Shenzhen Institutes of Advanced Technology, CAS, China at TRECVID INS 2016

Jin Ye, Linjie Xing, Xiaolong Fan, Changzhi Song, Diping Song
Cai-Zhi Zhu, Yu Qiao

Shenzhen Institutes of Advanced Technology, CAS, China

Abstract. We divide the task into person retrieval and location retrieval in TRECVID INS 2016, and then fuse the two results together with a simple . About person retrieval, we have two choices. One is based on face recognition. Firstly, deep convolutional networks are used to predict face and detect landmark location on it. Next, we use CNN with center loss to learn face features. Finally, the output features being L2-normalized are simply compared with cosine distance. The other choice is using person re-identification based on CNN. In the query process, we track the target person in the provided query videos so as to expand the query. We extensively compare eight strategies of aggregating multi-image search results in face recognition and person re-identification, and select the best. Location retrieval is based on our improved BOW system adopted in TRECVID 2014.

1 Instance Search

Our Instance Search system is designed following the guideline provided by TRECVID 2016 [1]. For this year's system, we divide the task into two different parts, location retrieval and person retrieval. For location retrieval, we propose similar SIFT-BoW based search framework as in this paper [2]. On the other hand, we use deep Convolutional Neural Network (CNN) framework instead of traditional BOW framework in person retrieval, in which two different choices are conducted, face recognition and person re-identification. We also tried to fuse both choices in one of our submissions. Finally, we combine the results of location retrieval and person retrieval. The whole system is shown in Figure 1.

1.1 Person retrieval

Face feature extraction Face detection is essential in the face recognition framework, as the accuracy of face localization can significantly influence subsequent feature extraction. However, large visual variations existing in face images, such as occlusions, large pose variations and extreme lighting conditions, impose great challenges on this task. This article [3] proposes a new framework to integrate face detection and face alignment using unified cascaded CNNs by multi-task learning. We follow the above-mentioned cascaded framework to generate five facial landmarks and bounding boxes. In training process, we collect

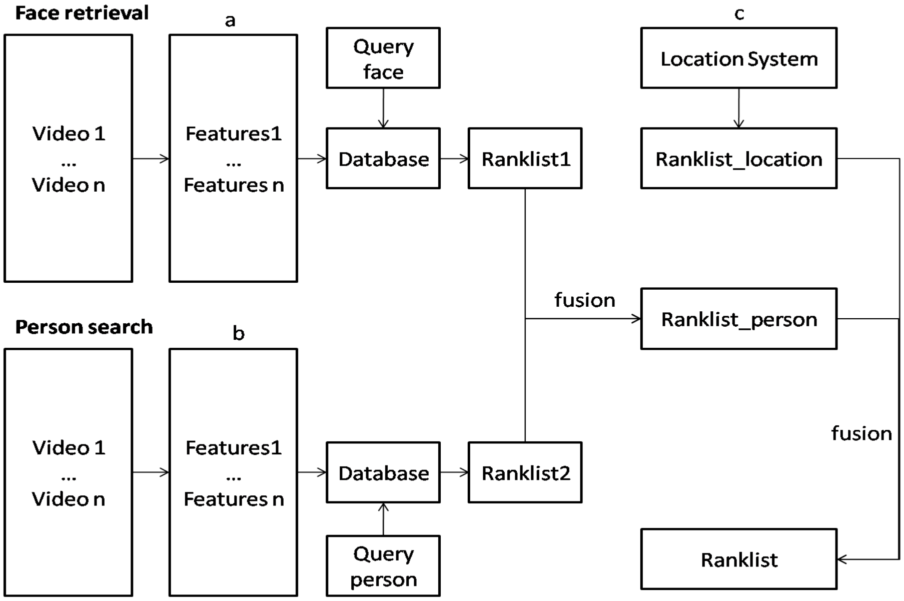

Fig. 1. The whole system

positive and negative patches from WINDER FACE dataset. Then, we crop faces from CelebA dataset as landmark faces.

After we have got facial landmarks and bounding boxes, an efficient feature extraction method must be taken. We use the method proposed in paper [4], in which a new center loss function is used to efficiently enhance the discriminative power of the learned CNN features. We adopt the loss function and train the networks on CELEBRITY+, CAISA-WEBFACE and CACD2000 three datasets.

Person feature extraction In person detection system, we use Faster R-CNN [5] to generate bounding boxes of persons, and train RPN on both PASCAL VOC 2007 and 2012 datasets. We extract person features from the bounding boxes using an end-to-end deep learning framework, and train a CNN network on the dataset as in this paper [6].

Tracking system Same object(face or pedestrian) appearing in a video can be grouped together to form an object track through the tracking system, as illustrated in Figure 2. With the tracking system, we can both achieve query expansion on the query side and correlate the same object appearing on the dataset side, so that the information of a single object can get richer from video sequence. For a query video, the initial bounding box of a query object is given by the corresponding mask image. Instead for a dataset video, the bounding box of the object is initialized by the object detection method [3, 5]. Our tracking module follows the idea of MDNet and FCNT [7, 8]. Specifically, object eigenvector, box overlap rate, box size, and target confidence information among sequential three frames are used for object correlation. The union subsystem combine the results of both the tracking module and the correlation module, if the ratio of the distance between the center of the tracking and detection boxes and the box size is greater than a certain threshold (e.g. >0.2 in our experiment), the object is confirmed lost. A new sequence will then be generated and a new tracker will be reinitialized.

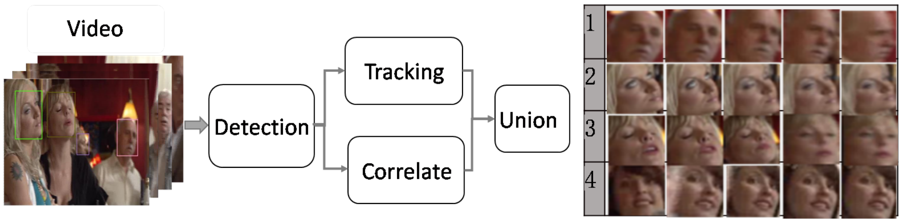

Fig. 2. Left:The Tracking System framework. Right:A sample result

Aggregate strategy In this paragraph, we use the same framework to treat regions of both face and person, collectively called the *object* in short. For each shot, after the tracking process, we can correlate the same object appearing in every frame of a video. We compared eight strategies of how to match the query object appearing in four query images with a database object tracked in multiple frames. Assuming four object features extracted from four query images, for each we compute their cosine distance between each detected object feature in one frame, this way we have two ways to choose the final distance among these four: mean or minimum. This aggregation happens in the feature level. Similarly in frame and shot level, we also have these two aggregation ways. In total we have eight strategies to measure the final distance between four query images and a database video, as shown in Figure 3.

1.2 Location retrieval

For location retrieval, we follow the previous TrecVid INS framework [2] based on multi-image aggregation [9] and practical spatial re-ranking [10], briefly summarized as follows. First SIFT features were extracted from database video frames and query images, then our own implementation of the Hamming Embedding based retrieval method [11] was used to yield initial ranking results, in which multi-image information on both query and database sides was aggregated. In the end, top 200 results were re-ranked by the practical spatial re-ranking method [10] and formed the location based ranking list ready for later person-location fusion.

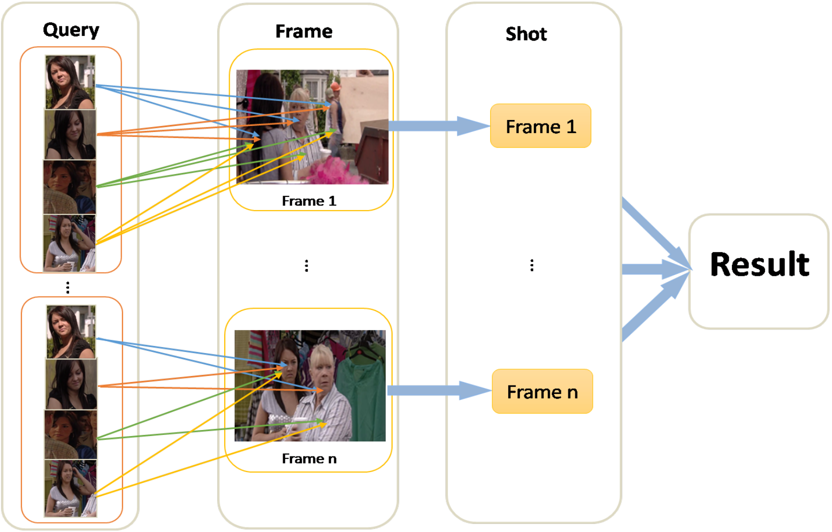

Fig. 3. Eight strategies of aggregating distances between multiple query images and a database video.

1.3 Person and Location fusion

After we get the similarity scores and ranking list based on person and location retrieval, respectively, we fuse them together to get the final list by Equation 1. Basically a Sigmoid function taking the ranking order of location retrieval as input will be used to weight similarity scores of person retrieval, the final score will be returned as the final ranking list.

$$fusion\_score(x) = \frac{person\_score(x)}{1 + e^{0.01*(location\_rank(x)-4000)}} \tag{1}$$

2 Experiment

We applied our approach on the Instance search 2016 task, and experiments were run on GeForce GTX TITAN GPU server. We extracted around 7.8 million keyframes at a frame rate of 5fps from all database shots. As for feature extraction, we got around 9.8 billion SIFT features, 10.8 million face CNN features and 14.7 million person CNN features.

2.1 Evaluation of eight aggregation strategies

We compare the eight aggregation strategies using the above face recognition method on all person related topics of TRECVID 20132015 INS datasets, as shown in Table 1. Here *v1*, *v2* and *v3* stand for three different face recognition

CNN networks, and all networks generate 1024d vectors. We can conclude that the best aggregation way is using mean-min-min in aggregating feature-frame-shot, which discovers a quite surprising conclusion, i.e. the retrieval results do not benefit from any object tracking/correlation method on the database video side! Therefore we give up tracking persons/faces in the final submission. Instead we observe some improvement by tracking the query object on the query video side, as shown in Table 2. It is easy to see that more images get the query object richer. Hereafter we use tracking in the F_E experiments.

Table 1. Result of eight aggregation strategies. 0 and 1 stands for mean and min respectively, i.e.,010 stands for mean-min-mean. *v1*, *v2* and *v3* denotes three different networks we adopt in face feature extraction. *bef* and *aft* mean before and after respectively.

	000		001		010		011		100		101		110		111	
	bef	aft	bef	aft	bef	aft	bef	aft	bef	aft	bef	aft	bef	aft	bef	aft
v1	0.22	0.19	0.25	0.35	0.30	0.18	0.37	0.37	0.21	0.18	0.23	0.36	0.28	0.17	0.36	0.36
v2	0.28	0.26	0.29	0.38	0.35	0.25	0.40	**0.40**	0.28	0.25	0.27	0.30	0.33	0.24	0.34	0.34
v3	0.29	0.27	0.29	0.38	0.35	0.26	0.40	**0.40**	0.28	0.25	0.26	0.30	0.33	0.4	0.37	0.37

Table 2. Evaluate how the object tracking influence retrieval performance on the query side.

Topic num	9084	9088	9092	9096	9104	9115	9116	9119	9124	9138	9143	all
Baseline	0.009	0.166	0.008	0.268	0.006	0.006	0.050	0.000	0.334	0.001	0.336	0.108
With tracking	0.009	0.175	0.032	0.315	0.006	0.000	0.062	0.071	0.297	0.004	0.402	**0.125**

2.2 Instance Search Task Results

We submitted eight runs to the INS task, as shown in Table 3. In our best submission *F_E_4*, we used object tracking on query videos to generate more query regions. We used these expanded queries only in face recognition method, and person re-identification results were excluded. In *F_A_1*, *F_E_A* and *F_E_3*, we fused both person retrieval and face recognition results, but we found that person retrieval actually drops the performance (compared with *F_E_4*). In our submission *F_A_3* and *F_A_4*, we used RPN[5] to detect bounding boxes of target object, and used the BoW framework [2] model to search. Not surprisingly the BoW framework was defeated by other CNN based methods.

3 Conclusion

By participating the instance search task in TRECVID 2016, we have the following conclusions:(1) Effective features are still vital in face recognition and person

Table 3. Final submission results

Run IDs	Mean Average Precision
F_A_1	0.123
F_A_2	0.133
F_A_3	0.028
F_A_4	0.030
F_E_1	0.129
F_E_2	0.043
F_E_3	0.126
F_E_4	**0.141**

re-identification; (2) We have proposed a quick idea to fuse location and person results, but it might not be (or even close to) the best way. We hypothesize the fusion method will influence the final performance largely.

References

1. George Awad, Jonathan Fiscus, Martial Michel, David Joy, Wessel Kraaij, Alan F. Smeaton, Georges Quenot, Maria Eskevich, Robin Aly, and Roeland Ordelman. Trecvid 2016: Evaluating video search, video event detection, localization, and hyperlinking. In *Proceedings of TRECVID 2016*. NIST, USA, 2016.
2. Cai-Zhi Zhu, Yinqiang Zheng, Ichiro Ide, Shinichi Satoh, and Kazuya Takeda. Nagoya university at trecvid 2014: the instance search task. *Participant Notebook Paper of TRECVID*, 2014.
3. Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. Joint face detection and alignment using multi-task cascaded convolutional networks. *CoRR*, abs/1604.02878, 2016.
4. Yandong Wen, Kaipeng Zhang, Zhifeng Li, and Yu Qiao. A discriminative feature learning approach for deep face recognition. In *European Conference on Computer Vision*. Springer, 2016.
5. Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. Faster R-CNN: towards real-time object detection with region proposal networks. *CoRR*, abs/1506.01497, 2015.
6. Tong Xiao, Shuang Li, Bochao Wang, Liang Lin, and Xiaogang Wang. End-to-end deep learning for person search. *CoRR*, abs/1604.01850, 2016.
7. Hyeonseob Nam and Bohyung Han. Learning multi-domain convolutional neural networks for visual tracking. *CoRR*, abs/1510.07945, 2015.
8. Kai Kang and Xiaogang Wang. Fully convolutional neural networks for crowd segmentation. *CoRR*, abs/1411.4464, 2014.
9. Cai-Zhi Zhu, Yu-Hui Huang, and Shin'ichi Satoh. Multi-image aggregation for better visual object retrieval. In *Proceedings of ICASSP*, 2014.
10. Xiao Zhou, Cai-Zhi Zhu, Qiang Zhu, Shin'ichi Satoh, and Yu tang Guo. A practical spatial re-ranking method for instance search from videos. In *Proceedings of ICIP*, 2014.
11. Herve Jegou, Matthijs Douze, and Cordelia Schmid. Hamming embedding and weak geometric consistency for large scale image search. In *Proceedings of ECCV*, 2008.