WHU-NERCMS AT TRECVID2018: INSTANCE SEARCH TASK

Dongshu Xu, Longxiang Jiang, Xiaoyu Chai, Jin Chen, Li Jiao, Jiaqi Li, Shichen Lu, Han Fang, Chao Liang*
National Engineering Research Center for Multimedia Software, School of Computer Science, Wuhan University
cliang@whu.edu.cn

February 28, 2019

ABSTRACT

We participated in the Trecvid Instance Search Task (INS) this year and submitted both automatic and interactive search results. As for automatic task, we retrieval scene and person respectively first and fuse them together later. When retrieving scene, we adopted deep CNN-based method to extract features from keyframes and measured similarity with queries to get scene scores. When retrieving person, we adopted face recognition model based and person re-identification model based method to get person score. In score fusion stage, we exploited weighting based balance and person identity based filter to joint person and scene score. As for interactive task, the initial results are the same as the automatic method, we traversed the top 50 ranks and dropped out the incorrect shots to get interactive results.

1 Introduction

The instance search (INS) is a special content based multimedia retrieval task. Given one or more visual examples of a specific item, which can be a person, an object, or a plane, the aim of the task is to find more video segments of the certain specific item [1]. In Trecvid 2018, the INS task contains automatic search and interactive search. The system task is, given a collection of test videos, a master shot reference, a set of known location/scene example images and videos, and a collection of topics (queries) that delimit a person in some example images and videos, locate for each topic up to the 1000 shots most likely to contain a recognizable instance of the person in one of the known locations[2]. As is shown in Figure 1, it asks to retrieval shots with Jane in cafe2 as many as possible (The metrial is copyrighted by BBC). We participated both automatic and interactive INS task.

For automatic task, we retrieved scene and person respectively with different methods. As for searching scene, we adopted CNN (Convolutional Neural Network) model trained off-the-self to extract global scene features of both probe images and keyframes. With the extracted features, we measured similarity between probes and keyframes to get scene ranking list based on similarity score. As for person, we adopted two methods, which is face recognition based method and person re-identification based method, the former focus on processing keyframes on which we can detect both body and face bbox of persons, the later, besides focus on what the former focus, also committed to process keyframes with detected person's body without face detected. Both methods can generate ranking lists with similarity scores of person. With both scene score and person score, we fusing them together by two fusing strategies to get final automatic results, namely are weight fusing and filter fusing.

For interactive task, we dropped out the false positive items from top 50 ranks of automatic results to get final interactive result.

With the proposed methods, We get 0.243 mAP in automatic task and 0.261 mAP in interactive task according the evaluation, ranks 5 among 8 teams.

*Corresponding author

Figure 1: One of the topics of Trecvid 2018 INS, asked to retrieval *Jane* in *cafe2*

2 Our Framework

The framework we proposed is shown in Figure 2, it consists of 5 modules. The first module is to retrieval scene, based on global feature extracted using CNN, the second and third modules are used to retrieval person, one is based on face recognition and the other is based on person re-identification. With the proposed scene and person retrieval, we get both scene and person retrieval scores. The forth module is to adjust the distribution of scores and the firth module is used to fusing the scores together to get final automatic results. The details of each module and related key technologies are demonstrated as following.

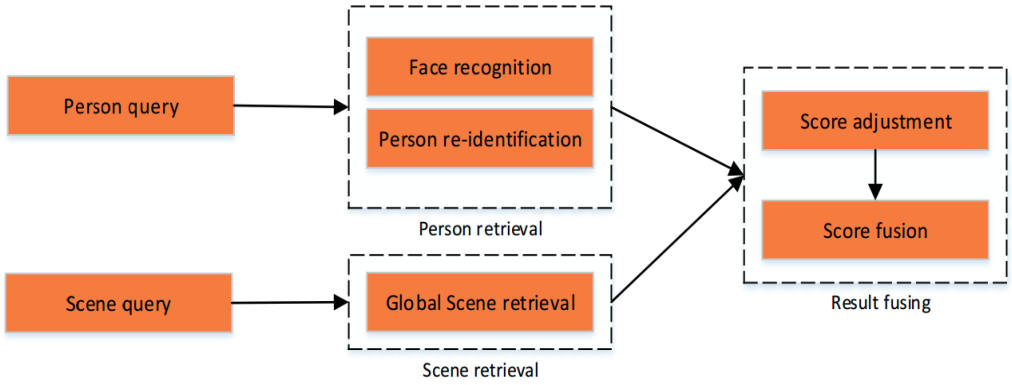


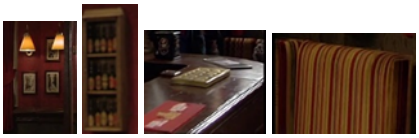
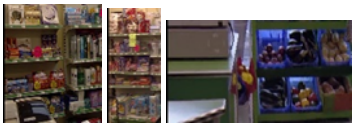
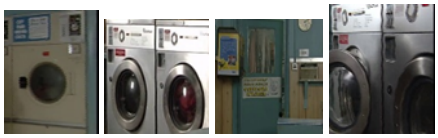
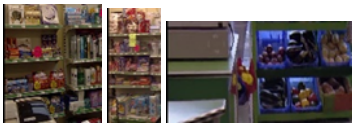
Figure 2: Our framework

2.1 Scene retrieval

Scene retrieval mainly includes the following two parts: 1) Local scene retrieval adopt deep CNN framework for target detection: Single Shot MultiBox Detector(SSD) [3]. We use the pre-trained VGG16 model [4] to initialize the proposed network. For different scenes, we use a certain iconic object to represent a scene. If the category of output is refrigerator, we can roughly judge this scene is the kitchen. The network includes 9 convolutional layers, 5 pooling layers and 2 fully connected layer. 2) Based on Places365 [5], we use the Resnet50 models [6] to extract the global feature. We save the output feature of the fully connected layers and used which we can compute the similarity scores using cosine distance. The network is a residual network with different parameter layers, it mainly replace each 2-layer block in the 34-layer net with this 3-layer bottleneck block.

In the optimization phase, we create the datasets which contain some specific targets to train the pre-trained network. Iconic objects can be seen in table 1.

Table 1: Selected objects for topic scene

Topic scenes	Iconic objects
Pub	
Cafe2	
Laun	
Market	

2.2 Face recognition based person retrieval

Face recognition is a core component of object retrieval in specific scene. Its main function is to determine the identity of the actors or actresses in the TV series through face analysis. We build our own face image library for face recognition as shown in Figure 3.

The quality of face images in the TV series are often affected by different pose, illumination, compression, resolution, etc., so the unconstrained conditions make a huge challenge for face recognition in TV series scenarios. To tackle this problem, we make two steps: First, we use MTCNN proposed in [7] to detect the faces location; Second, we use the Center-loss based method which was published in [8] to build a model for face recognition.

For face detection. Different scenes in the TV series or movies will cause different illumination conditions, poses and scales of target faces, which bring challenges for face detection.

In order to tackle the problems, we adopt MTCNN [7] face detection model, which is trained on a large-scale face detection dataset wilder face [9], which includes a high degree of variability in scale, pose and occlusion as depicted in the sample images. This model trained by the dataset will be more robust to the influence factors, and meanwhile, MTCNN using joint face detection and alignment multitask learning, which not only improve the accuracy of face detection but also can acquire the landmark information, this will be convenient for the subsequent pre-processing for face recognition, such as similarity transformation. Furthermore, MTCNN adopts cascaded network structure, which ensures the detection accuracy and decrease the computation cost, which is good for large scale data processing. In practice, we assume the height of face candidate region less than 60 pixel has high error probability which should be filtered.

For face recognition. The number of face identities in TV series is limited, however the illumination, pose and scale leading to inner-class change, and then effect the accuracy of recognition. In order to decrease the inner-class discrepancy of deep features, we adopt center loss + Softmax cost function proposed by [8] to further decrease inner-class variation. This network architecture using ResNet block to accelerate converge. The input of the network is a 96*112 RGB face image, after going through two convolutional layers it passed into 3 cascaded ResNet blocks, which is the standard ResNet block. The last is fully connected layer, which outputs 512-dim feature vector. During feature representation, we extract the features of original face image and its horizontal flip image, then forming a 1024-dim feature vector by concatenating the features together to represent a face.

In practice, we fine-tune the pre-trained model by using a filtered mixed face dataset, which involved CASIA FACE, YouTube Face, IJB-A and UMDFaces Datasets, to make it more consistent with the data distribution of video face. In order to further improve recognition accuracy, we collect a large number of the actors of the TV series images on the Internet which include different poses, ages, and so on (these images are not appeared in this EastEnders TV

series) as our reference dataset. The identity recognition pipeline is: first, we get the similarity matrix by querying the subjects with every image in the reference dataset, then the identity is confirmed by the maximum similarity score. Due to one identity in the reference dataset contains multiple face images, the postprocessing method can reduce the discrepancy of inner-class for identity feature representation, then improve the accuracy.

Figure 3: Part of our face library

2.3 Person re-identification based person retrieval

Generally speaking, when distinguishing the pedestrians or characters in the TV play, we will choose to identify the face of the pedestrian so as to identify his information. However, due to the occlusion situation, the pedestrian's back picture or the profile picture, etc., we cannot obtain the pedestrian's face, so we will adopt the person re-identification method. We adopt SSD [3] as our person detector, Specifically, we use the aligned-reid framework in [11] to recognize person identity, the framework of AlignedReID is in Figure 4. we generate a single global feature of the input image from the the Network, and adopt the Euclidean distance as the similarity metric. For each image, we use a Resnet50, to extract a feature map, which is the output of the last convolution layer (C*H*W , where C is the channel number and H*W is the spatial size, e.g., 2048*7*7 in Figure 4). Global feature (a C-d vector) is extracted by directly applying global pooling on the feature map [11]. At training stage, we trained a classification network, the base learning rate is 0.0002, and it costs about 5400 seconds per epoch. What's more, at test stage, we can acquire the global features of pedestrians and thus get similarity score. For the 1350 million pictures of training set, we divide them into 244 parts according to video from 0 to 243, extract features separately and get global features, and then, with the extracted features, get the similarity score of person re-identification by normalize the distance range from 0 and 1 and then using 1 minus the distance.

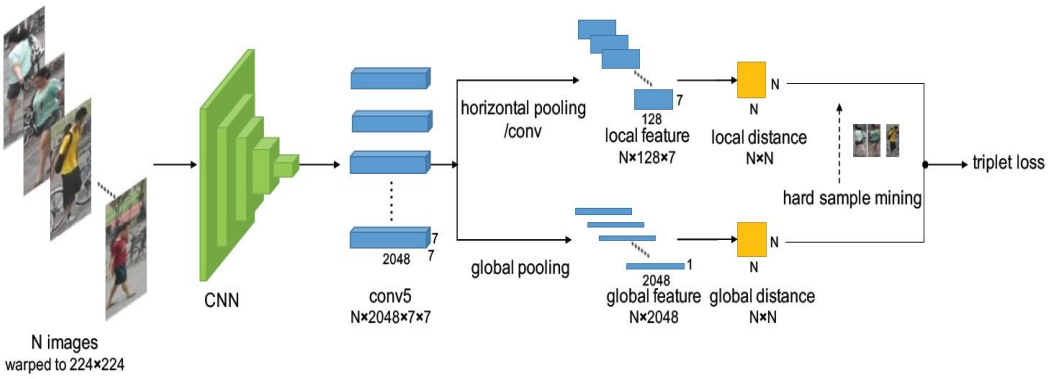


Figure 4: The framework of AlignedReID. Both the global branch and the local branch share the same convolution network to extract the feature map. The global feature is extracted by applying global pooling directly on the feature map. For the local branch, one 1*1 convolution layer is applied after horizontal pooling, which is a global pooling with a horizontal orientation. Triplet hard loss is applied, which selects triplet samples by hard sample mining according to global distances.

Table 2: Fusing Method

Score lists	Fusing method	Method index	Submit index
$f\_scene + f\_face$	weight	A	F_A_A_WHU_NERCMS_1
$f\_scene + f\_face + f\_reid$	weight	B	F_A_A_WHU_NERCMS_2
$f\_scene + f\_face$	filter	C	F_A_A_WHU_NERCMS_3
$f\_scene + f\_face + f\_reid$	filter+expansion	D	F_A_A_WHU_NERCMS_4

2.4 Score fusing

In this section, we will introduce our fusing methods, with which we get our final 4 automatic results. With the scene retrieval module, face recognition based person retrieval module and person re-identification based person retrieval module, we get 3 similarity score lists per topic, let them f_scene, f_face and f_reid respectively. We also propose two different fusing method, one is weight fusing and the other is filter fusing. The submitted 4 results are generated by combining score lists with the fusing methods, which are listed in table2.

Before fusing, we normalize all the score lists range from 0 to 1 by formula (1)

$$f = \frac{f - min(f)}{max(f) - min(f)} \tag{1}$$

For method A, we adopt weight method to fuse f_score and f_face together, the formula in show in(2) and the weight α is set to 0.5.

$$f = \alpha * f\_scene + (1 - \alpha) * f\_face + exp(-|f\_scene - f\_face|^2) \tag{2}$$

For method B, we adopt weight method to fusing f_score, f_face and f_reid together, the formula in show in (3) and the weight α is set to 0.5, β is 0.4 and γ is 0.1.

$$f = \alpha * f\_scene + \beta * f\_face + \gamma * f\_reid + exp(-|f\_scene - w_1 * f\_face - w_2 * f\_reid|^2) \tag{3}$$

Where, $w_1 = \beta/(\beta + \gamma)$ and $w_2 = \gamma/(\beta + \gamma)$.

For method C, we propose a face filter based method. Note that the face library has all the actors appeared in the TV series, the detected face must belong to a certain actor. Thus, firstly, we assign per detected face to a actor according to whose score is largest on it. By this way, given a shot and a target actor, we can conclude whether the actor appeared in the shot. Secondly we can filter out all the shots without target face, and then we can get fusing result by ranking the remained shots according to their f_scene.

For method D, thus person re-identification method can retrieval those shots which contain the target person but without face detected as previous mentioned. In this semantic meaning, the shots exclude by method C may still contain the target person, we mark the shots remained after the first step in method C as set S, and by the same way as the first step in C, we can get the shots filtered by person re-identification score, mark those as set Z, We traverse all the shots in S and find out the shot pairs which are near but not continues in time. For each pair, we resume the shot betwixt the pair member if the shot is not in set Z. With the expanded shots list, we can get fusion result by ranking the shots according to their f_scene.

3 Results and Analysis

Results of our submitted 4 automatic runs on Instance Search task of TRECVID 2018 are shown in table 3 and the interactive runs are in table 4. As we can see, the result of mehtod A is better than others. We can conclude that:

- The face recognition is a key method to identity person, but due to the complex environment in TV series, simply using the detect and recognition method is not sufficient As we can see in result F_A_A_WHU_NERCMS_3, it is based on face filter, there exists many false positives.
- The person re-identification method need modify and may not suitable for retrieval persons in TV series.
- The scene model need fine-tuning with the INS dataset, but how to mine the train items is worth considering.

Also, we get some suggestions and experiences to guide future work:

Table 3: Automatic Result

mAP	Method index	Submit index
0.243	A	F_A_A_WHU_NERCMS_1
0.174	B	F_A_A_WHU_NERCMS_2
0.211	C	F_A_A_WHU_NERCMS_3
0.182	D	F_A_A_WHU_NERCMS_4

Table 4: Interactive Result

mAP	Method index	Submit index
0.261	A	I_A_A_WHU_NERCMS_1
0.184	B	I_A_A_WHU_NERCMS_2
0.235	C	I_A_A_WHU_NERCMS_3
0.200	D	I_A_A_WHU_NERCMS_4

- The model used off-the-self can not suit well in INS dataset, thus need retrain or fine-tuning before making use of, how to train CNNs with unsupervised or weakly supervised way is worth considering.
- The fusing method is important for INS. However there is no better method except weight based method to our best knowledge, how to fuse also requires considering.
- The interactive method is not based on an effective algorithm. How to build a fast and accurate interactive algorithm to improve the results of large scale INS tasks is also waiting to be solved.

Acknowledgement

Thanks for the great support to our work by professor Ruimin Hu, professor Jun Chen and associate professor Chao Liang, who guided us from the begin to the end of the INS task. Thanks for Jiamei Lan in National Engineering Research Center for multimedia Software, School of Computer, Wuhan University, who gives us many ideas of results fusing. we also thank for BBC which provided the video material.

References

[1] Dongshu Xu, Jiamei Lan, Xiaoyu Chai, Yiyue Chen, Xiao Wang, Jiaqi Li, Longxiang Jiang, and Chao Liang. Whu-nercms at trecvid 2017: Instance search task.

[2] George Awad, Asad Butt, Keith Curtis, Yooyoung Lee, Jonathan Fiscus, Afzal Godil, David Joy, Andrew Delgado, Alan F. Smeaton, Yvette Graham, Wessel Kraaij, Georges Quénot, Joao Magalhaes, David Semedo, and Saverio Blasi. Trecvid 2018: Benchmarking video activity detection, video captioning and matching, video storytelling linking and video search. In *Proceedings of TRECVID 2018*. NIST, USA, 2018.

[3] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *European conference on computer vision*, pages 21–37. Springer, 2016.

[4] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[5] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017.

[6] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[7] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters*, 23(10):1499–1503, 2016.

[8] Yandong Wen, Kaipeng Zhang, Zhifeng Li, and Yu Qiao. A discriminative feature learning approach for deep face recognition. In *European Conference on Computer Vision*, pages 499–515. Springer, 2016.

[9] Shuo Yang, Ping Luo, Change Loy Chen, and Xiaoou Tang. Wider face: A face detection benchmark. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 5525–5533, 2016.

[10] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *European conference on computer vision*, pages 21–37. Springer, 2016.

[11] Xuan Zhang, Hao Luo, Xing Fan, Weilai Xiang, Yixiao Sun, Qiqi Xiao, Wei Jiang, Chi Zhang, and Jian Sun. Alignedreid: Surpassing human-level performance in person re-identification. *arXiv preprint arXiv:1711.08184*, 2017.