a simple but effective motion information embedding method is proposed. Given frame-level bounding boxes, we get a Motion clip by calculating the inter-frame displacement. Then RGB clip and Motion clip are concatenated as the input of the activity classifier. This improves the performance of vehicle-only activities by a large margin.

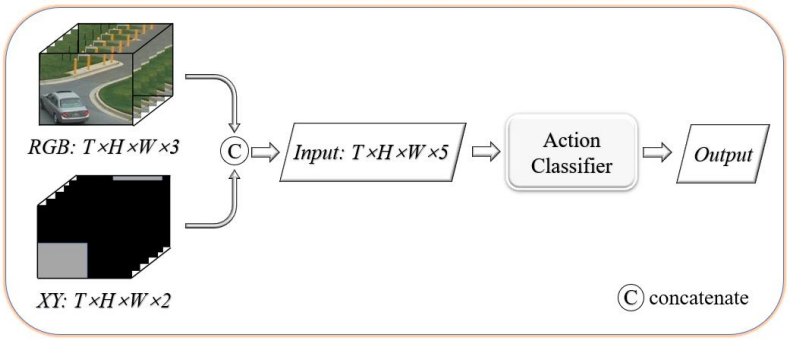

Figure 5. Our vehicle-only classifier.

For person-vehicle activities (figure 6), we exploit a GCN module to model the relationship between vehicle and person. Besides, a CNN module is used to fuse object features and a spatial attention module to add spatial information. All these 3 features are multiplied first and then concatenated with the original clip features.

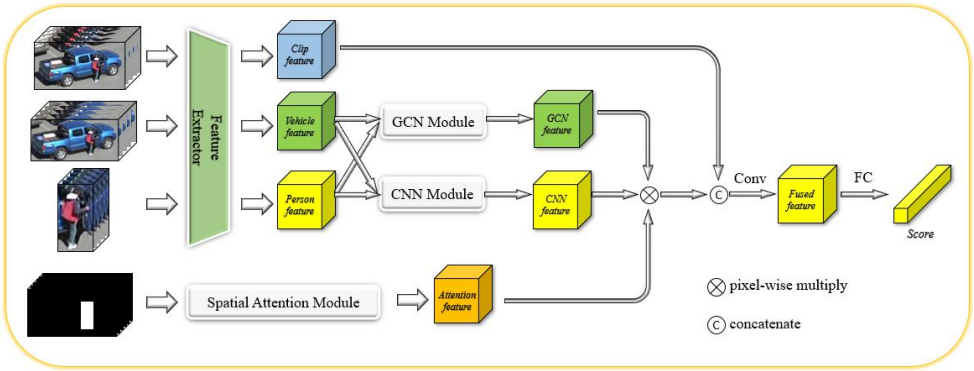

Figure 6. Our person-vehicle classifier.

For person-object activities (figure 7), we add an object regressor to make the backbone be aware of where the object is. It could be seen a kind of attention mechanism. Note that no object coordinate information is needed during inference.

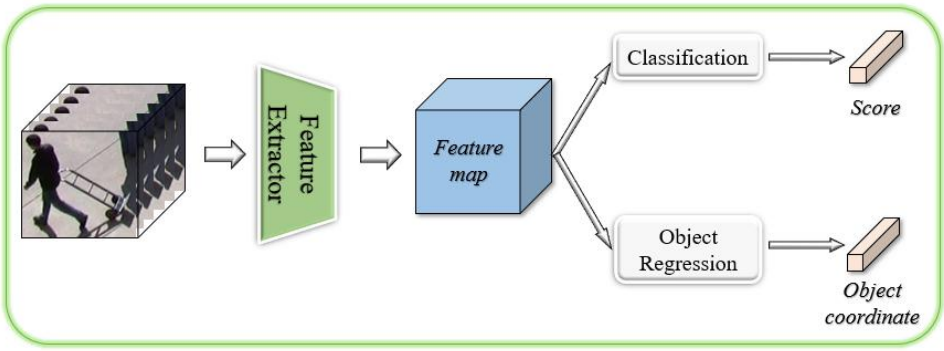

Figure 7. Our person-object classifier.

For person-person activities (figure 8), we add a suppression procedure after classification. If the number of people in that clip is less than 2, the clip is thought as background class.

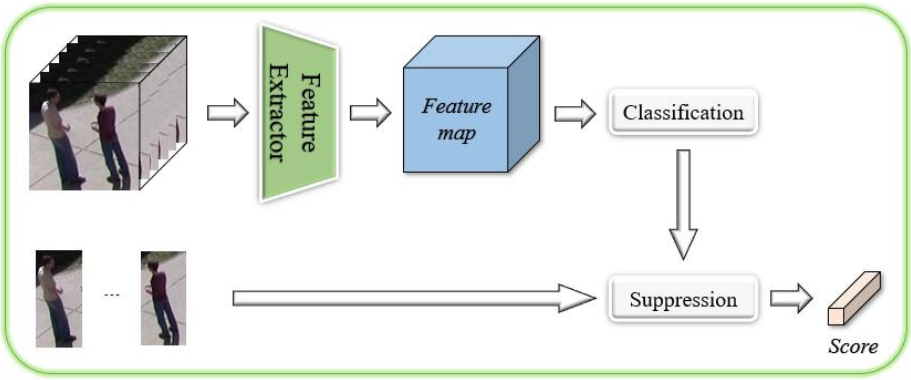

Figure 8. Our person-person classifier.

2. Results

Figure 9 and table 1 show the results of TRECVID 2021 ActEV challenge. Our system 215AD achieves nAUDC@0.2TFA=0.408 and wins the first place, which demonstrates the effectiveness of our approach.

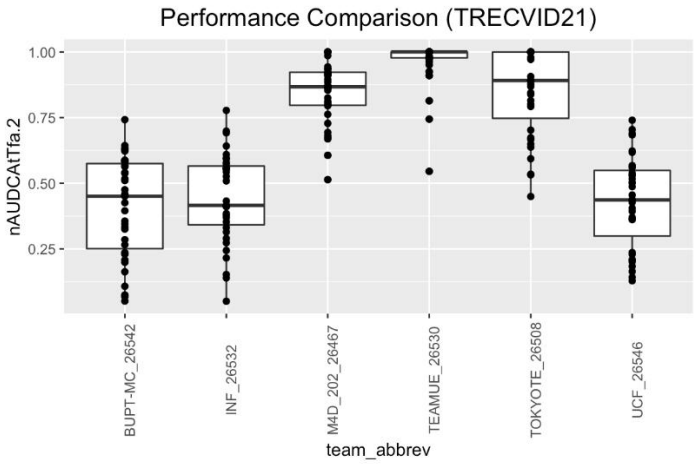

Figure 9. Box plots in TRECVID 2021 ActEV.

Table 1. Results in TRECVID 2021 ActEV.

Team	nAUDC
ours	**0.4085291**
UCF	**0.4305868**
INF	**0.4443625**
M4D	0.8465777
TTA	0.8515892
UEC	0.9640481

Reference

[1] George Awad, Asad A. Butt, Keith Curtis, Jonathan Fiscus, Afzal Godil, Yooyoung Lee, Andrew Delgado, Jesse Zhang, Eliot Godard, Baptiste Chocot,

Lukas Diduch, Jeffrey Liu, Yvette Graham, Gareth J. F. Jones, and Georges Qu´enot, “Evaluating multiple video understanding and retrieval tasks at trecvid 2021,” in Proceedings of TRECVID 20 21. NIST, USA, 2021.
[2] Cai Z, Vasconcelos N. Cascade r-cnn: Delving into high quality object detection[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 6154-6162.
[3] Wojke N, Bewley A, Paulus D. Simple online and realtime tracking with a deep association metric[C]//2017 IEEE international conference on image processing (ICIP). IEEE, 2017: 3645-3649.
[4] Li Y, Xu S, Cheng X, et al. An Effective Detection Framework for Activities in Surveillance Videos[J]. BUPT-MCPRL team report: https://www.mcprl.com/essay/BUPT-MCPRL_report_for _ActEV-PC. pdf, 2019.
[5] Bochinski E, Eiselein V, Sikora T. High-speed tracking-by-detection without using image information[C]//2017 14th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS). IEEE, 2017: 1-6.
[6] Zivkovic Z, Van Der Heijden F. Efficient adaptive density estimation per image pixel for the task of background subtraction[J]. Pattern recognition letters, 2006, 27(7): 773-780.
[7] Fan H, Xiong B, Mangalam K, et al. Multiscale vision transformers[J]. arXiv preprint arXiv:2104.11227, 2021.
[8] Feichtenhofer C. X3d: Expanding architectures for efficient video recognition[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020: 203-213.
[9] *https://www-nlpir.nist.gov/projects/tv2020/tv20.workshop.notebook/tv20.papers/inf.pdf*