Instance-based Video Search via Multi-task Retrieval and Re-ranking

Zhicheng Zhao[1,2], Guanyu Chen[1], Chong Chen[1], Xinyu Li[1], Xuanlu Xiang[1],
Yanyun 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
{zhaozc, loraschen, cc19456, lxinyu, jxncxxl, zyy, sufei}@bupt.edu.cn

Abstract

With the rapid growth of video data, instance-based video search (INS), i.e., retrieving videos according to specific objects, places, actions etc., has become more and more practical and important. In this paper, a novel INS framework based on multi-task retrieval and re-ranking is proposed to retrieve particular person doing specific action. Firstly, a face matching scheme is designed to match the target persons from videos. Secondly, an object detection network and an improved two-pathway key-pose estimation network (IECO) are introduced to explore semantic dependences between static visual object and person's behavior. Based on the dependences, an initial INS ranklist is obtained. Thirdly, via encoding absolute and relative positions of person's poses, a new relative pose representation (RPR) method is presented. Finally, regarding RPR as the input, a light action recognition network is constructed to re-rank INS results. The experimental results on HMDB, UCF101, JHMDB and BBC Eastenders datasets demonstrate the effectiveness of the proposed INS framework.

1. Introduction

Aiming at quickly retrieving specific person and action, instance-based video search (INS) [1] has attracted rising attention due to potential application prospect. Lots of methods were proposed to retrieve videos by learning video's spatio-temporal representations, where different 3D-CNNs were built. For example, action recognition networks ECO [24], I3D [2] and SlowFast networks [5] were converted from 2D-CNNs. However, these networks ignored the pose information - an important cue for action recognition. To solve this problem, some pose-based action detection models [4, 13, 23] were proposed and obtained promising results. We notice that above pose-based methods assumed that videos were shot from fixed views, thus only the absolute positions of human pose keypoints were used, i.e., they either focused on pose trajectory [4] or encoded human poses of a video into an image [13]. In fact, a large number of videos were dynamically shot, and global camera movements were common. The existing models would be disabled in similar condition.

Moreover, INS usually suffered from the following situation: many specific actions were too similar. For example, the differences of 'holding_glass', 'drinking', 'holding_plates' were not obvious, thus popular pose-based models and video representations couldn't handle.

To address above problems, we parse INS into three related subtasks, that is, face detection and matching, key-pose and object detection, action representation and detection, and then propose a novel multi-task retrieval and re-ranking framework. First, we retrieve specific persons based on faces. Second, rely on the results of key-poses and objects detection, the semantic dependences of target person and specific action are measured to rank the candidate videos. Third, an improved two-pathway ECO network (IECO) is designed to extract video representation. In the meantime, by encoding the absolute and relative positions of person's pose, a new relative pose representation (RPR) method is presented to alleviate the effect of camera movement. Regarding RPR as the input, a light action recognition network is constructed to get video representation. Finally we use video features generated from two methods to re-rank INS result list. Our contributions mainly include four aspects:

(1) Parse INS into multiple related visual subtasks, and propose a novel INS framework based on multi-task retrieval and re-ranking.

(2) An improved two-pathway ECO network (IECO) is designed to enhance pose representation.

(3) A new relative pose representation (RPR) is presented, and a light pose-based action recognition network is constructed to restrain the impacts of camera movement.

(4) The experimental results on four datasets demonstrate the effectiveness of the proposed INS framework.

The rest of paper is organized as follows. Related work is discussed in Section 2. The proposed framework and exper-

Figure 1. The overall flowchart of our framework.

imental results are given in Section 3 and 4. Failure analysis and conclusion are made in Section 5 and 6 respectively.

2. Related work

Face Detection. CascadeCNN [11] showed its discriminative capability and high performance. Additionally, Qin et. al. [15] proposed an end-to-end optimization method. UnitBox [20] introduced a new intersection-over-union loss function to boost the detection performance. S^3FD [22] proposed a scale-equitable face detection framework to handle different scales of faces. MTCNN [21] achieved a good balance between efficiency and accuracy. Dlib model [9] also showed high discrimination among faces.

Pose estimation. Human pose was an important cue for action recognition. Deep networks were widely used [3, 6, 14, 19], and most of them applied the classification network as the backbone, which usually decreased the resolution. To resolve this issue, Sun et. al. [18] proposed a High-Resolution Network (HRNet) and extracted high-resolution representation.

Action Recognition. A slice of successful action recognition models were converted from 2D-CNNs, where raw images or optical flow were input into networks to learn spatio-temporal features. Carreira et. al. [2] extended 2D-ConvNet and built a Two-Stream Inflated 3D ConvNet (I3D). Zolfaghari et. al. [24] proposed the ECO network to perfect the video representation. Feichtenhofer et. al. [5] presented a SlowFast network, which contained two pathways and processed videos in different frame rates, and achieved state-of-the-art results on Kinetics dataset [8].

Pose-based action detection. Since most action recognition models ignored human pose, a few recent methods were proposed to leverage pose information [4, 13, 23]. Choutas et. al. [4] introduced a novel video representation that encoded time information in keypoint heatmaps with color. Ludl et. al. [13] proposed a new pose representation that encoded position information of keypoints with color, and then transferred the video into an image whose every pixel was keypoints' position. However, above networks did not solve the problem of camera movement.

3. Approach

The proposed INS framework mainly consists of three components: face matching, key objects and poses detection, and action recognition. Figure 1 illustrates the flowchart of our framework.

3.1. Face matching

In our implementation, we firstly apply MTCNN [21] to detect faces. Then, we use Dlib model [9] to extract face features for similarity matching. Finally, an adaptive fusion scheme based on query extension (QE) is proposed to re-rank match lists, and it mainly includes three steps:

(1) Face matching is orderly carried out for query images, and ranklists are obtained.

(2) For each list, the mean similarity score of top-1000 results is calculated and normalized as the fusion weight.

(3) Ranklists are weighted and re-ranked.

3.2. Key object and pose detection

To search specific behavior, we explore the dependences between semantic objects and key-poses. Meanwhile, we apply object detection and pose estimation for the videos, existing in face matching ranklist.

In our implementation, we choose YOLOv3 [16] to detect key objects such as glass, bag, phone, person etc. Afterwards, we feed human bounding boxes into HRNet [18] to estimate person poses. In addition, we distinguish which pose belongs to the target person by comparing the estimated nose keypoints with target person's face position. Then, we simply calculate the distance between object po-

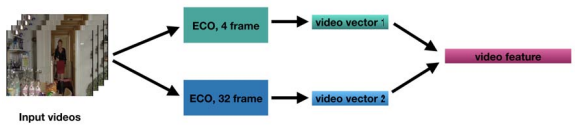

Figure 2. Video representation based on the improved two-pathway ECO network (IECO).

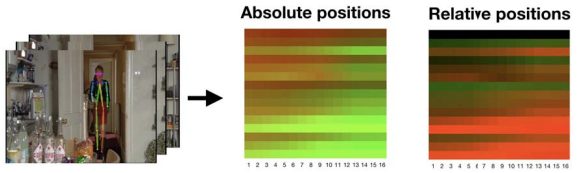

Figure 3. Two pose representation for human poses in a video.

sition and target person's nose position to measure the dependences of object-pose such as 'holding_glass', 'holding_phone' and 'carrying_bag'. Based on the dependences, an initial INS list can be determined.

3.3. Action retrieval

Spatio-temporal convolutional network. A great number of spatio-temporal networks were proposed to classify actions. We introduce action recognition network to extract video representation, and then carry out action retrieval.

In experiments, we choose ECO as the base network for feature extraction. Furthermore, inspired by SlowFast networks, where videos with different frame rates are input and video recognition performance is improved, we parallelly feed videos in different frame rates into ECO network to extract video representation. Figure 2 shows the flowchart of the improved two-pathway ECO (IECO).

Pose representation. Pose is crucial for action recognition since most of actions involve movements of human keypoints. However, common videos contain global camera motion, which will decrease the performance of pose-based action recognition model, where only absolute positions of poses are used. Here, we propose a new pose representation by considering both absolute and relative positions of poses, and alleviate the impact of camera movement.

As shown in Figure 3, we encode human pose joints in two ways. The left image shows the RGB feature map of absolute positions, where x, y positions of the k-th joint are encoded in red and green channels, and blue channel is set to zero. The right image shows the RGB feature map of relative distances and angles between keypoints. To calculate relative distances, we first compute the Euclidean distance dis_i between keypoint position pos_i and the nose position pos_n in a single frame:

$$dis_i = \|pos_i - pos_n\|_2 \tag{1}$$

Then, relative distance rela_dis_i in current frame is computed based on maximum normalization. This operation can eliminate the effect of different sizes of human bounding boxes.

$$rela\_dis_i = \frac{dis_i}{\max\left(dis_1, dis_2, ..., dis_k\right)} \tag{2}$$

As for relative angle rela_angle_i, we first calculate the angles between x-axis and the line that joints keypoint (x_i, y_i) and nose position (x_n, y_n), and then we divide angles by Equation 3.

$$rela\_angle_i = \begin{cases} \frac{\arccos \frac{x_i - x_n}{dis_i}}{360} & y_i \geq y_n \\ 1 - \frac{\arccos \frac{x_i - x_n}{dis_i}}{360} & y_i < y_n \end{cases} \tag{3}$$

In this paper, we use nose, neck, hip center, left shoulder, left elbow, left wrist, right shoulder, right elbow, right wrist, left hip, left knee, left ankle, right hip, right knee and right ankle as keypoints. The encoded joints are assigned in a $1 \times k \times 2$ matrix. k stands for the number of joints, and the third channel is zero. For all videos, we extract m frames to encode poses, thus a $m \times k \times 2$ matrix can be obtained, as shown in Figure 3. In experiment, m is set as 16. Finally, we get two pose representations for each video.

CNN on pose representations. To classify above pose representations we construct a light CNN, which consists of four convolutional layers followed by a fully connected layer (see Figure 4). After the second convolution layer, we apply a max-pooling layer to reduce the spatial resolution by a factor of 2. Like ECO method, instead of stacking two pose representations together, we separately feed them into the network and concatenate the vectors before the last layer. We train our network on JHMDB dataset [7] with shuffling about 60% of training images. Finally we use this network to extract video vectors for action retrieval.

4. Experiments

In this section, we evaluate the performance of the proposed INS framework on four datasets, i.e., HMDB [10], UCF101 [17], JHMDB [7] and BBC Eastenders [1]. Moreover, two pose-based action detection methods [4, 13] are compared.

Spatio-temporal convolutional networks. To evaluate the performance of IECO network on INS task, we conduc,t experiments on HMDB and UCF101 datasets. For each video in training set, we use IECO to extract video features and calculate the cosine distance between query and video features in test set. Figure 5 shows mAP of actions on HMDB and UCF101 respectively. We can see that IECO obtains good performance for the actions involving obvious movement. We also compare performance of IECO

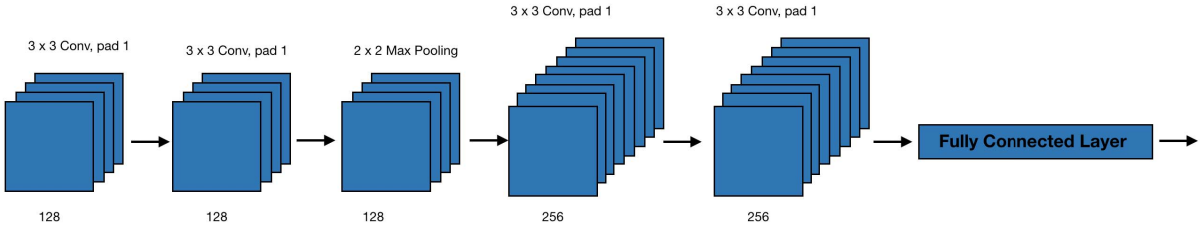

Figure 4. The network architecture for classifying pose representations. The number of channels of every layer is marked below.

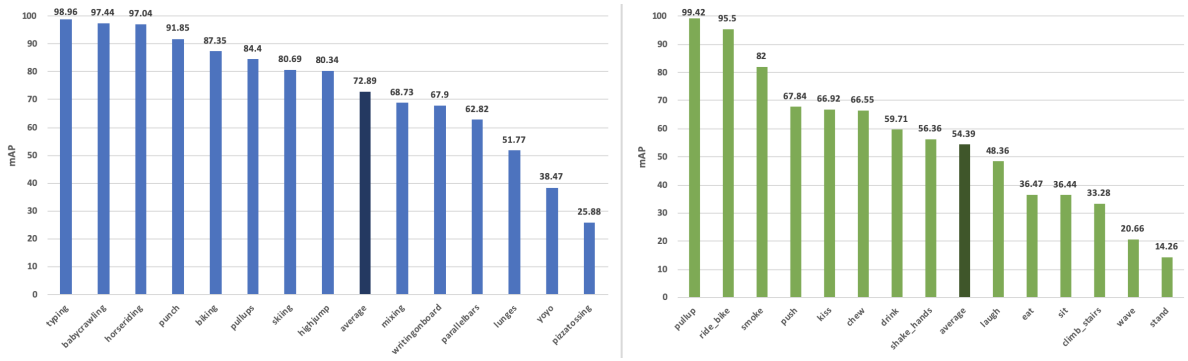

Figure 5. Actions and their mAP on UCF101 dataset (left) and HMDB dataset (right).

Pathway	HMDB (mAP)	UCF101 (mAP)
one (16 frame)	46.68	67.90
two (4 & 32 frame)	**54.39**	**72.89**

Table 1. Results on HMDB and UCF101 based on ECO with different pathways.

Architecture(channels)	JHMDB-1
(64, 128)	60.11 ± 2.81
(128, 256)	$\mathbf{62.29 \pm 2.5}$
(64, 128, 256)	60.49 ± 3.93
(128, 256, 512)	61.09 ± 4.08

Table 2. Results on JHMDB-1 with various channels and blocks.

with one pathway (frame rate is 16), and results are shown in Table 1, which indicates two-pathway is better than one pathway.

Posed-based action detection. We train pose representations with different architectures. Results are shown in Table 2. We can find that the best results on JHMDB dataset are achieved when 2 blocks (each block has two convolutional layers) and the number of channels is 128 and 256 respectively. Two state-of-the-art methods are also compared, and Table 3 gives the results, indicating that our network outperforms others on both two datasets.

INS based on face matching and action retrieval. Figure 6 visualizes a typical instance of face matching. Figure 7 gives several groups of visualization results, which shows our INS framework achieves promising action retrieval results on BBC Eastenders dataset.

Methods	JHMDB-1	JHMDB-1-GT
Choutas et. al. [4]	59.1	70.8
Ludl et. al. [13]	60.3 ± 1.3	65.5 ± 2.8
RPR(ours)	$\mathbf{62.29 \pm 2.5}$	$\mathbf{71.38 \pm 2.13}$

Table 3. Results on JHMDB-1 compared with two state-of-the-art algorithms. JHMDB-1-GT means using pose data given by JHMDB dataset to classify pose representations.

5. Failure Analysis

Face detection. We attempted to apply Dual Shot Face Detector (DSFD) [12] to detect faces. Compared with MTCNN, this model can correctly detect more faces in video frames, especially mini faces, while MTCNN's bounding boxes are more suitable.

Spatio-temporal convolutional network. I3D [2] was initially applied to extract video features and we tried to use optical flow to capture the movement information. However, it had high computation complexity since INS usually contained camera movement. We finally improve ECO network to obtain the tradeoff between runtime and accuracy.

Posed-based action retrieval. At first, we stacked two absolute position and relative position together and fed new feature maps into the action recognition network. However,

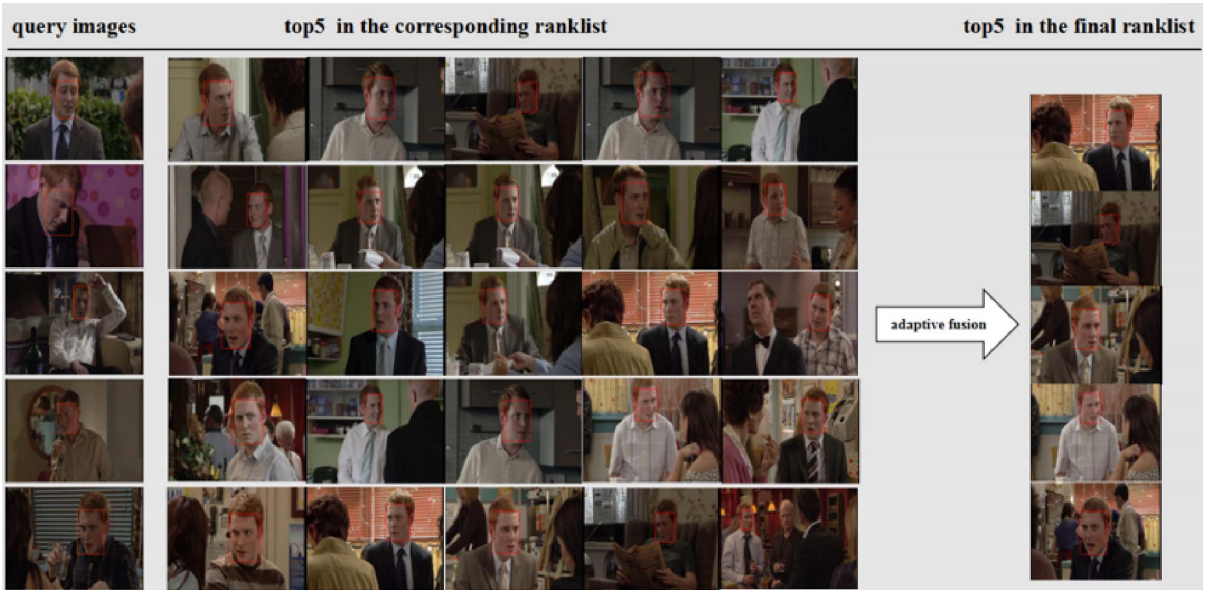

Figure 6. Visualization of face recognition result (face retrieval) results. Taking Bradley as an example, we show top 5 results for each rank list which is the retrieval result of one query image. Finally, we fuse five rank lists into single one.

Figure 7. Actions and their results for video retrieval on BBC Eastenders dataset.

Concatenation method	JHMDB-1-GT
stacked(one pathway)	68.51 ± 4.25
two pathway	$\mathbf{71.38 \pm 2.13}$

Table 4. Comparison of two different concatenation methods.

the results were even worse than only based on absolute position. Therefore, like SlowFast networks, we parallelly sent two pose representations into the network. Different from SlowFast, in which two pathways were relatively independent, our convolutional layers of two pathways shared the same weights. The comparison of two different concatenation methods on JHMDB-1-GT dataset is shown in Table 4, which indicates our method is better than SlowFast.

6. Conclusion

In this paper, we parse INS into three related subtasks, and propose a novel multi-task retrieval framework. First, we determine specific person based on face matching. Second, the semantic dependences of target persons and the corresponding behaviors are measured to rank the candidate videos. Third, a new relative pose representation (RPR) method is presented. Finally, a light pose-based action detection network and two-pathway ECO are constructed to re-rank INS result list. The experimental results on four datasets demonstrate the effectiveness of our INS framework.

Acknowledgements

This work is supported by Chinese National Natural Science Foundation (61532018, U1931202), and Key Laboratory of Forensic Marks, Ministry of Public Security of China.

References

[1] G. Awad, A. Butt, K. Curtis, Y. Lee, J. Fiscus, A. Godil, D. Joy, A. Delgado, A. Smeaton, and Y. Graham. Trecvid 2018: Benchmarking video activity detection, video captioning and matching, video storytelling linking and video search. 2018.

[2] J. Carreira and A. Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017.

[3] Y. Chen, Z. Wang, Y. Peng, Z. Zhang, G. Yu, and J. Sun. Cascaded pyramid network for multi-person pose estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7103–7112, 2018.

[4] V. Choutas, P. Weinzaepfel, J. Revaud, and C. Schmid. Potion: Pose motion representation for action recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7024–7033, 2018.

[5] C. Feichtenhofer, H. Fan, J. Malik, and K. He. Slowfast networks for video recognition. *arXiv preprint arXiv:1812.03982*, 2018.

[6] E. Insafutdinov, L. Pishchulin, B. Andres, M. Andriluka, and B. Schiele. Deepercut: A deeper, stronger, and faster multi-person pose estimation model. In *European Conference on Computer Vision*, pages 34–50. Springer, 2016.

[7] H. Jhuang, J. Gall, S. Zuffi, C. Schmid, and M. J. Black. Towards understanding action recognition. In *Proceedings of the IEEE international conference on computer vision*, pages 3192–3199, 2013.

[8] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, and P. Natsev. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017.

[9] D. E. King. Dlib-ml: A machine learning toolkit. *Journal of Machine Learning Research*, 10(Jul):1755–1758, 2009.

[10] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre. HMDB: a large video database for human motion recognition. In *2011 International Conference on Computer Vision*, pages 2556–2563. IEEE, 2011.

[11] H. Li, Z. Lin, X. Shen, J. Brandt, and G. Hua. A convolutional neural network cascade for face detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5325–5334, 2015.

[12] J. Li, Y. Wang, C. Wang, Y. Tai, J. Qian, J. Yang, C. Wang, J. Li, and F. Huang. Dsfd: dual shot face detector. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5060–5069, 2019.

[13] D. Ludl, T. Gulde, and C. Curio. Simple yet efficient real-time pose-based action recognition. *arXiv preprint arXiv:1904.09140*, 2019.

[14] A. Newell, K. Yang, and J. Deng. Stacked hourglass networks for human pose estimation. In *European conference on computer vision*, pages 483–499. Springer, 2016.

[15] H. Qin, J. Yan, X. Li, and X. Hu. Joint training of cascaded CNN for face detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3456–3465, 2016.

[16] J. Redmon and A. Farhadi. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*, 2018.

[17] K. Soomro, A. R. Zamir, and M. Shah. UCF101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012.

[18] K. Sun, B. Xiao, D. Liu, and J. Wang. Deep high-resolution representation learning for human pose estimation. *arXiv preprint arXiv:1902.09212*, 2019.

[19] B. Xiao, H. Wu, and Y. Wei. Simple baselines for human pose estimation and tracking. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 466–481, 2018.

[20] J. Yu, Y. Jiang, Z. Wang, Z. Cao, and T. Huang. Unitbox: An advanced object detection network. In *Proceedings of the 24th ACM international conference on Multimedia*, pages 516–520. ACM, 2016.

[21] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters*, 23(10):1499–1503, 2016.

[22] S. Zhang, X. Zhu, Z. Lei, H. Shi, X. Wang, and S. Z. Li. S3fd: Single shot scale-invariant face detector. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 192–201, 2017.

[23] M. Zolfaghari, G. L. Oliveira, N. Sedaghat, and T. Brox. Chained multi-stream networks exploiting pose, motion, and appearance for action classification and detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2904–2913, 2017.

[24] M. Zolfaghari, K. Singh, and T. Brox. Eco: Efficient convolutional network for online video understanding. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 695–712, 2018.