# PKU_WICT at TRECVID 2020:

# Instance Search Task

Yuxin Peng, Zhaoda Ye, Junchao Zhang, and Hongbo Sun
Wangxuan Institute of Computer Technology,
Peking University, Beijing 100871, China.
pengyuxin@pku.edu.cn

## Abstract

In TRECVID 2020, we participated all two types of Instance Search (INS) task, including automatic search and interactive search. For the automatic search task, we proposed a two-stage approach consisting of similarity computing and result re-ranking. Our approach recognized and obtained the action and person prediction scores separately, and then merged their scores for instance search. In *action-specific recognition*, our approach achieved action recognition from four aspects: frame-level action recognition, video-level action recognition, object detection and facial expression recognition. In the *person-specific recognition*, our approach adopted a deep neural network and applied top *N* query extension strategy. In the *instance score fusion*, we merged the action and person prediction scores and applied a re-ranking strategy to combine the information from action-specific recognition and person-specific recognition. For the interactive search task, the interactive query expansion strategy was applied to refine the results from automatic search. The official evaluations showed that our approach ranked 1st in both automatic and interactive search.

# 1. Overview

In TRECVID 2020[1], we participated all two types of Instance Search (INS) task, including automatic search and interactive search. We submitted totally 14 runs: 12 automatic runs and 2 interactive runs, and the official evaluation results are shown in Table 1. In both automatic search and interactive search, our team ranked 1st among all teams in the Main task and Progress task. Table 2 gives a brief description of the symbols used in Table 1. The overall framework of our approach is shown in Figure 1.

In our submitted runs, the notation "A" indicates the provided video examples were not used, while "E" is the opposite of "A". The notation "M" denotes the Main task and the "P" denotes Progress task. "Run3" contains all the components proposed in the automatic search, including frame-level and video-level action recognition, object detection, facial expression recognition, deep face recognition and top *N* query extension strategy. Compared with "Run3", the object detection and facial expression recognition were not adopted in "Run2", and "Run1" indicates the top *N* query extension strategy was not applied in the corresponding submitted runs. "Run4" indicates interactive runs with human feedback.

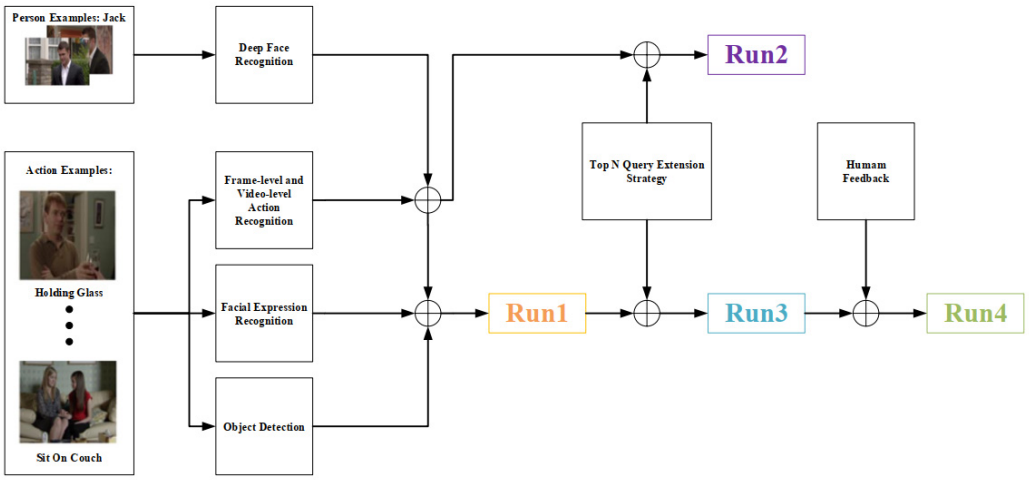


**Figure 1: Framework of our approach for the submitted runs.**

**Table 1: Results of our submitted 14 runs on Instance Search task of TRECVID 2020.**

| Task | Type | ID | MAP | Brief description |
|---|---|---|---|---|
| Main | Automatic | PKU_WICT_RUN3_ME | **0.252** | A+O+E+F+T |
| | | PKU_WICT_RUN3_MA | 0.247 | A+O+E+F+T |
| | | PKU_WICT_RUN2_ME | 0.195 | A +F+T |
| | | PKU_WICT_RUN2_MA | 0.196 | A +F+T |
| | | PKU_WICT_RUN1_ME | 0.245 | A+O+E+F |
| | | PKU_WICT_RUN1_MA | 0.207 | A+O+E+F |
| | Interactive | PKU_WICT_RUN4_M | **0.368** | A+O+E+F+T+H |
| Progress | Automatic | PKU_WICT_RUN3_PE | **0.178** | A+O+E+F+T |
| | | PKU_WICT_RUN3_PA | 0.171 | A+O+E+F+T |
| | | PKU_WICT_RUN2_PE | 0.165 | A +F+T |
| | | PKU_WICT_RUN2_PA | 0.163 | A +F+T |
| | | PKU_WICT_RUN1_PE | 0.174 | A+O+E+F |
| | | PKU_WICT_RUN1_PA | 0.123 | A+O+E+F |
| | Interactive | PKU_WICT_RUN4_P | **0.253** | A+O+E+F+T+H |

**Table 2: Description of our methods.**

| Abbreviation | Description |
|---|---|
| A | Frame-level and Video-level **A**ction Recognition |
| O | **O**bject detection |
| E | Facial **E**xpression Recognition |
| F | Deep **F**ace recognition |
| T | **Top** *N* Query Extension Strategy |
| H | **H**uman Feedback |

# 2. Our approach

## 2.1 Action-specific Recognition

In this year, 16 actions were required to be recognized, such as “hugging”, “kissing”, and “holding cloth”, etc. The main challenges are summarized as follows: (1) Some categories involve similar content that is hard to distinguish. For example, categories “open door enter”, “open door leave”, “close door without leaving” and “stand talk door” describe the actions that take place at the door, where the subtle details are crucial to distinguish opening the door from closing the door, or to distinguish entering from leaving. (2) The provided training data is scarce, since only 4~6 video clips are provided for per action category. However, current popular action recognition models based on deep learning always require a large number of training data.

Similar to the last year, our approach took multiple strategies for action recognition, including frame-level and video-level action recognition, object detection, as well as facial expression recognition. For collecting enough training data, on the one hand, existing relevant datasets were directly adopted to train the deep models for action recognition. On the other hand, we crawled images from the Internet as expanded training data for the fully automatic runs and interactive runs that state the source of training data with “E”, which were allowed to use external data for training purposes in INS task.

All the strategies were finally integrated to boost the action recognition, where we computed the fusion value of the all prediction scores of a shot as the final action prediction score $ActScore$.

### 2.1.1 Frame-level Action Recognition

Some action categories describe the state of involved persons, which can be directly recognized from the frames, such as “sit on couch”, “holding phone”, etc. In our approach, an image classification method was adopted for frame-level action recognition. The official guideline encouraged the participated teams to collect their own data for training purposes, thus we first collected web images and frames from relevant video datasets to construct the training data. Then the image classification model SENet was trained with a progressive training strategy, which was used to predict the category score of each video frame. Finally, the frame-wise prediction scores were merged as the recognition results of the corresponding video.

**(1) Training Data Collection**

According to the official guideline, the provided examples cannot include all possible variations of action appearance, thus it is necessary to collect extra data for training purpose. In our approach, the training data was collected in two ways. First, we exploited the images on the Internet. Considering NIST provided detailed text definitions of the action categories, we selected several keywords for each category and took the widely-used search engines, Baidu[1] and Bing[2], to search

[1] https://image.baidu.com
[2] https://cn.bing.com

related web images. Second, we exploited video frames from existing datasets. Existing datasets, such as Kinetics-400[2] , contain similar action categories in the INS task, which can be adopted to form the training data for frame-level action recognition.

**(2) Model Training and Testing**

The image classification technique was utilized to address the frame-level action recognition. Image classification methods[3][4] can distinguish the categories, which can capture the discriminative details of images. Specifically, SENet[5] is adopted as the frame-level action recognition model in our approach.

For model training, we exploited the progressive training strategy. Our approach used the collected web images and video frames as training data to train the SENet model. Then the trained SENet was employed to predict the action categories of video frames in INS task. According to the prediction scores, the top *N* frames were selected and added to training data. Then these augmented training data would be used to train the SENet model. The progressive training strategy increased the number of training data, and made the distribution of training data close to the data of INS task, which improved the recognition accuracy. The final well-trained SENet model was used to predict the category score of each frame, and the maximal score of frames in each shot was adopted to represent the category of the shot.

## 2.1.2 Video-level Action Recognition

Some action categories involve a long range of behavior across multiply frames, which should be recognized from multiple frames, such as “go up/down stairs”. Thus, our approach adopted two models, including StNet model[6] and Non-local network[7] for video-level action recognition.

**(1) Training Data Collection**

In the officially provided data, there are only 4~6 video clips for each category that can be taken as training data, which cannot support to train the deep learning based models. Because the examples cannot include all possible variations of action appearance, we collect the action development data for training purposes according to the suggestion of the official guideline. For collecting sufficient data for model training, we directly exploited the videos in Kinetics-400 dataset. Kinetics-400 is a video dataset that contains 400 human action categories, and each category includes more than 400 video clips, which can cover 11 relevant categories in INS task. Thus, we just took the videos of these categories as the training data. As shown in Table 3, data of the categories in the right column was used as the training data of the categories in the left column. The officially provided video clips in INS task were also included in the training data.

**(2) Model Training and Testing**

StNet model[6] and Non-local network[7] were adopted in video-level action recognition. StNet took ResNet[8] as the backbone, which exploited 2D and 3D convolutions to capture the local and global spatio-temporal information. Non-local network embedded non-local blocks in deep networks, which can capture long-range dependencies in the video. Both StNet model and Non-

local network used in our method took ResNet-50 network as the backbone. After training, our approach utilized the trained models to classify all the video shots in INS task, and the predicted category scores of StNet model and Non-local network were merged together for better recognition results.

**Table 3: Corresponding relationships of categories between INS task and Kinetics-400.**

| Categories in INS task | Categories in Kinetics-400 |
|---|---|
| holding phone | texting |
| drinking | drinking, drinking beer, drinking shots, tasting beer |
| crying | crying |
| laughing | laughing |
| go up/down stairs | climbing ladder |
| holding paper | folding napkins, folding paper, reading book, reading newspaper, ripping paper, shredding paper |
| holding cloth | folding clothes |
| smoking cigarette | smoking, smoking hookah |
| kissing | kissing |
| shouting | yawning |
| hugging | hugging |

## 2.1.3 Object Detection

Some action categories describe the interactions of humans and objects, for which the object can be taken as a strong clue for action recognition. Thus, we exploited object detection technique to facilitate action recognition. Table 4 lists some actions in INS task, which are related to the object categories in external datasets, including MS-COCO[9], Visual Genome[10], etc. MS-COCO includes 80 object categories, such as “bottle”, “couch”, etc., and Visual Genome includes 33877 object categories, such as “door”, “book”, etc. Both of them are widely-used benchmarks for object detection.

Mask R-CNN[11] model pre-trained on MS-COCO and Visual Genome datasets was adopted for object detection in our approach, which was directly employed to detect objects in all frames. In addition, for the action “smoking cigarette”, we did not find appropriate category in existing public datasets. So we collect related images from Internet for training purposes, and adopt YOLOv5 model[3] for cigarette detection.

After obtaining object detection scores of each frame, the maximal scores of the frames in a shot was regarded as the detection score of the corresponding shot.

## 2.1.4 Facial Expression Recognition

[3] https://github.com/ultralytics/yolov5

The facial expression recognition task was designed to identify the expressions of human, such as anger and happiness. In INS task, 3 actions were hugely related to human facial expressions, namely "shouting", "laughing" and "crying". Therefore, facial expression recognition was used here to assist action recognition.

**Table 4: Corresponding relationships between actions and objects.**

| Actions in INS task | Objects in external datasets | Dataset |
|---|---|---|
| sit on couch | couch | MS-COCO |
| holding phone | cell phone | |
| drinking | bottle, wine glass, cup | |
| go up/down stairs | stairs | Visual Genome |
| holding paper | newspaper, book, papers, envelope | |
| open door enter | door, gate, door knob, door handle, garage door | |
| open door leave | | |
| stand talk door | | |
| close door without leave | | |

**(1) Training Data Collecting**

Because the official guideline encouraged the participated teams to collect data for training purposes, the data from both widely-used datasets and search engines were gathered to form the training data for facial expression recognition in our approach. Concretely, the data was collected from two datasets, CK+[12] and FER2013[13], which were dedicated to facial expression recognition. We also employed an image crawler to image search engines, Baidu and Bing, with related keywords. Face detection model MTCNN[14] was utilized subsequently to crop human faces from the above images for constructing the training data.

**(2) Model Training and Testing**

VGG-19[15] model pre-trained on ImageNet dataset was chosen as our facial expression classifier. We further fine-tuned VGG-19 model with aforementioned training data. During the testing stage, our approach first detected human faces from video frames, and then took the cropped face images as input to extract features from VGG-19 model. Similarly, the maximal score of the frames in a shot was regarded as the prediction score of the shot.

# 2.2 Person Identification

For person identification, our approach first detected faces in query examples, and filtered out abnormal "bad" faces of low detection confidence and complemented with "good" faces of high detection confidence. Second, face features from queries and shots were extracted based on deep convolutional neural networks and calculated the similarity. Top *N* query extension strategy was conducted for further improving the retrieval results for specific person.

### 2.2.1 Face Detection and Complementation

Our approach detected faces of query examples and key frames of each shot by the pre-trained MTCNN, which could help localize the face area for feature extraction. We observed that not all the detected faces of query examples were helpful for retrieval performance. Sometimes, there existed low-quality abnormal faces namely "bad" faces, which were quite different from other faces due to the camera angle or light intensity and always corresponded to detected faces of low confidence. Thus, we filtered out "bad" faces and only adopt the "good" faces of high confidence automatically. In this way, the quality of query person faces was improved to facilitate retrieval performance.

### 2.2.2 Person Identification based on Deep Neural Network

After detecting faces from queries and shots in the last step, our approach extracted the feature vectors by FaceNet model[16] pre-trained on VGGFace2 dataset[17]. Based on the extracted face features, our approach calculated the prediction scores via cosine distance between query person examples and all the key frames of shots. The largest prediction score between the shot *i* and person *j* is denoted as $PerScore_{ij}$, which can be used to get the ranking list for specific person and realize preliminary person identification.

### 2.2.3 Top *N* Query Extension strategy

We observed that the initial query feature could not sufficiently represent specific person. Thus our approach tried to utilize relevant data from shot frames, which adopted a top *N* query extension strategy based on retrieval shot results of the first round. Specifically, our approach calculated the mean feature vector of top *N* nearest frame faces from the initial retrieval shot results. In this way, the query feature could be updated into a more discriminative feature, which was closer to real representation of specific person. Then, a new round of retrieval was conducted based on the new feature to refine the person identification results.

### 2.2.4 Text-based Search

The transcripts of video provided by NIST presented some clues to distinguish the speakers. To exploit the clues in transcripts, our approach adopted a text-based search on each topic in the transcripts. First, the people's information was extended through Wikipedia web pages (such as nicknames, role names, family members of specific people, and the names of his/her closest friends, etc.). Then, we retrieved the shots whose transcripts contain the keywords of topic. And the retrieval results can help us to adjust the prediction score in the score fusion process with a reward mechanism. See section 2.3 for more details.

## 2.3 Instance Score Fusion

In Sections 2.1 and 2.2, we have introduced the first stage for computing the prediction scores of action and person. In this section, we will introduce the re-ranking stage with two fusion strategies,

which fused the prediction scores of action and person in two aspects:

(1) Searching person based on candidate action shots. For each query, our approach first obtained the candidate action shot list with the action prediction scores mentioned in Section 2.1. Then a text-based reward mechanism was applied to adjust person prediction scores as follows:

$$s_1 = \mu \cdot PerScore \tag{1}$$

where $\mu$ is the reward parameter. We set $\mu > 1$ if the shot contain corresponding person according to the text-based search on the transcripts of videos, otherwise $\mu = 1$. Thus, it can utilize the information from the transcripts to improve the accuracy of prediction scores. Besides, the shots which were not included in top *N* action-specific results were dropped by setting $\mu = 0$. Finally, candidate action shots can be re-ranked according to the score $s_1$.

(2) Searching specific action based on candidate person shots. Similar to the previous step, our approach first obtained candidate shots with person according to the prediction scores between persons mentioned in Section 2.2, and defined the action prediction scores as follows:

$$s_2 = \mu \cdot ActScore \tag{2}$$

where $\mu$ is the parameter. We dropped the shots which were not included in top *N* person-specific results by setting $\mu = 0$, and $\mu = 1$ otherwise. Finally the candidate shots with persons could be re-ranked according to the score $s_2$. The fusion score of a shot would be calculated as:

$$s_f = \omega(\alpha s_1 + \beta s_2) \tag{3}$$

where $\alpha$ and $\beta$ are weight parameters, and $\omega$ is a bonus parameter. If the shot simultaneously existed in the top *N* action-specific results and top-*M* person-specific results, we set $\omega > 1$ to improve the significance of this shot, otherwise $\omega = 1$. The bonus parameter $\omega$ is designed to highlight those shots that involve the query in both action and person aspects. Finally, we obtained fusion scores by preserving information of both action-specific recognition and person aspects-specific recognition.

We also considered the continuity of videos, and proposed a time sequence based re-ranking strategy to refine the results, which can adjust the scores of each shot according to the adjacent frames. Concretely, our approach recalculated the score of each shot by its neighbor shots' score as follows:

$$s_f^{(i)} = \theta \sum\nolimits_{-T<k<T} s_f^{(i+k)} + (1-\theta) s_f^{(i)} \tag{4}$$

where $s_f^{(i)}$ denotes the score of $i$-th shot, $T$ defines the range of the adjacent frames, and $\theta$ is a parameter to adjust the prediction score.

# 3 Interactive Search

The strategy for interactive search is similar to our previous approach in INS 2019[18]. The interactive search was based on RUN3_ME/ RUN3_PE and adopted the top *N* (*N*>1000) results as the candidate shots. First, the user returned the positive or negative labels for each topic's top-ranked results in the candidate shots. Next, the positive samples (10 samples for each topic) were regarded

as expanded queries to compute the prediction scores of the other candidate shots. Finally, it fused the scores of expanded and original queries to re-rank the candidate shots, where the negative samples were discarded.

# 4 Conclusion

Through the INS task in TRECVID 2020, we concluded that: (1) Object and facial expression were helpful to improve the accuracy of action recognition. (2) More attention should be focused on action recognition to improve accuracy of the INS task in the future work. (3) Human feedback can significantly boost the accuracy of INS task.

# Acknowledgements

This work was supported by the National Natural Science Foundation of China under Grant 61925201 and Grant 61771025. For the using of BBC EastEnders video or images snapshots, we thank for the programme material copyrighted by BBC.

# 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, Proceedings of TRECVID 2020, 2020.

[2] João Carreira, and Andrew Zisserman. Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset. IEEE Conference on Computer Vision and Pattern Recognition, pp. 6299-6308, 2017.

[3] Yuxin Peng, Xiangteng He, and Junjie Zhao. Object-Part Attention Model for Fine-grained Image Classification. IEEE Transactions on Image Processing, 27(3): 1487-1500, 2018.

[4] Tianjun Xiao, Yichong Xu, Kuiyuan Yang, Jiaxing Zhang, Yuxin Peng, and Zheng Zhang. The Application of Two-level Attention Models in Deep Convolutional Neural Network for Fine-grained Image Classification. IEEE Conference on Computer Vision and Pattern Recognition, pp. 842-850, 2015.

[5] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation Networks. IEEE Conference on Computer Vision and Pattern Recognition, pp. 7132-7141, 2018.

[6] Dongliang He, Zhichao Zhou, Chuang Gan, Fu Li, Xiao Liu, Yandong Li, Limin Wang, and Shilei Wen. StNet: Local and Global Spatial-temporal Modeling for Action Recognition. AAAI Conference on Artificial Intelligence, pp. 8401-8408, 2019.

[7] Xiaolong Wang, Ross Girshick, Abhinav Gupta, Kaiming He. Non-local Neural Networks. IEEE Conference on Computer Vision and Pattern Recognition, pp. 7794-7803, 2018.

[8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. IEEE Conference on Computer Vision and Pattern Recognition, pp. 770-778, 2016.

[9] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: Common Objects in Context. European Conference on Computer

Vision, pp. 740-755, 2014.

[10] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li Jia-Li, David Ayman Shamma, Michael Bernstein, and Li Fei-Fei. Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations. arXiv, 2016.

[11] Kaiming He, Georgia Gkioxari, Piotr Dollar, and Ross Girshick. Mask R-CNN. IEEE International Conference on Computer Vision, pp. 2961-2969, 2017.

[12] Patrick Lucey, Jeffrey F. Cohn, Takeo Kanade, Jason M. Saragih, Zara Ambadar, and Iain A. Matthews. The Extended Cohn-kanade Dataset (ck+): A Complete Dataset for Action Unit and Emotion-specified Expression. IEEE Conference on Computer Vision and Pattern Recognition-Workshops, pp. 94-101, 2010.

[13] Challenges in Representation Learning: Facial Expression Recognition Challenge: https://www.kaggle.com/c/challenges-in-representation-learning-facial-expression-recognition-challenge/data.

[14] 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.

[15] Karen Simonyan, and Andrew Zisserman. Very Deep Convolutional Networks for Large-Scale Image Recognition. International Conference on Learning Representations, pp. 1-14, 2015.

[16] Schroff F, Kalenichenko D and Philbin J. FaceNet: A Unified Embedding for Face Recognition and Clustering. IEEE Conference on Computer Vision and Pattern Recognition, pp. 815-823, 2015.

[17] Qiong Cao, Li Shen, Weidi Xie, Omkar M. Parkhi, and Andrew Zisserman, Vggface2: A Dataset for Recognising Faces across Pose and Age. IEEE International Conference on Automatic Face & Gesture Recognition, pp. 67-74, 2018.

[18] Yuxin Peng, Zhang Wen, Yunkan Zhuo, Junchao Zhang, Zhaoda Ye, and Hongbo Sun, PKU_ICST at TRECVID 2019: Instance Search Task, TRECVID, Nov. 12-13, 2019.