Instance Search via Fusing Hierarchical Multi-level Retrieval and Human-object Interaction Detection

Wenhao Yang[1], Yinan Song[1], Zhicheng Zhao[1,2], Fei Su[1,2]
[1]Beijing University of Posts and Telecommunications
[2]Beijing Key Laboratory of Network System and Network Culture, China
{whyang78, songyn, zhaozc, sufei}@bupt.edu.cn

Abstract

Aiming to retrieve specific persons with specific actions, instance-based video search (INS) has attracted rising attention with the development of video understanding. In this paper, a novel hierarchical multi-task INS retrieval framework is proposed. Firstly, a multi-level action recognition framework and a face matching scheme are introduced to obtain initial action and person retrieval scores separately. In particular, a novel graph-based human-object interaction (HOI) detection model, named interaction-centric graph parsing network (iCGPN), is proposed to recognize interactions between human and objects. Secondly, an improved query extension strategy is adopted to re-rank the initial person retrieval results. Thirdly, more elaborate action features are extracted to recognize complicated actions. Finally, a specially designed fusion strategy is used to integrate the retrieval results of persons and actions to generate the final INS ranking list. The experimental results show the effectiveness of the proposed framework.

1. Introduction

With the development of the Internet and the vastly increasing amount of video data, searching in video is a common task in many areas, such as media and entertainment. TRECVID [2] is dedicated to promoting the video understanding with one of the key subtasks, i.e., instance-based video search (INS). The core goal of INS is to locate shots containing specific person doing one of the predefined actions in given a set of videos.

INS usually suffers from the following situations: (1) different actions are very similar and hard to distinguish, for example, “holding glass” and “drinking” both contain the action of holding glass. (2) some actions related to doors, such as “open door enter”, are rarely involved in existing external datasets of action recognition. (3) the result of the action recognition related to ‘holding’ has a great relationship with the performance of the object detection model, such as “holding paper” and “holding cloth”.

To solve the above problems, a novel hierarchical multi-task retrieval framework is proposed, and the task is parsed into two main subtasks, that is, person retrieval and action retrieval. We first retrieve specific persons in video frames based on facial feature representations of query person images. Then we propose a hierarchical multi-level action recognition framework, including frame-level, clip-level and video-level to handle the complicated action retrieval. In addition, a lightweight Convolutional Neural Network (CNN) pretrained on CK+ [15] and FERPLUS [1] is applied to recognize emotion related actions, such as “shouting”. Moreover, an action similarities based fusion strategy is presented to fuse multi-level results. Finally, we design a re-ranking strategy to merge the action and person retrieval results together to get the final ranking list.

To summarize, our contributions are as follows:

(1) Propose a novel hierarchical action recognition based multi-task retrieval framework.

(2) To enhance the performance of action recognition, a multi-level (frame-, clip-, video-level) action recognition framework is adopted.

(3) Propose a graph-based HOI detection model iCGPN to recognize interactions between human and objects.

(4) Design action similarities based and re-ranking based fusion strategies to merge hierarchical multi-task results.

The rest of this paper is organized as follows: Section 2 discusses related work. Section 3 introduces the proposed methods. The experimental results and failure analysis are discussed in Section 4 and Section 5 respectively. Section 6 summaries the conclusions.

2. Related Work

Action Detection. Feichtenhofer et. al. proposed a SlowFast [7] network, consisting of two pathways with different frame rates to process spatial semantic information and motion information separately, and achieved state-

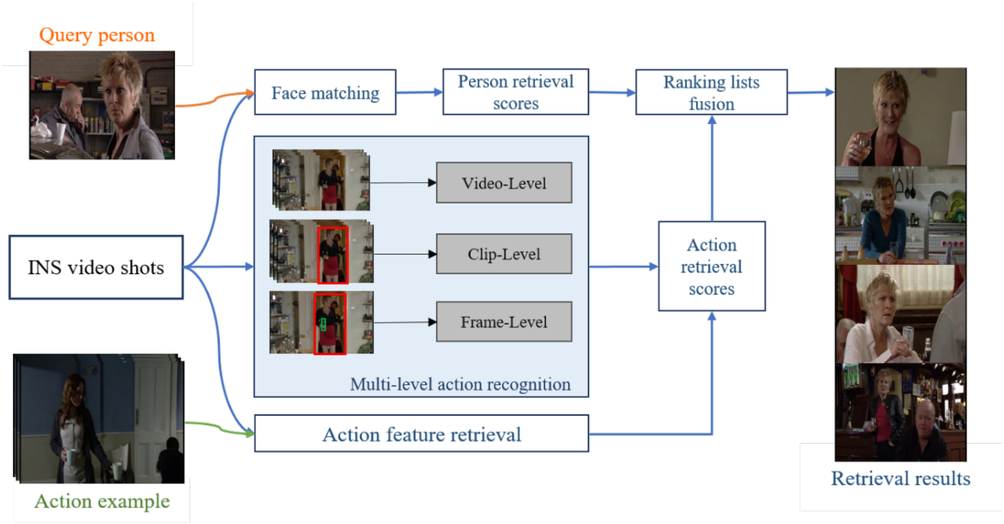

Figure 1. Overview of our INS framework.

of-the-art accuracy with lower computational complexity for video action recognition on several video recognition benchmarks such as Kinetics-400 [10], Kinetics-600 [4] and AVA-Kinetics [12] dataset.

Human-Object Interaction Detection. Human-object interaction (HOI) detection aims to localize human and objects, as well as to identify the complex interactions between them, so it is usually represented by a triplet $\langle human, action, object \rangle$. GPNN [16] was designed to equally treat all detected objects as the graph nodes to learn node relationships by applying graph convolutional networks (GCNs). AGRR [14] performed relation reasoning on human-object pairs with a graph network, where each graph node contained the information of each human-object pair. However, the dominant role of humans should be obvious, so we propose a heterogeneous graph iCGPN that models humans and objects as different kinds of nodes in this paper. We denote each person as the central node, and then construct a fully-connected graph and predict interactions for the central human. Experiments show our proposed model obtain the best performance on the HICO-DET [5] compared with some state-of-the-art methods.

3. Method

The proposed INS framework mainly consists of two subtasks: person retrieval and action retrieval. Figure 1 illustrates the details of our framework.

3.1. Action retrieval

We proposed a multi-level action recognition framework, including frame-level, clip-level and video-level to enhance the performance of action recognition. In addition, elaborate action feature retrieval methods are adopted to improve the recognition accuracy of those actions which are rarely involved in action recognition datasets.

3.1.1 Action recognition

Video-level action recognition. Video-level action recognition aims to recognize actions from videos and we adopt SlowFast model pretrained on Kinetics-400 [10] dataset to roughly judge whether the action occurs in video shots of INS database. Then we take the shot scores as the scores of all key frames in the shot.

Clip-level action recognition. The goal of clip-level action recognition is to localize persons in key-frames, and meanwhile, recognize actions from video clips. Compared with the video-level methods, it can obtain action scores of specific persons at key-frames level. We firstly train SlowFast model on AVA-Kinetics [12] dataset. Then we use Cascade R-CNN [3] pretrained on COCO dataset to locate the persons in key-frames of INS video shots. With the pretrained model above, the action scores of each detected person in key-frames are obtained.

Frame-level HOI detection. It aims to recognize actions from a single frame, which contain obvious human-object interactions (HOI), such as ‘sit on couch’ and ‘holding glass’. A HOI detection model, named iCGPN, which trained on HICO-DET [5] dataset is adopted. As shown in Figure 2, the node feature is represented by visual features and spatial information between human and objects. Then a graph convolutional network is applied to update node features according to the connectivity matrix and predict the final HOI labels. In addition, HOI detection consists of two main steps: object detection and HOI prediction. And the performance of HOI detection model relies on the object

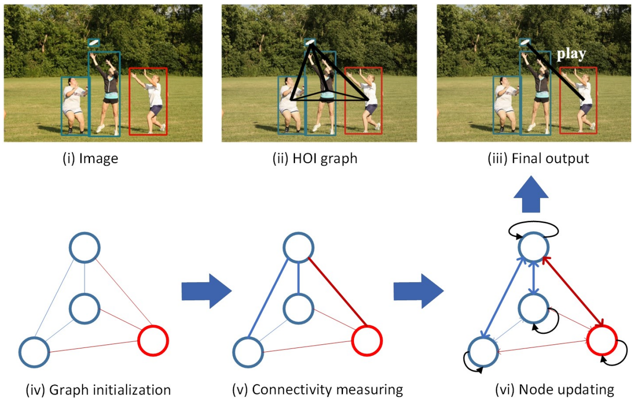

Figure 2. Illustration of the proposed interaction-centric graph parsing networks. Aiming to recognize the interaction between the right human (red box) and frisbee, we treat red node as central human node and blue nodes as object nodes.

detection model to a great extent. Thus, we conduct a new object detection dataset to train Cascade R-CNN. Table 2 shows the dataset sources of all object classes. With the trained Cascade R-CNN and iCGPN, we can get the interaction scores for each human-object pair.

Scores fusion. The final fusion multi-level score of action is defined as equation (1):

$$a = \mu_v a_v + \mu_c a_c + \mu_f a_f, \ \ s.t. \, \mu_v + \mu_c + \mu_f = 1 \quad (1)$$

where a_v, a_c, a_f respectively refer to video-level, clip-level, frame-level prediction scores, and μ_v, μ_c, μ_f are weight hyperparameters. Specifically, the hyperparameters will set to 0 if the action classes are not in the level of action recognition according to Table 1. Additionally, we set μ_v smaller for all actions because video-level prediction scores are originally shot scores and it is inaccurate to serve as the scores of shot key frames. μ_f is set larger for human-object interactions. Moreover, we also apply facial expression recognition methods for recognize “shouting”, additionally, we fuse “angry” scores and “talk to” scores from the clip-level method or ”talk with person” scores from the frame-level method to generate the prediction scores of “shouting”. Finally, we obtain action scores of all shot key-frames for all actions except for the four actions related to doors.

3.1.2 Action feature retrieval

Some actions related to doors, such as “open door enter”, are rarely involved in existing external datasets of action recognition. Furthermore, we cannot get prediction scores by the multi-level action recognition framework. Thus, we adopt action feature retrieval methods to handle these actions.

The SlowFast [7] model pretrained on AVA-Kinetics [12] is adopted to extract action features of key frames in

Actions in INS task	Kinetics-400	AVA-Kinetics	HICO-DET
sit on couch	-	sit	sit on couch
holding phone	-	answer phone	hold cell phone
holding glass	opening bottle	carry/hold	hold wine glass
carrying bag	-	carry/hold	hold backpack
holding paper	reading book	carry/hold	hold book
holding cloth	folding clothes	dress	hold cloth
drinking	drinking	drink	drink with cup
kissing	kissing	kiss	kiss person
shouting	-	talk to	talk with person
hugging	hugging	hug	hug person

Table 1. The corresponding relationships between INS task and Action Recognition Datasets.

query action examples and INS shots. Then we calculate the cosine similarity between them and set maximal similarity as the shot key-frames similarity. Let $\{x_i\}_{i=0}^n$ and $\{y_j\}_{j=0}^m$ denote the L2-norm features of shot key-frames and action example key-frames, where $x_i, y_j \in R^d$, the shot key-frame maximal similarity is:

$$s_i = \underset{j}{max}(x_i y_j^T), \ i \in [0, n], \ j \in [0, m] \qquad (2)$$

3.2. Person retrieval

We firstly adopt RetinaFace [6] to detect faces in all shot key frames and query person examples, and then extract face landmarks using PFLD [8] for face alignment. In order to alleviate the problem of undetected faces when persons are away from the camera lens, we introduce DeepSORT [21] to track the persons. Then, FaceNet [18] is used to obtain facial feature representation for cosine similarity matching. To reduce the impact of low-quality query person examples on the generated ranking lists, the search results with high similarity (S_h) are chosen to extend the query examples. In addition, the person retrieval ranking lists are re-ranked by top N α-weighted query extension (αQE) [17] strategy. In experiments, S_h=0.6.

3.3. Ranking lists fusion

To obtain the better performance, we apply late fusion in the post-processing stage. For the i-th key frame, the fusion score is calculated as follows:

$$f_i = \beta_1 p_i + \beta_2 a_i, \ \ s.t. \, \beta_1 + \beta_2 = 1 \qquad (3)$$

where f_i, p_i, a_i respectively denote to the final scores, person retrieval scores and action retrieval scores, and β_1, β_2 are weight hyperparameters. In experiments, we set β_1 and β_2 to 0.5. Finally, according to the final scores of all shot key-frames, we take the maximal score of the key-frames in each shot as the shot scores to generate the shot ranking lists for all person-action pairs of INS task.

4. Experiments

Performance of iCGPN. To evaluate the effectiveness of the proposed HOI model, we test HICO-DET [5] dataset

Objects in INS task	Dataset sources
couch	couch(COCO[13])
phone	cell phone(COCO), telephone(Objects365[19])
glass	cup(COCO), bottle(COCO), wine glass(COCO)
bag	backpack(COCO), handbag(COCO), suitcase(COCO), briefcase(Objects365)
paper	book(COCO), newspaper(Manual Labeling)
cloth	towel(Objects365), clothing(OpenImage[11]), jacket(OpenImage), coat(OpenImage)
person	person(COCO)

Table 2. Dataset sources of all object classes in INS task.

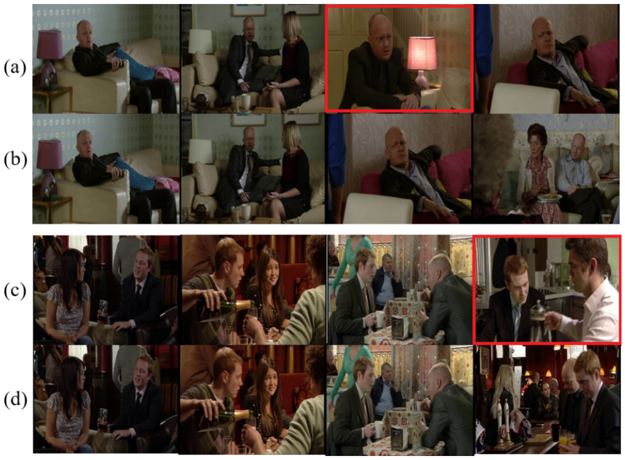

Figure 3. Two groups of visualization video retrieval results.

on five different models. We evaluate the HOI detection performance using the commonly used role mean average precision (role mAP [9]). A $\langle human, action, object \rangle$ triplet is considered as a true positive if the predicted action matches the ground-truth, and both predicted human and object bounding boxes have $IOUs \geq 0.5$ with reference to GT boxes. And we report the role mAP over three different HOI category sets: all 600 HOI categories in HICO-DET (Full), 138 HOI categories with less than 10 training instances (Rare), and 462 HOI categories with 10 or more training instances (Non-Rare). Table 3 shows that our proposed iCGPN achieves the best performance in default mode among other state-of-the-art methods. Particularly, iCGPN outperforms some graph-based methods by a significant margin, such as GPNN [16] and AGRR [14]. This model helps to improve the action recognition results.

Action retrieval results. Figure 3 gives several visualization retrieval results. (a) and (b) show Max is sitting on couch and (c) and (d) show Bradley is holding glass. The results of (a) and (c) are generated only by frame-level action recognition scores and the video annotated in red is wrong. The confusing spatial layouts between person and objects may lead to wrong HOI detection results. However, we also introduce the "sit" and "carry/hold" scores from clip-level action recognition to frame-level scores, thereby we get correct results as shown in (b) and (d). The visualization results show that the action recognition scores fusion strategy can effectively exclude some wrong results.

Method	Full	Non Rare	Rare
GPNN(ECCV2018)[16]	13.11	14.23	9.34
PMFNet(ICCV2019)[20]	17.46	18.00	15.65
AGRR(IJCAI2020)[14]	16.63	18.22	11.30
In-GraphNet(IJCAI2020)[22]	17.72	19.31	12.93
iCGPN(**ours**)	**18.78**	**19.80**	**16.03**

Table 3. Performance comparison on the HICO-DET test set.

5. Failure Analysis

Person retrieval. We attempt to fusion ranking lists using matching results of original query person examples. However, we find some query person examples are helpless to match correct results according to visualization results. Hence, we remove these "bad" queries and search correct matching results with high similarity to extend query examples. The final matching results are greater than before.

Frame-level HOI detection. We adopt a HOI detection model to identify the interactions between humans and objects. At first, we use the detection results from COCO pretrained models and we find that lots of objects cannot be detected, such as cloth and paper. However, the performance of HOI detection model relies on the object detection model to a great extent. We conduct a new object detection dataset to train Cascade R-CNN which is more effective to detect object related to INS task than pretrained model.

6. Conclusion

In this paper, we propose a novel multi-level INS framework, where specific persons and actions retrieval are accomplished and HOI is introduced to improve action recognition, and the results are fused by two ranking schemes. First, person retrieval scores are obtained by weighted ranking lists. Second, action retrieval scores are computed based on the proposed multi-level action recognition framework and action feature retrieval methods. Finally, action and person retrieval scores are merged to obtain the final shot ranking lists. The experimental results demonstrate the effectiveness of our INS framework.

7. Acknowledgement

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

References

[1] Training deep networks for facial expression recognition with crowd-sourced label distribution. 1

[2] George Awad, Asad A Butt, Keith Curtis, Jonathan Fiscus, Afzal Godil, Yooyoung Lee, Andrew Delgado, Jesse Zhang, Eliot Godard, Baptiste Chocot, et al. Trecvid 2020: A comprehensive campaign for evaluating video retrieval tasks across multiple application domains. *arXiv preprint arXiv:2104.13473*, 2021. 1

[3] Zhaowei Cai and Nuno Vasconcelos. Cascade r-cnn: Delving into high quality object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6154–6162, 2018. 2

[4] Joao Carreira, Eric Noland, Andras Banki-Horvath, Chloe Hillier, and Andrew Zisserman. A short note about kinetics-600. *arXiv preprint arXiv:1808.01340*, 2018. 2

[5] Yu-Wei Chao, Yunfan Liu, Xieyang Liu, Huayi Zeng, and Jia Deng. Learning to detect human-object interactions. In *2018 ieee winter conference on applications of computer vision (wacv)*, pages 381–389. IEEE, 2018. 2, 3

[6] Jiankang Deng, Jia Guo, Yuxiang Zhou, Jinke Yu, Irene Kotsia, and Stefanos Zafeiriou. Retinaface: Single-stage dense face localisation in the wild. *arXiv preprint arXiv:1905.00641*, 2019. 3

[7] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6202–6211, 2019. 1, 3

[8] Xiaojie Guo, Siyuan Li, Jinke Yu, Jiawan Zhang, Jiayi Ma, Lin Ma, Wei Liu, and Haibin Ling. Pfld: A practical facial landmark detector. *arXiv preprint arXiv:1902.10859*, 2019. 3

[9] Saurabh Gupta and Jitendra Malik. Visual semantic role labeling. *arXiv preprint arXiv:1505.04474*, 2015. 4

[10] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017. 2

[11] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, et al. The open images dataset v4. *International Journal of Computer Vision*, 128(7):1956–1981, 2020. 4

[12] Ang Li, Meghana Thotakuri, David A Ross, João Carreira, Alexander Vostrikov, and Andrew Zisserman. The ava-kinetics localized human actions video dataset. *arXiv preprint arXiv:2005.00214*, 2020. 2, 3

[13] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 4

[14] Xue Lin, Qi Zou, and Xixia Xu. Action-guided attention mining and relation reasoning network for human-object interaction detection. In *IJCAI*, pages 1104–1110, 2020. 2, 4

[15] Patrick Lucey, Jeffrey F Cohn, Takeo Kanade, Jason Saragih, Zara Ambadar, and Iain Matthews. The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression. In *2010 ieee computer society conference on computer vision and pattern recognition-workshops*, pages 94–101. IEEE, 2010. 1

[16] Siyuan Qi, Wenguan Wang, Baoxiong Jia, Jianbing Shen, and Song-Chun Zhu. Learning human-object interactions by graph parsing neural networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 401–417, 2018. 2, 4

[17] Filip Radenović, Giorgos Tolias, and Ondřej Chum. Fine-tuning cnn image retrieval with no human annotation. *IEEE transactions on pattern analysis and machine intelligence*, 41(7):1655–1668, 2018. 3

[18] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823, 2015. 3

[19] Shuai Shao, Zeming Li, Tianyuan Zhang, Chao Peng, Gang Yu, Xiangyu Zhang, Jing Li, and Jian Sun. Objects365: A large-scale, high-quality dataset for object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8430–8439, 2019. 4

[20] Bo Wan, Desen Zhou, Yongfei Liu, Rongjie Li, and Xuming He. Pose-aware multi-level feature network for human object interaction detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9469–9478, 2019. 4

[21] Nicolai Wojke, Alex Bewley, and Dietrich Paulus. Simple online and realtime tracking with a deep association metric. In *2017 IEEE international conference on image processing (ICIP)*, pages 3645–3649. IEEE, 2017. 3

[22] Dongming Yang and Yuexian Zou. A graph-based interactive reasoning for human-object interaction detection. *arXiv preprint arXiv:2007.06925*, 2020. 4