BUPT-MCPRL at TRECVID 2020: INS*

Qi Zhang, Jiacheng Zhang, Zhicheng Zhao, Yanyun Zhao, Fei Su

Multimedia Communication and Pattern Recognition Labs,
Beijing Key Laboratory of Network System and Network Culture,
Beijing University of Posts and Telecommunications, Beijing 100876, China
{zhaozc, zyy, sufei}@bupt.edu.cn

Abstract

In this paper, we describe BUPT-MCPRL Instance Search (INS) algorithm and evaluation results at TRECVID 2020. Only visual information is used. Specifically, person retrieval includes face recognition and person tracking, and three kinds of action retrieval methods, i.e., emotion-related actions detection, human-object interactions identification and general actions recognition are implemented. We submit four runs for automatic INS, a brief description is as follows:

- **F_M_A_E_BUPT_MCPRL_1:** Extract action features by ECOfull only
- **F_P_A_E_BUPT_MCPRL_2:** Extract action features by ECOfull only
- **F_M_A_E_BUPT_MCPRL_3:** Extract action features by ECOfull and STGCN, Re-ranking actions with QE
- **F_P_A_E_BUPT_MCPRL_4:** Extract action features by ECOfull and STGCN, Re-ranking actions with QE

The final results are summarized in Table 1. More details will be given in the following sections.

Table 1. Results for each run

Run ID	mAP
F_M_A_E_BUPT_MCPRL_1	13.5
F_P_A_E_BUPT_MCPRL_2	13.0
F_M_A_E_BUPT_MCPRL_3	14.2
F_P_A_E_BUPT_MCPRL_4	12.7

1. Method

To balance the efficiency and effectiveness on such a large dataset, we extract video key frames with a sample rate of 5 fps for every shot. To retrieve specific persons doing specific action, we consider how to achieve satisfying results in person retrieval and action recognition respectively and fuse them in an appropriate manner.

1.1 Persons retrieval

For persons retrieval, the goal is to find all the video shots that contain the queried person and track its position in all frames within the shot. To achieve this goal, it will first use face recognition to find the target person, and then track its position in all the frame of the video shot using object tracking algorithms.

*This work is supported by Chinese National Natural Science Foundation (U1931202, 62076033).

1.1.1 Face recognition

To do the face recognition, we take the following steps: (1) Detect faces at every 5 frames in shots. (2) Describe faces being detected with designed features. (3) Select metric method to distinguish different faces.

For the first step, we adopt the MTCNN method[1] which has proved its accuracy and efficiency in face detection. For the second step, we extract a 128-dim vector representing the face feature for every detected faces with the help of dlib library[2]. Finally, we conduct cosine distance between the target person's features and detected person's feature. We can get five rank lists for a target person because each target person has five example images. We employ adaptive fusion scheme[3] to the five rank lists for each target person. This method make progress obviously by our own observation(Figure 1).

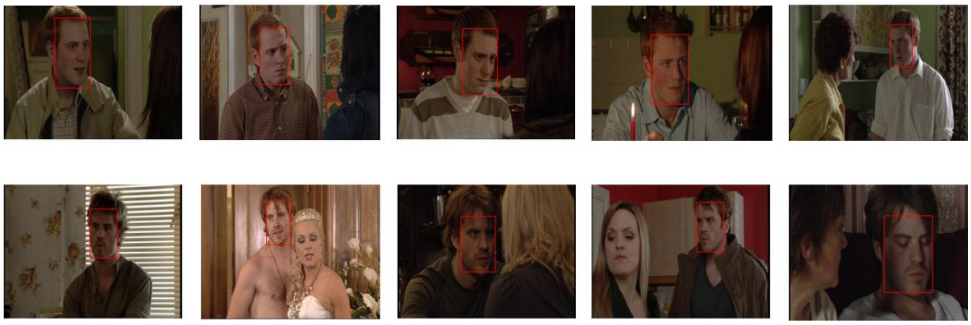

Figure 1. Face recognition results display. We select Bradley and Sean as an example. From the left column to the right column are the 1st, 1000-th, 3000-th, 5000-th and 10000-th retrieval results respectively.

1.1.2 Person tracking

Since the face recognition method is based on images, if we can't see the target person's face clearly (for example, when one turns his back to the camera), this frame may not be detected. This may cause us to miss many correct results. So we adopted person tracking to address this problem(Figure 2).

For every person in each shot, the first frame where the target person appears is used as the initialization frame, and the object tracker is initialized according to the bounding box of the person. The bounding box can be obtained by extending the detected face box with a specific ratio. ResNet50-SiamRPN is chosen as the object tracking model. In this way, we can get the location sequences of all the target persons in the shot, which is used for subsequent feature extraction and action recognition.

1.2 Instance Retrieval

For instance retrieval, we divide instances into three categories: emotion-related actions, human-object interactions and general actions.

1.2.1 Emotion-related action retrieval

We group crying, laughing and shouting together as the first category. We use emotion

recognition to retrieve scenes in which someone is sad, happy or angry. We train emotion recognition models based on Region Attention Network[4] while taking FERPlus[5] and CK+[6] datasets as main training set. Data augment is also performed, including horizontal flip and random cropping.

1.2.2 Human-object interactions retrieval

To retrieve human-object interactions, we explore the dependences between semantic objects and human keypoints by using object detection and pose estimation models.

For the object detection, we choose YOLOv4 trained on COCO to detect key objects such as couch, phone, person etc. However, for the detection of objects that not in COCO (such as cigarettes), we used few-shot detection [7] due to the lack of data. For each class, we downloaded and annotated 30 images as a training set.

For the pose estimation, we feed the sequence of human bounding boxes gained from person tracking into HRNet to estimate human poses (Figure 2). Then, we simply calculate the distance between object location and target person's interactive keypoint to measure the dependences of object-pose such as 'holding_glass', 'holding_phone' and 'sit_on_couch'. Based on the strength of the dependences, we divide the initial ranklist into several groups.

Figure 2. The person retrieval and pose estimation result of a sequence of key frames from one shot. In this shot, heather is opening the door and entering the room. As we can see, in the first four frames, heather's face can not be seen. Besides, her face in the 5th frame is not clear enough. Only the face in the 6th frame can be detected with a high score. To get enough information to recognize this action, we must backtrack heather's position in the first 5 frames using the detected box in the 6th frame.

1.2.3 General actions retrieval

Other instances involve human-human interactions and complex actions, such as kissing, hugging and go up-down stairs, could be easily recognized by action recognition models.

In experiments, we choose two different models to extract the feature for action recognition. The first one is image-based method named ECOfull, it can extract action features directly from the raw video. However, in the daily life scene, there will be a lot of information irrelevant to the action of persons, which will have certain interference to the prediction effect of the model. Therefore, we introduced STGCN to extract features only base on the human pose. Since we have estimated the human pose using HRNet, this approach is easy to implement. In practice, both models are trained on the Kinetics-600 dataset.

Given the action to be retrieved, if similar actions are included in the Kinetics-600 dataset, the final output probability is directly used for ranking and retrieval; Otherwise, the cosine distance

between the probability feature output by the penultimate layer of the model and the probability feature of the example video is calculated, and ranking and retrieval are performed according to the cosine distance. The long clips are divided into multiple short shots. For the retrieval results of multiple shots, we take the most similar retrieval result as the final result. There are multiple example videos for the same action, and each example video can be calculated to obtain a retrieval list, which are weighted to obtain the final retrieval list. In practice, due to the small number of example videos, we use query expansion to simulate more example videos in order to obtain a higher recall rate. Specifically, we select the ten most similar shots retrieved for each queried action, and then calculate the probability features of these shots and take the average value as a new query.

In addition, in order to improve the final retrieval result, the retrieval results of all the actions in the first two categories described in 1.2.1 and 1.2.2 are re-ranked according to the retrieval list obtained by the action recognition.

2. Conclusion

This year, we improve person retrieval by introducing object tracking algorithm, which helps to backtrack the important frames that can't be retrieved only by face recognition method. We also introduce STGCN to improve ECOfull in action retrieval task. Moreover, object detection and emotion recognition models are updated to achieve better performance. In the next year, we will try to mine more information from text and audio.

Acknowledgment

In this paper, BBC EastEnders frames are used for non-commercial individual research and private study use only. BBC content included courtesy of the BBC.

References

[1] Zhang, Kaipeng , et al. "Joint Face Detection and Alignment Using Multitask Cascaded Convolutional Networks." IEEE Signal Processing Letters 23.10(2016):1499-1503.

[2] http://dlib.net/files/dlib-19.18.tar.bz2.

[3] L. Zheng, S. Wang, L. Tian, F. He, Z. Liu, and Q. Tian. Queryadaptive late fusion for image search and person re-identification. In Computer Vision and Pattern Recognition, 2015. 1

[4] Wang K, Peng X, Yang J, et al. Region attention networks for pose and occlusion robust facial expression recognition[J]. IEEE Transactions on Image Processing, 2020, 29: 4057-4069.

[5] Barsoum E, Zhang C, Ferrer C C, et al. Training deep networks for facial expression recognition with crowd-sourced label distribution[C]//Proceedings of the 18th ACM International Conference on Multimodal Interaction. 2016: 279-283.

[6] http://www.pitt.edu/~emotion/ck-spread.htm

[7] Wang X, Huang T E, Darrell T, et al. Frustratingly Simple Few-Shot Object Detection[J]. arXiv preprint arXiv:2003.06957, 2020.

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