WHU-NERCMS at TRECVID 2017: Instance Search Task

Dongshu Xu, Jiamei Lan, Xiaoyu Chai, Yiyue Chen,
Xiao Wang, Jiaqi Li, Longxiang Jiang, Chao Liang*
National Engineering Research Center for Multimedia Software, School of Computer Science, Wuhan University
cliang@whu.edu.cn

Abstract

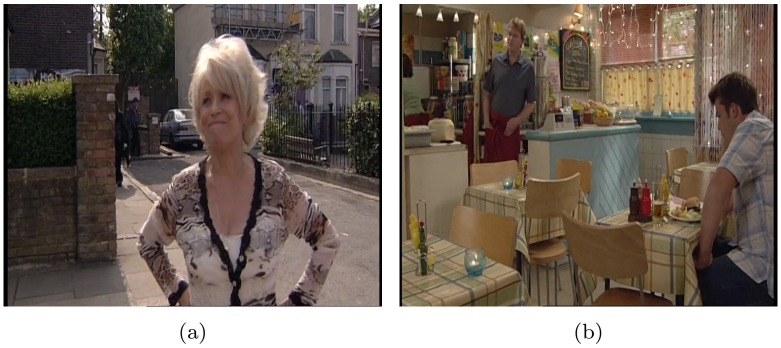
This report introduces our work in instance search (INS) task of TRECVID 2017. The INS task aims at retrieving specific persons in specific locations this year. For this task, we first retrieve person and scene respectively, then combine both results to get the final result. Our system contains four modules, e.g. shots filtering, person retrieval, location retrieval and result fusion. The filter is to delete irrelevant shots. The person retrieval module is to find target persons based on face recognition and speaker identification. The location retrieval module is to identify the target locations by multi-landmark retrieval and several CNN networks. And the result fusion module is to combine the results of person search and scene retrieval to obtain the ranking result at first and then to optimize it to get the final result. We participated in all two types of INS task: automatic search and interactive search. In automatic INS task, we utilized the person retrieval module the same as that in last year but enlarged the face library. Compared with the system in last year, the most different thing is that we adjusted the scene retrieval module and adopted a new expansion method. The automatic results show that they took a better performance. In interactive INS task, we selected effective expansion of person retrieval and scene retrieval ranking result and deleted the obviously wrong shots. The interactive results show that expansions and interactive selection of the ranking results are useful. In our whole system, scene retrieval module is still the weakness due to its CNN part not training on the target.

1 Introduction

The Instance Search (INS) Task in TRECVID [1] is a special content based multimedia retrieval task, which is given a collection of test videos, a master shot reference, a set of known location/scene example videos, and a collection of topics (queries) that delimit a person in some example 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 shown in Fig.1 (Programme material copyrighted by BBC). For the task aims at retrieving specific persons in specific locations, we can search the person and the scene respectively at the first, and then combine the results of them. According to the past participants' results, face recognition is effective to search the person. But in scene retrieval,

*Corresponding author

it's hard to identify the target scene accurately as scene can be occluded and disturbed by other factors when view changes. Previously, On one hand, the BoW [8] (Bag-of-Words) model is proved to have a good effect. However, the BoW model is not robust if there are few objects in some shots especially those simple-decorated ones. On the other hand, using the CNNs (Convolutional Neural Networks) to extract features made an obvious improvement as it is prone to extract robust global features. But the CNNs neglect some key information if they are trained on the images without the objects in the scenes. The BoW model and the CNN model are combined together. Moreover, an expansion method was performed on the ranking results to pull in more corrective shots as the target may be occluded or queried error in sequential frames. In above, we can conclude that: (1) Face recognition is necessary to person retrieval. (2) Scene retrieval is worth exploring. We would fuse BoW model and multiple CNN models to obtain a better result. (3) Query adaptive expansion method can optimize the final result.



(a) (b)

Figure 1: Topic 9189, a task to find shots with the person Peggy (a) in location Cafe1 (b)

For INS task,there are three problems attracting our attention:

- In person retrieval, faces may be similar from different persons, which causes similarity score nearby in different persons.
- In scene retrieval, global feature extraction methods are hard to cope with similar background especially in the simple-decorated scene such as CNN based method. Moreover, local feature extraction methods are not effective to deal with the changed views.
- Some sequential targets are lost for the occlusion of the target, changed views or other reasons.

To solve the problems mentioned above, we adopted some methods as follows:

- In person retrieval, we built the face library with more than ten face images of each actor in *Eastenders*.
- In scene retrieval, we adopted SSD [5] network to detect objects automatically and selected the ones that have the most density of SIFT points. Then BoW model was used to extract features from the key objects. What's more, global features were extracted by five CNN models named *Inception-Resnet-v2* [4], *Inception-v4* [9], *ResNet-v1-152* [3], *VGG16* and *VGG19* [7].

- In the end, the query adaptive expansion method was adopted to adjust the result score close to its nearby shots.

2 Our Framework

2.1 Automatic Instance Search

Our framework contains four parts as shown in Fig.2. The first is shot filter module, which removed the most irrelevant shots from gallery. The second is human retrieval module, in which face recognition is the main part with speaker identification as supplement. The third is scene retrieval module including multi-objects retrieval and global scene retrieval. In multi-objects retrieval, we used the SSD to detect objects and the BoW model to get a local feature based ranking list. In global scene retrieval, five CNN models were performed to extract features which would lead to the global feature based ranking list. The final part is result optimization module which combined all the ranking list and expanded them to get the final result. The details are as follows:

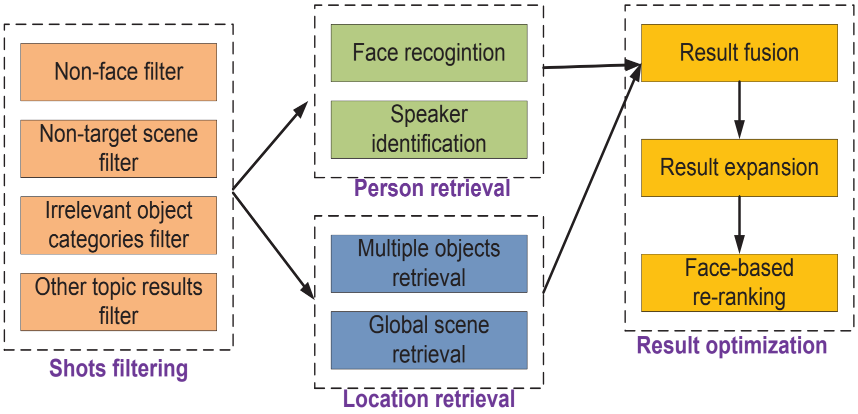


Figure 2: The framework we used this year

Face recognition. Face recognition is core of person retrieval. It consists of two steps: Firstly, we detected where the face locates by a Scale-Adaptive Deconvolutional Regression (SADR) network based on Faster R-CNN. Secondly, A Deep Embedding Network was utilized to conduct face identification after face detection and alignment with 78 landmarks. This network includes 9 convolutional layers, 5 pooling layers and 2 fully connected layers. The network was trained on our collected IVA-WebFace with 80 thousand identities and each has about 500-800 face images [10].

To match the target faces, a face library is needed. We built the library by collecting the images of all the actors appeared in the TV series *Easterners* including all the target faces and non-target faces, as shown in Fig.3 (Programme material copyrighted by BBC). There are 192 actors in the series, each of which has about twelve pictures. Finally, the library has 2664 images. Once one of the actor's faces is recognized, we deem the actor appeared in the shot.

Multiple objects retrieval. There are different objects in different scenes. In general, some typical objects only appear in specific scenes. For example, washing machine only exists in laundry and kitchenware only in kitchen. We can identify scenes by landmarks retrieval. BoW model is utilized to conduct object search. Firstly, we detected the objects automatically with SSD network, applied hessian affine-SIFT feature to describe each detected object. Secondly, we calculated the Root-SIFT and its density to select the typical objects. Thirdly, we trained the 1 million dimensional codebook by adopting Approximate K-means (AKM) algorithm. Fourthly, we quantized each 128 dimensional SIFT features into one of the codes ranging from 1 to 1000000. Lastly, hard assignment method is used to quantize SIFT features of keyframes and soft assignment method is exploited to quantize SIFT features of query images. After getting the feature of both keyframes and query images, we adopted Query Adaptive Similarity Measure to calculate the similarity between keyframes and query images [10].

Global scene retrieval. The output of fully connected layers of a CNN can be regarded as global features. Five CNN models were adopted to extract features from the keyframes [11], which are *Inception-Resnet-v2*, *Inception-V4*, *ResNet-v1-152*, *VGG16* and *VGG19*. Specifically, a shot contains several keyframes. We extracted features of each keyframe and used L2 distance to measure the similarity score and got the maximum score of the keyframes in each shot.

Non-target face filter. In fact, there are tens of thousands of shots without target faces, bringing a lot of noises to scene retrieval. Based on the high accuracy of face recognition, we can easily find out the shots only with non-target faces and delete them from the gallery.

Non-target scene filter. We can also filter non-target shots with the scores of CNN models. If a shot has high similarity with irrelevant objects, it will be filtered out. In fact, a shot can be moved out directly in the non-target scenes, such as *cafe2, foyer, kitchen1, living room1, pub*. For example, when retrieving the scene *cafe1*, all the shots in other target scenes are filtered out, such as *living room2, kitchen2, laun, market.*

Irrelevant object categories filter. From the given topics, we found that irrelevant shots contain some typical objects which could not appear in relevant shots, such as bicycle, bus and tree. We used five CNN models to classify the objects detected by SSD network, integrated the results together and removed the irrelevant shots according to the classification result. In this way, we deleted most irrelevant shots.

Speaker identification. Besides images, audio information can also help us find shots. Speaker identification module is a supplement to face recognition. We used a typical model to identify the specific person. Firstly, we extracted all the audio and filtered out the noisy parts through a low pass filter. Secondl y, the audio corresponding to each shot was segmented and the BIC hop point detection method was used to subdivide the audio. Thirdly, we extracted 39 dimensions of MFCC features. Fourthly, The GMM-UBM [6] method was used to establish the background sound model

Figure 3: Part of our face library

and the role model. Lastly, we calculated the similarity of the speech model with the query speech, and then arranged the similarity score from high to low to get the final speech ranking result.

Score adjustment. There are some keyframes in which the characters make up a large area, so the extracted features can not represent the characteristics of the shots. When retrieving the scenes, the scores of those shots are lower than the real ones. What's more, the keyframes around the one with high scores should also have higher scores as sequential keyframes probably involve the same person or scene. In order to overcome the problem, we arranged all the keyframes in the sequence according to their index, selected the keyframes which have high scores and then adjusted the score of nearby keyframes higher than before.

Result fusion. We combined shot scores accquired in face recognition, BoW model and five CNN models to generate final score results. Specifically, we got three vectors whose values are from 0 to 1 representing the score of shots, and then assigned weights according to aforementioned filter. For example, we set the deleted shots score to 0. Firstly we multiplied the face vectors and the CNN models vectors one by one to get the first result. Then, we mutiplied the BoW vectors and the first result vectors one by one to get the second result. Above all, we got two final ranking lists.

Result expansion According to the similarity of adjacent shots, we adjusted the score of ranking list as well. Firstly, we took a threshold for sorting score. If the score lower than the threshold, set the score to 0. Secondly, based on the new score we did gaussian shape expansion over 8 keyframes, and took the highest one as the current shot score. In this way, we got another two final ranking lists.

2.2 Interactive Instance Search

In order to exert judgement on the expansion of feature extraction models, we first expanded the ranking results of face recognition and CNN based scene retrieval respectively. Then each preferable ranking list was selected, in which the obvious errors ahead would be deleted interactively. In our opinion, this is a process of feature selection and result reranking.

3 Results and Analysis

3.1 Automatic Instance Search

The automatic INS task in TRECVID 2017 has also 30 topics this year, searching 8 persons in 5 locations. Based on our last year's work, we have the following improvement: (1) The scene retrieval method based on the landmarks is automatic this year, such as SSD, an automatic object selection method, marking out the key objects in the scenes. (2) The global scene features are extracted by more CNN networks and combined to be more representative. (3)The result expansion strategy is proposed to renew the result scores. Hence, we designed 4 strategies on automatic INS task, which adopted different modules. 4 runs corresponding to the strategies were submitted, which are listed as Tab.1. We can conclude that the modules are effective in the task. In particular, the result expansion strategy can improve the result about 5% as it can pull in many ground truth near the query answers, and the fusion of BoW model and CNN models can bring about 2% improvement by their compliment in feature extraction mechanism.

No.	Method	MAP
1	CNNs + Face + Filters	0.147
2	CNNs + Face + Filters + Expansion	0.191
5	CNNs + BoW + Face + Filters	0.167
6	CNNs + BoW + Face + Filters + Expansion	0.214

Table 1: Results of our automatic INS runs

3.2 Interactive Instance Search

The interactive INS task in TRECVID 2017 has 20 topics this year, chosen from the 30 topics in automatic task. Based on the automatic strategies, 4 runs on interactive INS task are submitted including fusion selection and expansion on different ranking lists, as shown in Tab.2. indicating that expansions on both face recognition ranking list and scene retrieval ranking lists, and interactive selection are effective.

No.	Method	MAP
3	Expanded CNN + BoW + Face + Feature selection	0.165
4	CNN + BoW + Expanded face + Feature selection	0.172
7	Expanded CNN + BoW + Expanded face + Feature selection	0.217
8	Expanded CNN + BoW + Expanded face + Feature selection + Reranking	0.262

Table 2: Results of our interactive INS runs

In total, we selected many modules to construct our framework. The filter module kicked out nearly a half shots to alleviate the computing complexity. The face recognition module could find out the target persons with a high accuracy. But the speaker identification module only found few and was time consuming. Despite of the difficulty of scene retrieval, two ways were adopted to improve the results. On one hand, we combine the SSD and BoW model to search the scene based on landmarks. On the other hand, we fuse multiple CNN models to make the global feature representative. A result expansion method performed on both face recognition and CNN search ranking lists took a good effect on the result. Moreover, the interactive feature selection gave some improvement on the final result.Although the optimization module can lift up the final mAP value, the module on scene retrieval is the weakness of the framework.

According to the analysis above, we will introduce the feature extraction based on landmarks to a CNN model in a follow-up experiment and train it on the standard scene datasets. Metric learning method is also worth adopting instead of the traditional L2 distance measurement. What's more, the interactive strategy on the whole score distribution especially the hard samples should be used as the final insurance.

Acknowledgement. We express our sincere appreciation to professor Ruimin Hu and professor Jun Chen who give us a lot of ideas and guidance. Thank the DaQian Information co.,Ltd for providing us the source of face recognition. Thank BBC for providing programme material. Thank all the former INS members who accumulated a great deal of experiences for us.

References

[1] George Awad, Asad Butt, Jonathan Fiscus, David Joy, Andrew Delgado, Martial Michel, Alan F. Smeaton, Yvette Graham, Wessel Kraaij, Georges Qunot, Maria Eskevich, Roeland Ordelman, Gareth J. F. Jones, and Benoit Huet. Trecvid 2017: Evaluating ad-hoc and instance video search, events detection, video captioning and hyperlinking. In *Proceedings of TRECVID 2017*. NIST, USA, 2017.

[2] George Awad, Wessel Kraaij, Paul Over, and Shinichi Satoh. Instance search retrospective with focus on trecvid. *International Journal of Multimedia Information Retrieval*, 6(1):1–29, 2017.

[3] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Computer Vision and Pattern Recognition*, pages 770–778, 2016.

[4] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, pages 448–456, 2015.

[5] 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, 2016.

[6] Jack Mclaughlin, Douglas A. Reynolds, and Terry P. Gleason. A study of computation speedups of the gmm-ubm speaker recognition system. In *European Conference on Speech Communication and Technology, Eurospeech 1999, Budapest, Hungary, September*, 1999.

[7] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *Computer Science*, 2014.

[8] Josef Sivic and Andrew Zisserman. Video google: A text retrieval approach to object matching in videos. In *IEEE International Conference on Computer Vision*, page 1470, 2003.

[9] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alex Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning. 2016.

[10] Zheng Wang, Yang Yang, Shuosen Guan, Chenxia Han, Jiamei Lan, Rui Shao, Jinqiao Wang, and Chao Liang. Whu-nercms at trecvid2016: Instance search task. 2016.

[11] Mang Ye, Bingyue Huang, Lei Yao, Jian Qin, Jian Guan, Xiao Wang, Bo Luo, Zheng Wang, Dongjing Liu, and Zhuosheng Zhang. Whu-nercms at trecvid2014:instance search task. 2014.