UEC at TRECVID 2020: INS and ActEV

Sosuke Mizuno Keiji Yanai
The University of Electro-Communications, Tokyo , Japan
{mizuno-s,yanai}@mm.inf.uec.ac.jp

***Abstract*—In this paper, we describe our systems and evaluation results at TRECVID 2020. We participated two tasks, INstance Search (INS) and Activity in Extended Video (ActEV).**

I. INS: INSTANCE SEARCH

The purpose of this task is to search for videos related to the behavior of a particular person and action. Our method consists of two parts: the first is person retrieval; the second is action retrieval. The action retrieval is performed only on the videos at the top of the person retrieval. The action retrieval is based on a combination of facial expression recognition, object detection, and general action recognition. Figure 1 shows the overall of our approach.

A. Person retrieval

The person retrieval compares the person in the query with the facial features of the people in each video to obtain a person similarity score. First, the faces in each frame are detected by RetinaFace[1] and cropped. Then, the facial features are extracted from the cropped images using ArcFace[2]. For efficiency, we perform person retrieval on videos with person similarity scores of 0.3 or higher.

B. Action retrieval

Action retrieval consists of three parts: Emotion-related action retrieval, Human-Object Interaction retrieval, and General action retrieval. Each component is selected according to the action class of the query. Table I shows the corresponding relationships between components of action retrieval and action classes.

TABLE I
THE CORRESPONDING RELATIONSHIPS BETWEEN COMPONENTS OF ACTION RETRIEVAL AND ACTION CLASSES.

Retrieval method	Classes in INS
Emotion	crying, laughing, shouting
Human-Object	sit on couch, holding paper drinking, holding cloth, holding phone
General	smoking cigarette, go up down stairs kissing, open door enter, hugging open door leave, stand talk door close door wo leaving

1) Emotion-related action retrieval: Some action classes, such as "crying", can be identified only by recognizing facial expressions. These classes are identified by the method in [3]. We used a model trained on the FER2013 dataset. The mapping between the FER2013 dataset and the INS classes is shown in Table II.

TABLE II
THE MAPPING BETWEEN FER2013 DATASET AND INS CLASSES.

FER2013	INS
sad	crying
happy	laughing
angry	shouting

2) Human-Object interaction retrieval: Some of the action classes are related to human-object interaction, such as "holding phone". It detects objects around a human and counts the number of frames detected. The interaction score is the ratio of the number of object counts to the number of frames. We used EfficientDet[4], which was pre-trained in MS-COCO.

3) General action retrieval: Other general classes are recognized by SlowFast[5]. We fine-tune the SlowFast pre-trained by Kinetics-600 with INS data.

C. Results

INS has two submission types, i.e., Fully Automatic (F) runs and Interactive (I) runs, depending on human intervention is involved or not. In this time, we focus on the Fully Automatic runs. And all teams should demonstrate the types of training data by the notations of "A" and "E", in which "A" means video examples are not used while 'E' is the opposite. We used the images and videos provided by NIST for training. The table III shows the results for each team. "UEC-1" is the result obtained by the proposed method and "UEC-2" is the result of a random selection. Table IV shows the accuracy for each action. From Table IV, we can see that the retrieval by facial expression recognition is working to some extent.

TABLE III
INS RESULTS.

Type		Team	mAP
F	E	PKU-WICT	0.252
		WHU-NERCMS	0.151
	A	PKU-WICT	0.247
		BUPT-MCPRL	0.142
		NII-UIT	0.091
		UEC-1(ours)	0.022
		UEC-2(ours)	0.0
I	E	PKU-WICT	0.368

D. Discussion

From Table III, we can see that our results are less accurate than other teams. This may be due to the failure of the fine-tune models in the general action category. In this time, we

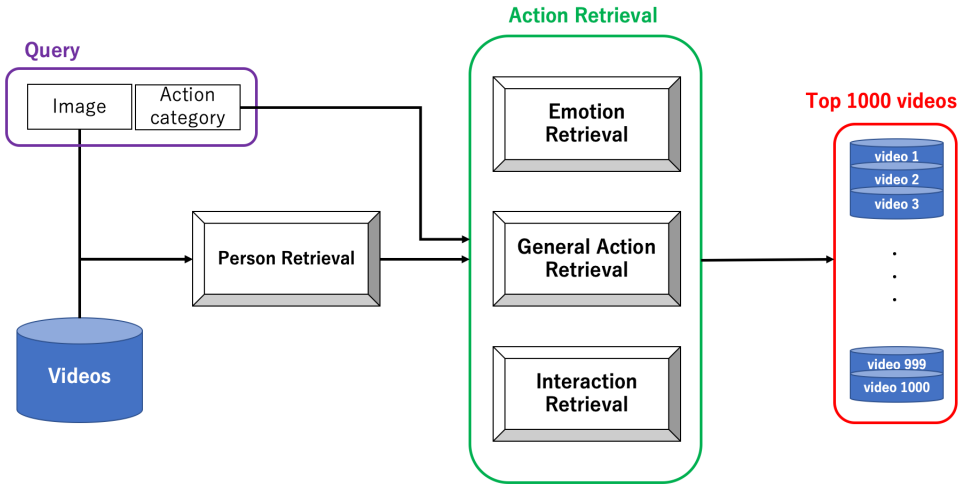

Fig. 1. The overall of our approach

TABLE IV
THE ACCURACY PER ACTION CATEGORY.

action	mAP
laughing	0.080
crying	0.051
holding phone	0.049
sit on couch	0.014
smoking cigarette	0.008
drinking	0.006
holding paper	0.005
holding cloth	0.004
go up down stairs	0.002

fine-tuned it using the videos provided by NIST and some of the Kinetics. The cause of failure is assumed to be the lack of training data and imbalance. In order to improve the accuracy, we believe that it is necessary to extract action features from the detected person's bounding box.

II. ACTEV: ACTIVITY IN EXTENDED VIDEO

ActEV is a very challenging task because it requires precise spatial and temporal localization. Our approach consists of three parts, proposal generation, action classification, and post-processing. Figure 2 shows the overall of our approach.

A. Proposal generation

Here, we extract candidate areas for action from the input video. First, we use Faster R-CNN[6] to detect humans and cars from the input frame. We utilize Fater R-CNN with feature pyramid network[7] on ResNet-101. The model trained on the COCO dataset was fine-tuned using the VIRAT dataset. Next, we use deep SORT[8] to generate a tracking trail for each object. Finally, we generate event proposals from the trajectory of a single object, a person and a car. An event proposal can be treated as a row of bounding boxes cut out of each frame. In this study, we classify each of the proposals into one of three categories, Person, Vehicle and Person-Vehicle. "Person" category includes only events that occurred in a single person. "Vehicle" category includes only events occurring in a single vehicle. "Person-Vehicle" category proposes events in relation to a human and a vehicle. If the spatial distance between the human trajectory and the vehicle trajectory is less than the threshold, a bounding box containing a human and a vehicle is proposed.

B. Activity Classification

1) Feature Extract: We extracted features for action classification in a 3D-ResNet[9] model. We used a 3D ResNet-101 model pre-trained with Kinetics-600.

2) Spatial-Temporal Classification: We utilize a bi-directional LSTM[10] to perform temporal classification to localize activities within spatial-temporal proposals.

C. Post-processing

Candidates after localization and classification may be spatially and temporally overlap. We employ a spatially-temporal NMS to avoid overlapping candidates.

D. Results

Figure 3 and 4 shows the evaluation and validation results of ActEV. Table V shows the results of our systems, "UEC" and "UEC-Test". "UEC-Test" is the result of the proposed method. "UEC" is the result of re-tracking the bounding box obtained from the proposed method. We set all the confidence scores to 1 in "UEC".

TABLE V
OUR SYSTEM RESULTS.

System	nAUDC
UEC	0.95168
UEC-Test	0.96374

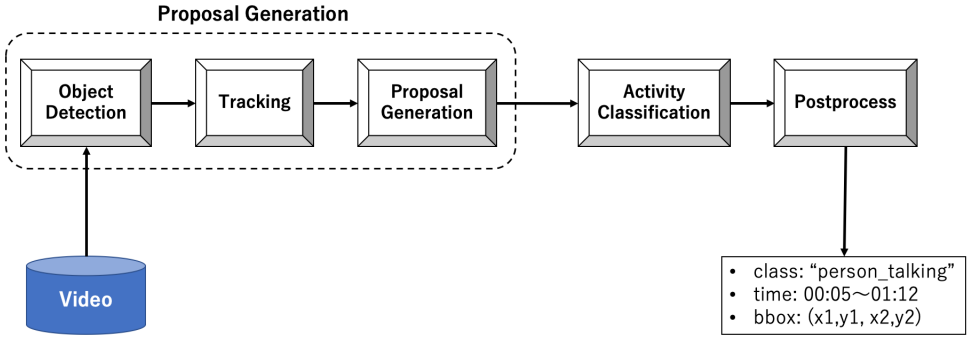

Fig. 2. The overall of our approach.

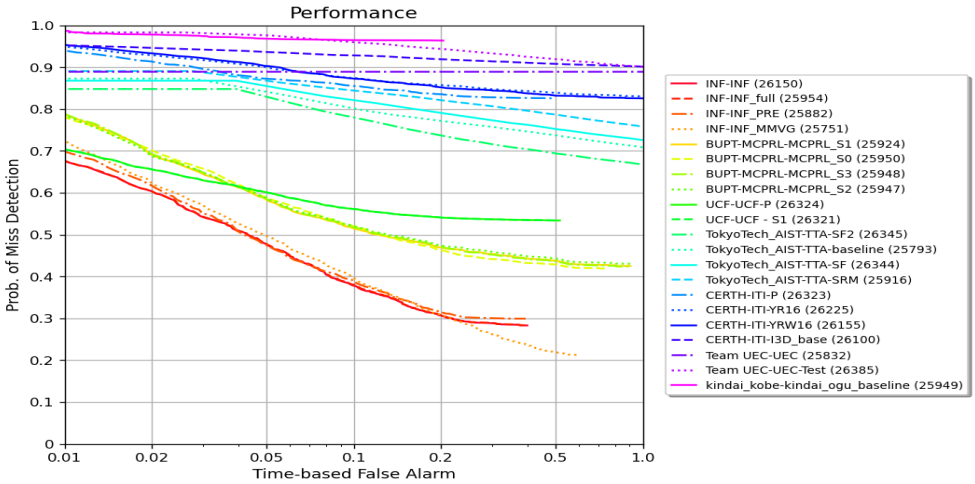

Fig. 3. ActEV evaluation results.

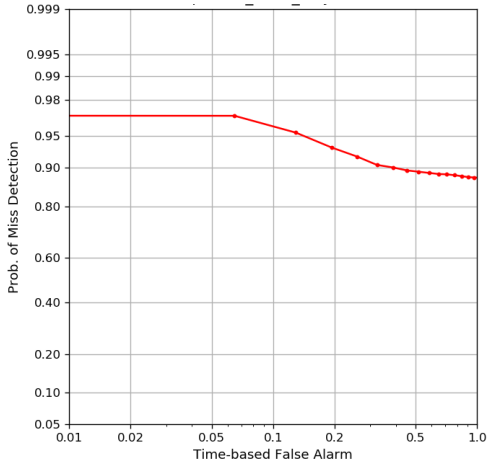

Fig. 4. ActEV validation results.

E. Discussion

We find that our method is less accurate than other methods. This may be due to the fact that our method is hardly able to detect action classes with little training data. Another reason could be the poor recognition of action classes where person and vehicles interact with each other. In order to improve accuracy, these problems need to be solved.

III. Conclusion

This was our first time participating in the INS and ActEV task in TRECVID [11]. This year, we experimented with a baseline method. Our results of this year were not as good as those of other teams. We will keep improving our system for TRECVID 2021.

References

[1] J. Deng, J. Guo, Y. Zhou, J. Yu, I. Kotsia, and S. Zafeiriou. RetinaFace: Single-stage dense face localization in the wild. In *arXiv:1905.00641*, 2019.

[2] J. Deng, J. Guo, N. Xue, and S. Zaferiou. ArcFace: Additive angular margin loss for deep face recognition. In *Proc.of IEEE Computer Vision and Pattern Recognition*, 2019.

[3] L Pham and T. A. Tran. Facial expression recognition using residual masking network. 2020.

[4] M. Tan, R. Pang, and Q. V. Le. Efficientdet: Scalable and efficient object detection. In *Proc.of IEEE Computer Vision and Pattern Recognition*, 2020.

[5] C. Feichtenhofer, H. Fan, J. Malik, and K. He. Slowfast networks for video recognition. In *Proc.of IEEE International Conference on Computer Vision*, 2019.

[6] S. Ren, K. He, R. Girshick, and J. Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In *Proc.of Neural Information Processing System*, 2015.

[7] T. Lin, P. Dollar, R. Girshick, K. He, B. Hariharan, and S. Belong. Feature pyramid networks for object detection. In *Proc.of IEEE Computer Vision and Pattern Recognition*, 2017.

[8] 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)*, pages 3645–3649. IEEE, 2017.

[9] K. Hara, H. Kataoka, and Y. Satoh. Learning spatio-temporal features with 3d residual networksfor action recognition. In *Proc.of IEEE International Conference on Computer Vision*, 2017.

[10] S. Hochreiter and J. Schmidhuber. Long short-term memory. In *Neural computation 9*, page 1735–1780, 1997.

[11] G. Awad, A. A. Butt, K. Curtis, Y. Lee, J. Fiscus, A. Godil, A. Delgado, J. Zhang, E. Godard, L. Diduch, J. Liu, Alan F. Smeaton, Yvette Graham, G. J. F. Jones, W. Kraaij, and G. Quénot. TRECVID 2020: comprehensive campaign for evaluating video retrieval tasks across multiple application domains. In *Proc. of TRECVID 2020*, 2020.