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

Wenhao Yang¹, Yinan Song¹, Zhicheng Zhao¹,², Fei Su¹,²
¹Beijing University of Posts and Telecommunications
²Beijing 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 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 summarizes 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-
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 \( \langle \text{human}, \text{action}, \text{object} \rangle \). 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

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

Table 1. The corresponding relationships between INS task and Action Recognition Datasets.
Objects in INS task | Dataset sources
--- | ---
couch | couch(COCO)
phone | cell phone(COCO), telephone(Objects365)
glass | cup(COCO), bottle(COCO), wine glass(COCO)
backpack | backpack(COCO)
handbag | handbag(COCO)
suitcase | suitcase(COCO), briefcase(Objects365)
bag | backpack(COCO), handbag(COCO), suitcase(COCO), briefcase(Objects365)
bottle | bottle(COCO)
wine glass | wine glass(COCO)
towel | towel(Objects365), clothing(OpenImage)
jacket | jacket(OpenImage), coat(OpenImage)
coating | coat(OpenImage)
coat | coat(OpenImage)
paper | book(COCO), newspaper(Manual Labeling)
person | person(COCO)

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

![Visualization results](image)

Figure 3. Two groups of visualization video retrieval results.

We evaluate the HOI detection performance using the commonly used role mean average precision (role mAP). A \( \langle \text{human}, \text{action}, \text{object} \rangle \) 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.

5. Failure Analysis

**Person retrieval.** We attempt to fuse 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


