WHU-NERCMS AT TRECVID2020: INSTANCE SEARCH TASK

Jingyao Yang[1], Kang'an Chen[1], Yanrui Niu[1], Xinyao Fan[1], Chao Liang[*]

[1]National Engineering Research Center for Multimedia Software, School of Computer Science, Wuhan University

cliang@whu.edu.cn

ABSTRACT

In the TRECVID 2020 Instance Search task, the participants are required to retrieve specific persons doing specific actions from 468,533 shots. The retrieval target is divided into two parts: person retrieval and action retrieval. As for person retrieval, a face detection and recognition model pre-trained on wilder face dataset is adopted to compute person retrieval scores. As for action retrieval, we utilize a common action recognition model(TSM) pre-trained on kinetics dataset and a human-and-object-interaction model(PPDM) pre-trained on HICO-DET dataset to compute action retrieval scores. We obtain the final results by fusing person scores and action scores. The approaches of person retrieval in two runs we submitted are the same, however there is something different in action retrieval:

- F_M_E_A_WHU_NERCMS.20_2: We adopt PPDM for some specific actions and C3D that we used last year for the others.
- F_M_E_A_WHU_NERCMS.20_1: We replace C3D with TSM.

The MAP increases by a percentage of 5.2 after C3D is replaced, which means that TSM works much better than C3D in action retrieval. Our best result achieved 0.151 mAP in automatic task according to the evaluation.

1 Introduction

The TRECVID 2020 [1] Instance Search task [2] is retrieving shots of a specific person doing a specific action from 468,533 video clips according to a set of person-action examples called as a topic. Finally, the top 1000 shots that best fit the topic should be given. As is shown in figure 1, the system is asked to retrieve shots with Stacey holding phone.

Person and action are retrieved separately, which will be fused at last. As for person retrieval, a face recognition model is adopted to get person scores. As for action retrieval, we utilize a common action recognition model(TSM) and

[*]Corresponding author

Stacey

Holding phone

Stacey is Holding phone

Figure 1: An example for retrieval (programme material copyrighted by BBC)

a human-and-object-interaction model(PPDM) to get action scores. After that, we exploited two fusion strategies for the combination of person retrieval and action retrieval: weighting based balance and person identity based filter. A ranking list of shots will be generated after fusion operation. Independent retrieval of people and actions makes the retrieval task clearer and easier, so that we can complete complex task in a simple way.

2 Our Framework

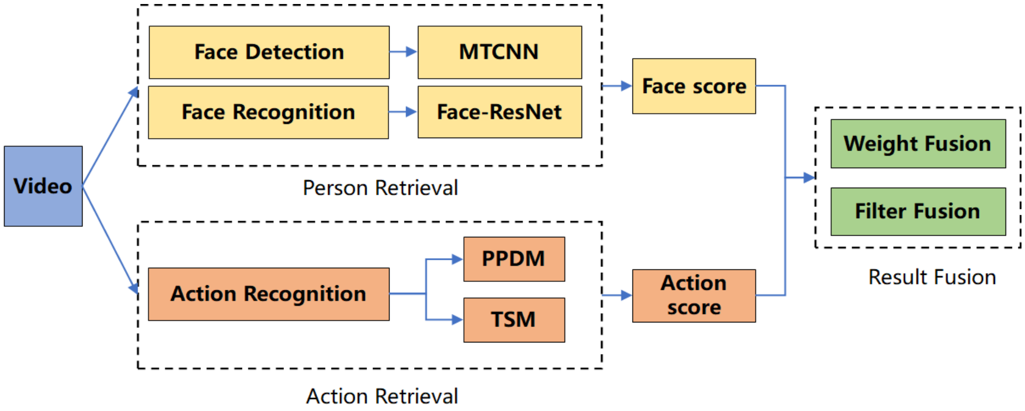

Figure 2: Our framework

The proposed framework of our scheme contains three parts as shown in Figure 2. The first part is person retrieval module, which is based on face detection and face recognition. We choose MTCNN as the framework for face detection and select Face-ResNet as the backbone network for face recognition. The second part is action retrieval module, which is based on TSM and PPDM. With the two modules, we get both person and action retrieval scores. The third is result fusion module, we fuse the scores together to get the ranking result. The details of each module and related key technologies are demonstrated as following.

2.1 Person retrieval

The different postures, illumination conditions, scales and other factors often affect the quality of face images in movies and TV plays, which poses a huge challenge to face recognition in movies and TV plays. To solve this problem, we do two steps in the person retrieval module. Firstly, we use the MTCNN model proposed in [3] to detect faces. Secondly, we use the method based on Center-loss proposed in [4] to build a face recognition model.

To match the target faces, We collected a large number of images of actors in TV series from the Internet as our reference dataset, which are not appeared in this EastEnders TV series.

For the face detection. We use MTCNN [3] face detection model trained on large-scale face detection dataset wilder face [5]. The model has several advantages. Firstly, the dataset contains the high variability of scale, posture and occlusion in the sample images, so it has strong robustness to the influencing factors. Secondly, MTCNN adopts cascade network structure, which reduces the computational load and ensures the detection accuracy, so it is conducive to large-scale data processing. Furthermore, MTCNN uses joint face detection and alignment multitask learning, which improves the accuracy of face detection.

For the face recognition. In order to decrease the internal-class discrepancy of deep facial features, we use the center loss + Softmax cost function proposed in [4] to build a face recognition model. The network has two convolution layers, three cascaded ResNet blocks, followed by a full connection layer which outputs a 512-dimensional feature vector. In the process of feature representation, firstly the features of the original face image and its horizontal flip image are extracted, and then the features are connected together to form 1024-dimensional feature vectors to represent the face.

With the above parts, the similarity matrix is obtained by querying each image in the reference data set first, and then the identity is determined according to the maximum similarity. Since an identity in a reference dataset contains multiple face images, we decrease the discrepancy of inner-class for the identity representation based post-processing method.

2.2 Action retrieval

As for action retrieval, we exploited two methods: TSM and PPDM. TSM is aimed at videos, while PPDM is aimed at frames.

2.2.1 TSM

Some actions are not relevant to any objects such as laughing and crying. And some actions can not be recognized by a single frame such as open_door_in and open_door_out. To solve the problems above, we used TSM [6] (short for 'Temporal Shift Module') to learn spatiotemporal features. TSM is a Convolutional network with the cost of 2D CNN and the ability to well model the temporal information. TSM defines an operation named 'shift': shifting, along the temporal dimension, 1/8 part of the channels by -1, and another 1/8 part by +1, leaving the rest un-shifted. The backbone of TSM is Resnet50. Given a video, 2D CNN sampling several frames and then process each of the frames individually and give the final prediction. As for TSM network, the frames are running just like the 2D CNNs during the inference of convolution layers. The difference is that 'shift' is inserted for each residual block, which enables temporal information fusion at no computation. To extract TSM feature, we used the Resnet50 model which is pre-trained on kinetics dataset [7] to initialize the network. Then the keyframes extracted were delivered to the TSM network and we got a sort list of probability score. The procedure shows as figure3

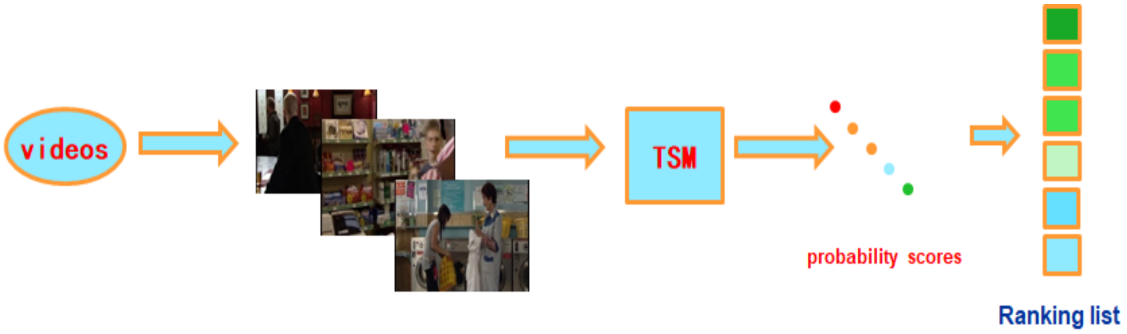

Figure 3: TSM experiment framework (programme material copyrighted by BBC)

2.2.2 PPDM

The interaction between people and objects is quite common in reality, and topics to be retrieved also include this type of action. Different from common action recognition, human and object interaction recognition aimed to recognize the actions with obvious objects, such as "holding paper" and "sit on couch". In this part, we adopted PPDM [8] , the state of the art human and object interaction model, to recognition actions. PPDM model is a real-time human-object interaction detection framework based on frames. The network mainly uses a common heat map prediction networks DLA-34[9] as the feature extractor. The framework is presented in figure 4.

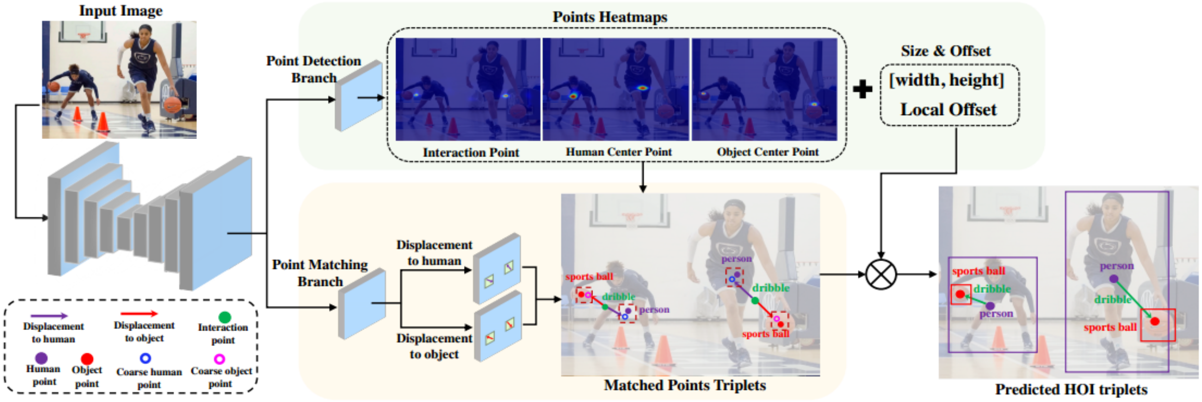

Figure 4: PPDM Framework [8]

During the retrieval, we used PPDM which is pre-trained on HICO-DET dataset [10] to initialize the network and split each shot into keyframes to preprocess the dataset. As shown in figure 5, we tried two different experiments for the recognition of actions. One method is classifying actions of the gallery keyframes directly. The other method is extracting features of frames and then computing feature similarity between gallery frames and probe frames. Corresponding to the first method, keyframes are passed to the network to get output of the classify layers. Corresponding to the second method, keyframes are passed to the network to get output features before the final classify layers, we use these features to compute the action similarity scores based cosine distance. After obtaining the prediction scores of shot keyframes, we took the maximal score of the keyframes in each shot as the prediction score of this shot for each category.

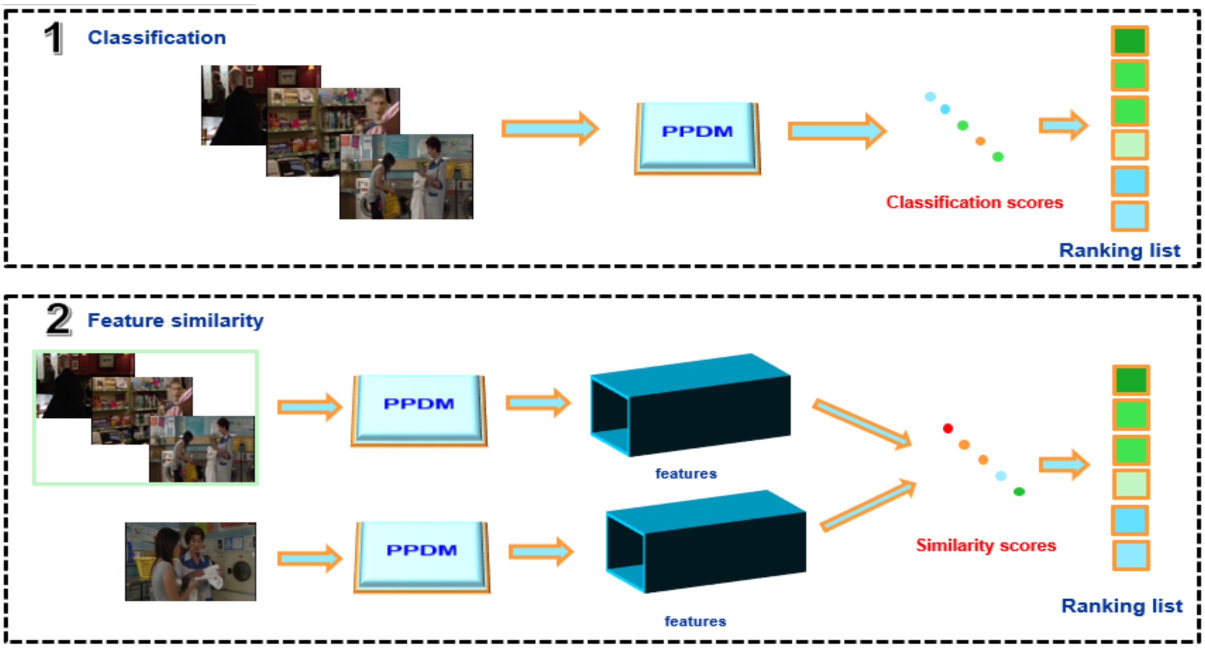

Figure 5: PPDM experiment framework (programme material copyrighted by BBC)

2.2.3 Result fusion

In result fusing module, we propose two different fusing methods. One is weight fusing and the other is filter fusing. With the person retrieval and action retrieval, we get two similarity score list for every topic, let them f_face and f_action respectively. We exploited weighting based and person identity filter based method to combine the results to get the ranking result. Our submitted two results are generated by combining score lists with the above two fusing methods.

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 fuse f_face and f_action together using weight fusion, the formula is shown in 2 and the weight α is set to 0.75.

$$f = \alpha \times f\_face + (1 - \alpha) \times f\_action \tag{2}$$

For method B, we adopt face filter method to fuse the scores. Note that the face library has all the actors appeared in the TV series, the detected face must belong to a certain actor. According to the largest score, we assign detected face to a actor. However, the face is not always recognized correctly due to the complex situation in the video, thus we refined the actor ID of the face which tack smooth, that is to say, we combine the bounding boxes together which have large Intersection over Union and have large possibility to be the same person, and then we assigned actor ID to all the face boxes. Thus, given a shot and a target actor, we can conclude whether the actor appeared in the shot. Based on above, we filter out all the shots without target face, and then according to the f_action, we rank the remained shots to get fusing result.

3 Results and Analysis

The final results of our submitted runs on Instance Search task of TRECVID 2020 are shown in Table 1. Through the results, our analysis and summary are as follows:

- As shown in table 1, we adopt PPDM for some specific actions and C3D that we used last year for the others in runid of F_M_E_A_WHU_NERCMS.20_2. However, in runid of F_M_E_A_WHU_NERCMS.20_1, we replace C3D with TSM and the MAP increases by a percentage of 5.2. It means that TSM works much better than C3D in action retrieval.
- We found that some retrieval results of specific topic have poor performance because of inaccurate face recognition, and the choice of person based filter fusion method may exacerbate this kind of errors. So the improvement of face recognition algorithm and selection of fusion method are worthy of serious consideration.
- The lack of fine-tuning for TSM in actions asked for retrieval may limit its performance. Doing fine-tuning work is our next consideration.

Table 1: Automatic Result

Runid	ptype	exampleSet	method	mAP
F_M_E_A_WHU_NERCMS.20_1	F	E	PPDM+TSM	0.151 [11]
F_M_E_A_WHU_NERCMS.20_2	F	E	PPDM+C3D	0.099 [11]

References

[1] George Awad, Asad A. Butt, Keith Curtis, Yooyoung Lee, Jonathan Fiscus, Afzal Godil, Andrew Delgado, Jesse Zhang, Eliot Godard, Lukas Diduch, Jeffrey Liu, Alan F. Smeaton, Yvette Graham, Gareth J. F. Jones, Wessel Kraaij, and Georges Quénot. Trecvid 2020: comprehensive campaign for evaluating video retrieval tasks across multiple application domains. In *Proceedings of TRECVID 2020*. NIST, USA, 2020.

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

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

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

[6] Ji Lin, Chuang Gan, and Song Han. Tsm: Temporal shift module for efficient video understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

[7] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, and Andrew Zisserman. The kinetics human action video dataset. 2017.

[8] Yue Liao, Si Liu, Fei Wang, Yanjie Chen, and Jiashi Feng. Ppdm: Parallel point detection and matching for real-time human-object interaction detection. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[9] Fisher Yu, Dequan Wang, Evan Shelhamer, and Trevor Darrell. Deep layer aggregation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[10] Y. Chao, Y. Liu, X. Liu, H. Zeng, and J. Deng. Learning to detect human-object interactions. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 381–389, 2018.

[11] Alan F. Smeaton, Paul Over, and Wessel Kraaij. Evaluation campaigns and trecvid. In *MIR '06: Proceedings of the 8th ACM International Workshop on Multimedia Information Retrieval*, pages 321–330, New York, NY, USA, 2006. ACM Press.