Inf@TRECVID 2019: Instance Search Task

En Yu[1,2], Wenhe Liu[2], Guoliang Kang[2], Xiaojun Chang[3], Jiande Sun[1] and Alexander Hauptmann[2]

[1]Shandong Normal University, [2]Carnegie Mellon University, [3]Monash University

Abstract

We participated in one of the two types of Instance Search task in TRECVID 2019: Fully Automatic Search, without any human intervention. Firstly, the specific person and action are searched separately, and then we re-rank the two sorts of search results by ranking the one type scores according to the other type, as well as the score fusion. And thus, three kinds of final instance search results are submitted. Specifically, for the person search, our baseline consists of face detection, alignment and face feature selection. And for the action search, we integrate person detection, person tracking and feature selection into a framework to get the final 3D features for all tracklets in video shots. The official evaluations showed that our best search result gets the 4th place in the Automatic search.

1 Task Description

In TRECVID 2019 [1], a query type in Instance Search task is proposed to retrieve specific person doing specific actions [2]. And it also derives two submission types, i.e., Fully Automatic (F) runs and Interactive (I) runs, depending on human intervention is involved or not. In detail, 30 queries are released, and each of them includes 4 example images for person containing the person of interest and example videos for the corresponding action. Besides, a mass of video shots segmented from BBC Eastenders test videos are given as the retrieved samples, while the type of training data can be chosen according to official requirements by ourselves. 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 only focus on the Fully Automatic search and the video examples are also used in our method. And thus, Table 1 shows all of our submissions and the evaluation results (MAP) according to different ranking strategies. Also, the comparisons with other teams are illustrated in Table 2 under the same setting, and we can find our team gets the 4th place in Fully Automatic task.

2 Our Method

In this time, we focus on the Fully Automatic runs and design two baselines for person search and action search respectively, and also three re-ranking strategies are used to get the final submissions [3].

2.1 Person Search

For person search, as shown in Figure 1, we firstly utilize the MTCNN model [4] to detect and crop the faces from frames, and then the cropped faces are fed into the face recognizer VGG-Face2 [5] for feature selection. Finally, we use the Cosine Distance to measure the similarities between queries and retrieved samples.

Table 1: Results of our submissions

Task Type	Submission ID	MAP	Ranking Type
F	Inf_run1_E	0.017	Person-Based
	Inf_run2_E	0.013	Action-Based
	Inf_run3_E	0.001	Fusion-Based

Table 2: Comparison with other teams at the same setting

Task Type	Team+Submission (best)	MAP
F	PKU_ICST + run_F_E	0.239
	BUPT_MCPRL + run_F_E	0.119
	NII_Hitachi_UIT + run_F_E	0.024
	Inf + run_F_E (ours)	0.017
	WHU_NERCMS + run_F_E	0.017
	HSMW_TUC +run_F_E	0.009

2.2 Action Search

For action search, from Figure 2, we can find that the Faster-RCNN [6] model pre-trained on MSCOCO dataset [7] is firstly used for person detect. In order to include the actions and objects completely, the proposals generated by person detection are expanded by 15% to the periphery. And then the tracklets for each person are generated via DeepSort [8] Tracking algorithm. After that, we fine-tune the RGB benchmark of I3D [9] model on the combination Charades dataset [10] and the offered video shots to extract the features of tracklets. Similarly, the Cosine similarity is used for action ranking.

2.3 Re-Ranking

Since we have got the person ranking and action respectively, three re-ranking methods are proposed to obtain the final results. Here we regard the query, 'Ian + Holding_Glass', as an example to describe these strategies. (1) Person-Search-Based: this strategy aims at using the person ranking to re-rank the action search rank list. We first select the shotID list about 'Ian' from person search list and the ShotID list about 'Holding_Glass' from action search list, respectively. And then the intersection of these two shotID lists. Finally, the final rank list is based on the intersection from the selected person search rank list. (2) Action-Search-Based: the first step is same with Person-Search-Based strategy, while the third step is based on the intersection from the selected action search rank list. (3) Fusion-Based: it is re-ranked by the average similarities between person search and action search.

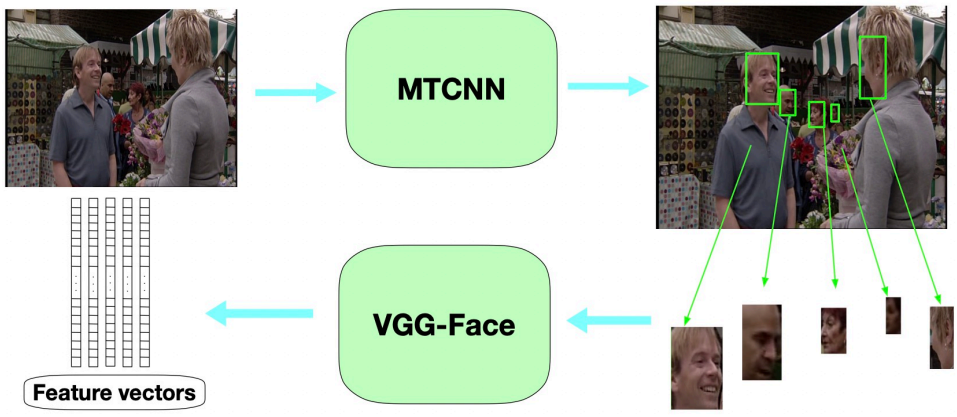

Figure 1: Framework of the face feature selection

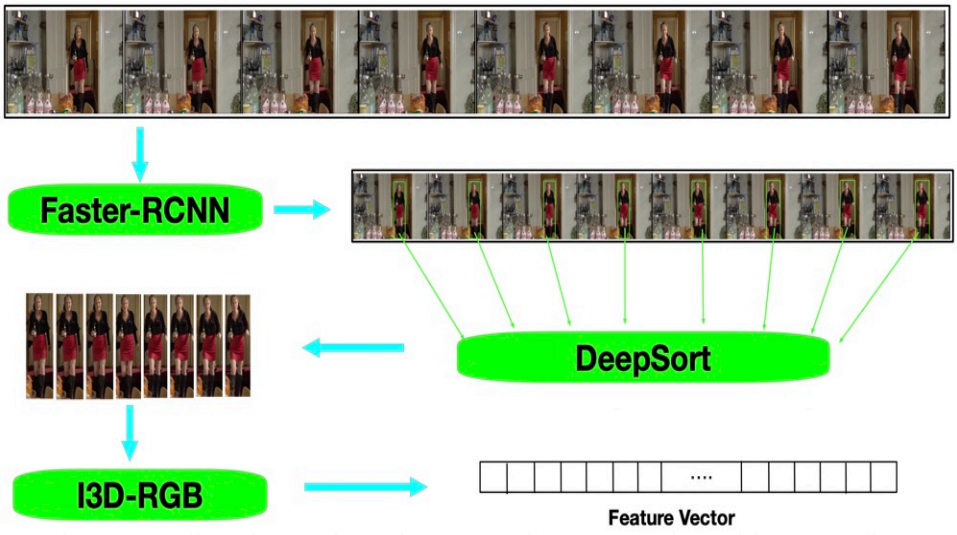

Figure 2: Framework of the action feature selection

3 Conclusion

We designed the person search and action search baseline for feature selection, and three re-ranking strategies were used for final submission. By doing the INS task in TRECVID 2019, we find that many target people only have side faces or it is very blurred in the testing dataset. So in future work, we will add super-resolution processing and person ReID technology to improve the accuracy of person search.

From the evaluation results, we can find the Person-Search-Based strategy got a better performance compared with the other two results. It demonstrated that our method can get accurate result for person search, while action search baseline can not reach the accurate representations for actions. Thus, we will also improve the action search baseline in future works.

Acknowledgments

This work was supported by Natural Science Foundation for Distinguished Young Scholars of Shandong Province (JQ201718), Natural Science Foundation of China (U1736122), Shandong Provincial Key Research and Development Plan (2017CXGC1504), Intelligence Advanced Research Projects Activity (IARPA) via Department of Interior/ Interior Business Center (DOI/IBC) contract number D17PC00340.

References

[1] Awad G, Butt A, Curtis K, et al. TRECVID 2019: An evaluation campaign to benchmark Video Activity Detection, Video Captioning and Matching, and Video Search & retrieval. *Proceedings of TRECVID 2019*, 2019.

[2] Awad G, Kraaij W, Over P, et al. Instance search retrospective with focus on TRECVID. *International journal of multimedia information retrieval*, 2017, 6(1): 1-29.

[3] Yu E, Sun J, Li J, et al. Adaptive semi-supervised feature selection for cross-modal retrieval. *IEEE Transactions on Multimedia*, 2018, 21(5): 1276-1288.

[4] Zhang K, Zhang Z, Li Z, et al. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters*, 2016, 23(10): 1499-1503.

[5] Cao Q, Shen L, Xie W, et al. Vggface2: A dataset for recognising faces across pose and age. *IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, 2018: 67-74.

[6] Ren S, He K, Girshick R, et al. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 2015: 91-99.

[7] Lin T Y, Maire M, Belongie S, et al. Microsoft coco: Common objects in context. *European conference on computer vision*, 2014: 740-755.

[8] Wojke N, Bewley A, Paulus D. Simple online and realtime tracking with a deep association metric. *IEEE International Conference on Image Processing (ICIP)*, 2017: 3645-3649.

[9] Carreira J, Zisserman A. Quo vadis, action recognition? a new model and the kinetics dataset. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017: 6299-6308.

[10] Sigurdsson G A, Varol G, Wang X, et al. Hollywood in homes: Crowdsourcing data collection for activity understanding. *European Conference on Computer Vision*, 2016: 510-526.