TRECVID2020@UEC

Sosuke Mizuno Yanai Keiji
The University of Electro-Communications

INS: Instance Search

Method

A. Person Retrieval

(1)Detect faces in each frame by *RetinaFace[1]*.
(2)Extract the facial feature by *ArcFace[2]*.
(3)Calculate the similarity score.

B. Action Retrieval

(1) Emotion-related action retrieval.

Some action classes, such as “crying”, can be identified only by recognizing facial expressions by the method in [3].

(2) Human-Object interaction retrieval.

Some of the action classes are related to human-object interaction, such as “holding phone”.The interaction score is the ratio of the number of object around a human counts to the number of frames. We used EfficientDet, which was pre-trained in MS-COCO.

(3) General action retrieval.

Other general classes are recognized by SlowFast. We fine-tune the SlowFast pre-trained by Kinetics-600 with INS data.

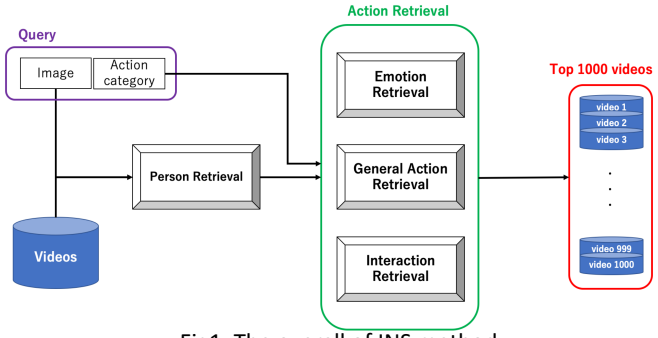

Fig1. The overall of INS method

Results

- UEC-1: the proposed method
- UEC-2: random selection

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, because we use only the videos provided by NIST and some of the Kinetics.
→ We believe that it is necessary to extract action features from the detected person’s bounding box.

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

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

ActEV: Activity in Extended Video

Method

A. Proposal generation

(1)Detect humans and cars by Faster R-CNN.
(2)Generate a tracking trail by deep SORT[4].

B. Activity Classification

(1)Extract features by 3D ResNet-101.
(2)Temporal localization by bi-directional LSTM.

C. Post-processing

We employ a spatially-temporal NMS to avoid overlapping candidates.

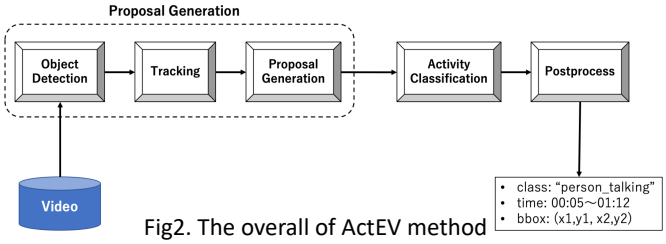

Fig2. The overall of ActEV method

Results

- UEC-Test: the proposed method
- UEC: re-tracking the bounding box obtained from the proposed method

We find that our method is less accurate than other methods
→ This may 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.

System	nAUDC
UEC	0.95168
UEC-Test	0.96374

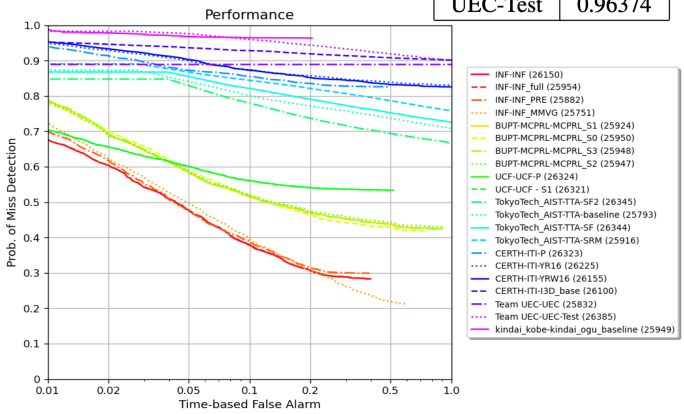

Fig3. ActEV evaluation results.

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. *https://github.com/phamquiluan/ResidualMaskingNetwork*. 2020.
[4] 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.