A Spatio-Temporal Identity Verification Method for Person-Action Instance Search in Movies

Yanrui Niu[1,2,3], Jingyao Yang[1,2,3], Chao Liang[1,2,3](✉), Baojin Huang[1,2,3], and Zhongyuan Wang[1,2,3]

[1] National Engineering Research Center for Multimedia Software (NERCMS), Wuhan, China
[2] Hubei Key Laboratory of Multimedia and Network Communication Engineering, Wuhan, China
[3] School of Computer Science, Wuhan University, Wuhan, China
cliang@whu.edu.cn

Abstract. As one of the challenging problems in video search, Person-Action Instance Search (P-A INS) aims to retrieve shots with a specific person carrying out a specific action from massive amounts of video shots. Most existing methods conduct person INS and action INS separately to compute the initial person and action ranking scores, which will be directly fused to generate the final ranking list. However, direct aggregation of two individual INS scores ignores spatial relationships of person and action, thus cannot guarantee their identity consistency and cause identity inconsistency problem (IIP). To address IIP, we propose a simple spatio-temporal identity verification method. Specifically, in the spatial dimension, we propose an identity consistency verification (ICV) step to revise the direct fusion score of person INS and action INS. Moreover, in the temporal dimension, we propose a double-temporal extension (DTE) operation to further improve P-A INS results. The proposed method is evaluated on the large-scale NIST TRECVID INS 2019–2021 tasks, and the experimental results show that it can effectively mitigate the IIP, and its performance surpasses that of the champion team in 2019 INS task and the second place teams in both 2020 and 2021 INS tasks.

Keywords: Person instance search · Action instance search · Double-temporal extension · Identity consistency verification

1 Introduction

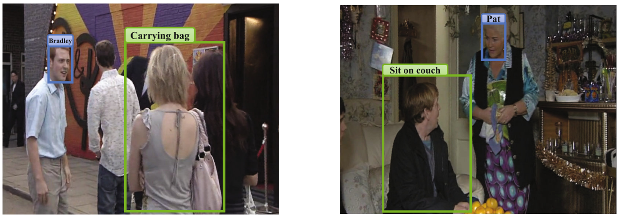
With the rapid development of multimedia technology in recent years, videos have flooded our life. Finding specific targets from massive amounts of videos,

Y. Niu and J. Yang—These authors contribute equally to this work.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/978-3-031-27077-2_7.

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023
D.-T. Dang-Nguyen et al. (Eds.): MMM 2023, LNCS 13833, pp. 82–94, 2023.
https://doi.org/10.1007/978-3-031-27077-2_7

i.e., video instance search (INS), is becoming increasingly important, and movies are suitable breeding grounds for it. Early INS research in movies mainly focuses on a single target, *i.e.*, single concept INS, such as finding a specific object [15,21], person [13,25,29], or action [7,17,26]. Recently, researchers started to investigate the more challenging combinatorial-semantic INS, which aims at retrieving specific instances with multiple attributes simultaneously. Representative works in this field include Person-Scene (P-S INS) [1,2,6] and Person-Action (P-A INS) [3–5]. The former aims at finding shots about the specific person in a specific scene, while the latter aims at finding shots about the specific person doing a specific action. In this paper we study the P-A INS in movies.



(a) "Bradley is carrying bag"

(b) "Pat is sitting on the couch"

Fig. 1. Examples of IIP in P-A INS. The blue and green boxes mark the target person and action, respectively. (Color figure online)

Existing methods [16,23,32] often adopt two different technical branches for person INS and action INS. Specifically, in the person INS branch, face detection and identification are conducted to compute ranking scores of video shots concerning the target person. In the action INS branch, the action recognition is conducted to compute ranking scores of video shots about the target action. Thereafter, two-branch INS scores are directly fused to generate the final ranking result. However, direct aggregation of scores cannot guarantee the identity consistency between person and action. For example, in Fig. 1(a), given "Bradley is standing" and "Danielle is carrying bag", the system [14] mistakes it as "Bradley is carrying bag" since the person "Bradley" and action "carrying bag" appear simultaneously; similar case happens in Fig. 1(b). We call it *identity inconsistency problem* (IIP).

To address the above problem, we propose a spatio-temporal identity verification method. In spatial dimension, we propose an identity consistency verification (ICV) scheme to compute the spatial consistency degree between face and action detection results. The higher spatial consistency degree means the larger overlapping area between the bounding boxes of face and action, thus the more likely that face and action belong to the same person. Furthermore, we find many face and action detection failures due to complex scenarios, such as non-frontal filming or object occlusion, hindering ICV from getting basic detection information. Considering the continuity of video frames in a shot and temporal continuity of some actions in adjacent shots, we propose a double-temporal extension (DTE) operation in the temporal dimension. The detection information of the interval frames is shared with intermediate frames through intra-shot DTE, and the fusion scores of adjacent shots are adjusted by inter-shot DTE.

The main contributions of this paper are as follows:

- We discover and study the IIP in the combinatorial P-A INS of movies. It shows that direct aggregation of single concept INS scores cannot always guarantee the identity consistency between person and action, which leads to the degraded performance of previous works.
- We propose a spatio-temporal identity verification method to address IIP, which uses ICV in the spatial dimension to check identity consistency between person and action, and DTE in the temporal dimension to share the detection information in successive frames in a shot and transferring the score information among adjacent shots.
- We verify the effectiveness of the proposed method on the large-scale TRECVID INS dataset. The performance surpasses the champion team in the 2019 INS task and the second place teams in both 2020 and 2021 INS tasks.

2 Related Work

2.1 Person INS

Person INS in videos aims to find shots containing a specific person from a video gallery, which is also termed as person re-identification. Most of the previous research working on person re-identification mainly focus on surveillance videos, where dresses rather than faces are more robust for identity discrimination [9,31,35]. But in movies, due to massive amounts of close-up shots and frequent clothing changes, faces are more stable than dresses for person re-identification. Therefore, most of the existing works in movies mainly use face detection and face recognition algorithms for person INS [13,25,29].

2.2 Action INS

Existing research on action INS mainly relies on action recognition or detection technology [14,16,22,23,32]. The difference between them is that the former only recognizes the category of action, whereas the latter can provide the location bounding boxes of action, and we focus on action detection. According to different implementation strategies, action detection can be generally divided into image-based and video-based methods. The former is mainly designed for actions with obvious interactive objects but without rigorous temporal causality, *e.g.*, “holding glass” and “carrying bag”. This corresponds to a specialized action detection task, *i.e.*, human-object interaction (HOI) detection [19,20,24], which aims to recognize the action (interaction) category, and meanwhile, locate human and object bounding boxes from images. The latter targets actions with rigorous temporal causality. *e.g.*, “open the door and enter” and “go up/down stairs”. Hence, it usually works on successive multiple video frames, and representative methods are [27,28,30].

2.3 Fusion Strategy

For combinatorial-semantic P-A INS, the difficulty lies in how to combine the results of different branches. Most of the existing studies adopt a strategy of retrieving two instances separately and then aggregating individual scores in some ways [14,16,22,23,32]. For example, NII fuses scores of person INS and action INS by direct weighted summation [16]. Instead, WHU adopts a stepwise strategy of searching for the action based on a candidate person list. It first builds an initial candidate person shot list with person INS scores, then sorts the list according to scores of action INS [14,32]. PKU adopts a strategy of searching for the person based on a candidate action list [22,23]. However, direct aggregation of person INS and action INS results without checking their identity consistency may incur serious IIP. To solve this problem, Le *et al.* [18] raises a heuristic method. They calculate the distance between the target face and desired object, and assume that the shorter distance means a more positive relationship between person and action with the desired object. The method indirectly judges the identity consistency by the distance between the related object and the target face, but can not sufficiently prove the identity consistency of the target face and

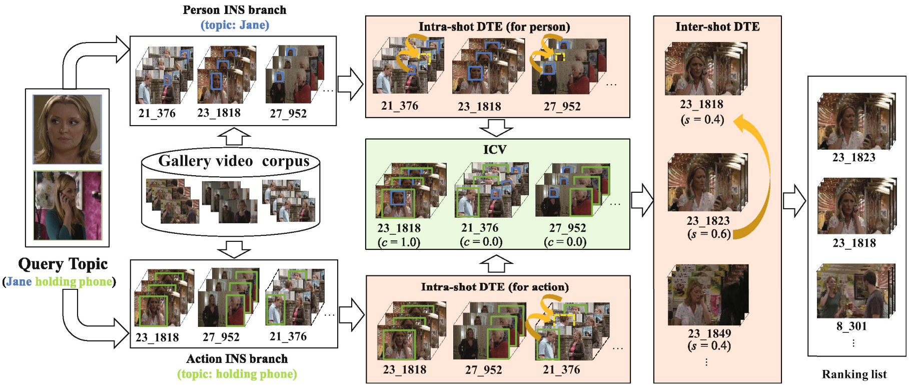


Fig. 2. The overall scheme of the spatio-temporal method for P-A INS. First, Person INS and action INS are conducted. Then intra-shot DTE recovers face and action detection information in the keyframes with detection failure. After that, ICV is conducted to filter out IIP shots. Then inter-shot DTE is used to adjust the final ranking of shots. At last, the ranking list is obtained by sorting all shots’ scores. The yellow dotted boxes represents the recovered boxes, the orange arrows represent the directions of interpolation operation, c means the consistency degree of person and action, and s means shot score. (Color figure online)

specific action. Moreover, it works based on object detection, which means that it does not work for actions without obvious interactive objects, *e.g.*, “walking” and “standing”.

Different from [18], we propose a spatio-temporal identity verification method for P-A INS, which can determine the identity consistency of the P-A pair without additional dependence on objects. Hence, it can be applied to both HOI and object-free actions.

3 Method

The overall scheme of our method is shown in Fig. 2. Given a topic and a video corpus, uniform sampling at an interval of 5 frames is first carried out to extract representative keyframes from shots of the video corpus. Then, person INS and action INS are conducted. Note that we apply detection in INS branches so we can obtain face/action detection scores as well as bounding boxes. Next, in the temporal dimension, intra-shot DTE is firstly conducted on failed detection shots, providing more detection information for ICV. Thereafter, in the spatial dimension, the ICV method is applied to check identity consistency between person and action, which filters out erroneous IIP shots. Finally, the maximum fusion score of all keyframes in a shot is taken as the INS score of the shot, and the inter-shot DTE is conducted to adjust the scores of shots, then the ranking list is obtained by sorting INS scores of all shots.

3.1 Preliminary

Assume that there are L shots in video gallery. For the l-th shot, K keyframes can be extracted. We denote the k-th keyframe in the shot l as $P^{(l,k)}$, where $l \in [1, L]$ and $k \in [1, K]$. For the convenience of the following discussion, the subscript signs k and l are temporarily omitted from all variables when they do not cause confusion.

For a keyframe P, assume that there are m faces and n actions detected in the person INS and action INS branches. The detection and identification results of i-th face can be expressed as $\langle ID_i, Conf_i, Box_i\rangle_{i=1}^{m}$, where ID_i represents the face id, $Conf_i$ records the confidence score of face identification, $Box_i = \langle x_{min_i}, y_{min_i}, x_{max_i}, y_{max_i}\rangle$ contains the horizontal and vertical coordinates of upper-left and lower-right corners of the face bounding box. Similarly, the result of j-th action can be expressed as $\langle ID_j, Conf_j, Box_j\rangle_{j=1}^{n}$, with similar notation definitions.

3.2 Identity Consistency Verification (ICV)

In order to address the IIP, we propose ICV to verify the identity consistency between person and action in the spatial dimension.

Specifically, for a keyframe P, we calculate spatial consistency degree matrix $\mathbf{C} = [c_{i,j}] \in \mathbb{R}^{m \times n}$ based on face and action bounding boxes obtained from person and action INS branches, in which $c_{i,j}$ is defined as:

$$c_{i,j} = \frac{\textbf{Intersection}\left(Box_i^{\text{face}}, Box_j^{\text{action}}\right)}{\textbf{Area}\left(Box_i^{\text{face}}\right)}, \tag{1}$$

where $\textbf{Intersection}(\cdot,\cdot)$ is the function of computing the intersection area of two bounding boxes, $\textbf{Area}(\cdot)$ is the function of computing the area of a bounding box.

Next, the proposed spatial consistency degree is applied to optimize the fusion score. Two representative fusion strategies are adopted.

- One simple strategy is the weighted fusion method ($Fusion_{wet}$) [14,16,32], which can be optimized as:

  $$s_{i,j} = c_{i,j} \times \left[\alpha \times Conf_i^{\text{face}} + (1-\alpha) \times Conf_j^{\text{action}}\right], \tag{2}$$

  where $s_{i,j}$ stands for the fusion score of the i-th person and the j-th action, $\alpha \in [0,1]$ is the fusion coefficient, which is a hyperparameter.
- The other effective fusion strategy, *i.e.*, searching for the specific action based on a candidate person list ($Fusion_{thd}$), is widely used [14,22,23,32]. It can be improved by the proposed spatial consistency degree as:

  $$s_{i,j} = c_{i,j} \times \left[\mathbf{F}_\delta\left(Conf_i^{\text{face}}\right) \times Conf_j^{\text{action}}\right], \tag{3}$$

  where $\mathbf{F}_\delta(\cdot)$ is a threshold function, δ is the threshold for face scores to determine whether the target person exists in the keyframes, *i.e.*, $\mathbf{F}_\delta(x) = 1$ if $x \geq \delta$, otherwise 0.

3.3 Double-Temporal Extension (DTE)

To address the detection failure problem caused by complex filming conditions, we propose DTE to transfer the information in the temporal dimension, which includes intra-shot DTE and inter-shot DTE.

intra-shot DTE shares the detection information among keyframes. We conduct the intra-shot DTE to recover face and action detection information in the keyframes with detection failure by linear interpolation. The shared detection information including confidence scores and coordinates of detection bounding boxes.

Inter-shot DTE shares the score information among shots. Because some actions have time continuity and can last more than one shot, the same query may appear in adjacent shots. Therefore, we adjust the final ranking of shots by diffusing the fusion scores of adjacent shots. The Gaussian curve is used to guide the score diffusion between shots with different distances:

$$\hat{s}_{i,j}^{l} = s_{i,j}^{l} + \sum_{-\gamma \leq d \leq \gamma} \boldsymbol{F}_{dis}(d) \times \textbf{max}\left(s_{i,j}^{l+d} - s_{i,j}^{l}, 0\right), \tag{4}$$

$$\boldsymbol{F}_{dis}(d) = \theta \cdot \frac{1}{\sqrt{2\pi}\sigma} exp\left(-\frac{d^2}{2\sigma^2}\right), \tag{5}$$

where $\hat{s}_{i,j}^l$ is the revised fusion score of i-th person conducting j-th action in the l-th shot after inter-shot DTE, $s_{i,j}^l$ is original fusion score, d is the distance between two shots, $\mathbf{max}(\cdot,\cdot)$ is used to limit diffusion direction, and $\boldsymbol{F}_{dis}(\cdot)$ is a distance based weight function, which decreases with the increase of shot distance. θ is used to adjust the contribution of distance, and σ is used to adjust the range of score diffusion, which determines the value of γ ($\gamma \approx 3 \cdot \sigma$).

3.4 Generating Ranking List

After obtaining fusion scores of all keyframes, the fusion score of the i-th person conducting the j-th action in l-th shot is the maximum score of keyframes in the shot:

$$s_{i,j}^l = \max_{k=1,\cdots,K} s_{i,j}^{(l,k)}. \tag{6}$$

Based on the fusion scores of all shots, we perform an inter-shot DTE in Sect. 3.3 to obtain the revised fusion scores. Then the ranking list concerning the topic of the i-th person conducting the j-th action is obtained by sorting the revised fusion scores of all shots. The complete flowchart of the proposed spatio-temporal identity verification method is presented in Algorithm D.1 in the supplementary material.

4 Experiments

4.1 Dataset and Evaluation Criteria

The TRECVID INS Dataset [5] comes from the 464-hour BBC soap opera "EastEnders", which is divided into 471,527 shots, containing about 7.84 million keyframes. NIST selects 70 topics based on it as representative samples for TRECVID 2019–2021 INS tasks [3–5]. The details of the dataset and topics are presented in Table A.1 and Table B.1-B.3 in the supplementary material.

According to the official evaluation criteria of TRECVID, Average Precision (AP) is adopted to evaluate the retrieval quality of each topic, and mean AP (mAP) is used to describe the overall performance among the given set of P-A INS topics. For each topic, only 1,000 shots at most can be evaluated.

4.2 Implementation Details

Person INS Branch. We adopt the RetinaFace detector [10] trained on the WIDER FACE [33] to obtain the face detection bounding boxes for each keyframe. and utilize the ArcFace [11] trained on the MS1Mv2 [11] to extracted 512-dimension features from normalized face images based on the detected face bounding boxes. Cosine similarity is used to calculate the face scores.

Action INS Branch. In the action INS branch, we especially apply two different action detection methods, *i.e.*, HOI detection on images and action detection on videos, according to different action characteristics. For topics with actions with obvious objects, we adopt PPDM [20] pre-trained on HICO-DET [8] (heatmap prediction network is DLA-34 [34]) to conduct HOI detection on images. For topics with actions lasting for a long time, we adopt ACAM [28] to conduct action detection on videos, which is trained on the AVA dataset [12].

Fusion Strategy. We test the effect of the parameters α and δ of two fusion methods, *i.e.*, $Fusion_{wet}$ and $Fusion_{thd}$, and compare the best performance of them. As shown in Figure C.1 in the supplementary material, $Fusion_{thd}$ is better than $Fusion_{wet}$, so we choose $Fusion_{thd}$ in the baseline model.

Double-Temporal Extension. We get the best parameters (refer to Figure C.2 in the supplementary material), where $\theta = 3$ and $\sigma = 5$.

4.3 Ablation Study

In this section, we evaluate the effectiveness of DTE and ICV on the NIST TRECVID 2019–2021 INS tasks.

Table 1. Ablation study results on NIST TRECVID 2019–2021 INS tasks. The black bold values mark the best value for each column (%).

Method			**2019**			**2020**			**2021**		
Base	***DTE***	***ICV***	**P-A$_i$**	**P- A$_v$**	**P- A**	**P-A$_i$**	**P-A$_v$**	**P-A**	**P-A$_i$**	**P-A$_v$**	**P-A**
✓			29.34	3.94	20.51	35.16	4.37	21.69	39.63	8.69	30.35
✓	✓		31.68	4.12	22.10	37.31	5.49	23.39	43.33	8.59	32.91
✓		✓	31.57	4.48	22.15	37.09	5.52	23.28	43.36	**9.54**	33.21
✓	✓	✓	**34.13**	**4.73**	**23.90**	**39.28**	**6.89**	**25.11**	**47.70**	9.27	**36.17**

Base. We construct a baseline model referred to as Base by eliminating all proposed methods. Specifically, in the Base model, the face and action scores are fused with $Fusion_{thd}$ to get scores of keyframes. Thereafter, the maximum score of keyframes is taken as the shot score. Finally, the ranking list is obtained by sorting the shot scores for each topic.

Then we add DTE and ICV gradually to Base. Note that we have two P-A INS combination methods since we adopt two action detection methods in the action INS branch, *i.e.*, image-based P-A INS (P-A$_\mathrm{i}$ INS) and video-based P-A INS (P-A$_\mathrm{v}$ INS). Table 1 shows ablation study results in 2019–2021 INS tasks. The mAP of topics corresponding to P-A$_\mathrm{i}$ and P-A$_\mathrm{v}$ columns are computed respectively, and the mAP of all topics is shown in the final P-A column.

Evaluation of DTE. We add DTE to Base, referred to as Base+DTE. In 2019 INS task, Base+DTE gains 1.59% (7.75% relative growth) improvement over the Base method. Similarly, in 2020 and 2021 INS tasks, the improvements are 1.70% (7.84% relative growth) and 2.56% (8.43% relative growth), which confirms the effectiveness of DTE.

Evaluation of ICV. We add ICV to Base, referred to as Base+ICV, which gains 1.64% (8.00% relative growth), 1.59% (7.33% relative growth), and 2.86% (9.42% relative growth) improvements over Base in 2019–2021 INS tasks respectively, confirming the effectiveness of ICV.

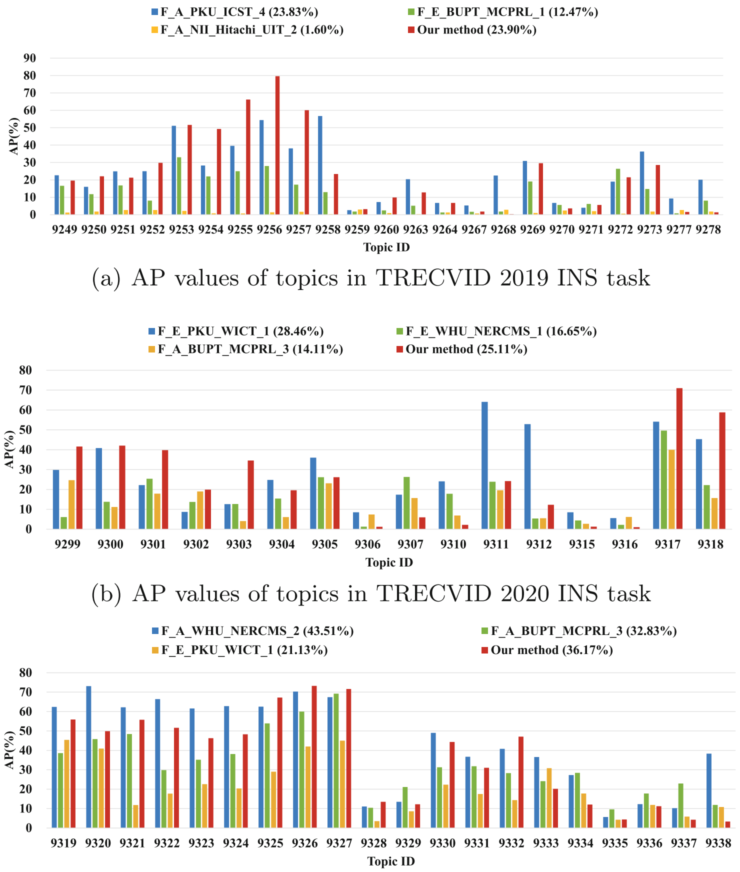
Evaluation of DTE and ICV. Furthermore, Base+DTE+ICV achieves the best performance in both experiments, which gains 3.39% (16.53% relative growth), 3.42% (15.77% relative growth), and 5.82% (19.18% relative growth) improvements over Base in 2019–2021 INS tasks. It can be seen that with the proposed method, the mAPs of P-$\mathrm{A_i}$ INS and P-$\mathrm{A_v}$ INS both improve, which proves the effectiveness of the proposed method is consistent, and the method works for both P-A INS branches on images and videos.

The visualization results of DTE and ICV are shown in Figure E.1 and E.2 in the supplementary material.

4.4 Comparison with Other Methods

We compare the proposed method with state-of-the-art methods on NIST TREC- VID 2019–2021 INS tasks. According to the official evaluation settings, each team is allowed to submit several runs for evaluation, and we select the best run of each top-3 team for comparison, whose details are shown in Section F in the supplementary material.

Figure 3 demonstrates the comparative results of our method and previous evaluation runs. As shown in Fig. 3(a), our method achieves the best performance on 10 topics and competitive performance on 9 topics. In Fig. 3(b), our method achieves 7 best and 4 competitive performance. And in Fig. 3(c), our method achieves 5 best and 11 competitive performance. The performance is relatively poor on other topics. Through observing the results of three-year INS tasks, we find that the reason for those relatively poor-performance topics is due to detection errors of some difficult action topics. For example, the actions in topics *9268, 9278, 9315* and *9316* are all “go up or down stairs”, the actions in topics *9267, 9277, 9335* and *9336* are all “open the door and enter/leave”, and the actions in topics *9306, 9307, 9337* and *9338* are all “holding cloth”. It can be seen that the difficulty of action INS is an important factor limiting the performance of P-A INS. In general, we propose a simple INS method, compared with other methods with many tricks, our method still gets considerable performance. The mAP of our methods surpassed the state-of-the-art in 2019 INS task and the best runs of second place in 2020–2021 INS tasks.



(a) AP values of topics in TRECVID 2019 INS task

(b) AP values of topics in TRECVID 2020 INS task

(c) AP values of topics in TRECVID 2021 INS task

Fig. 3. Comparisons with other P-A INS methods. The legend shows the mAP values. Blue, green and orange represents the best run of the first place, second place and third place team of INS 2019–2021 tasks, while red represents our method. (Color figure online)

5 Conclusion

We study the IIP between person and action in P-A INS in movies, and propose a simple but effective spatio-temporal identity verification method. The experimental results of our method on the large-scale TRECVID INS dataset verify its effectiveness and robustness. In the future, we will concentrate on improving the accuracy of identity verification by trying more methods, such as using other appearance-based features within the bounding boxes to infer identity consistency, or using human posture information to locate the face position in the action bounding boxes. And we will extend our method to more combinatorial-semantic INS tasks, e.g., the Person-Action-Scene INS.

Acknowledgement. This work is supported by the National Natural Science Foundation of China (No. U1903214, 61876135). The numerical calculations in this paper have been done on the supercomputing system in the Supercomputing Center of Wuhan University.

References

1. Awad, G., et al.: TRECVID 2018: Benchmarking video activity detection, video captioning and matching, video storytelling linking and video search. In: Proceedings of TRECVID 2018 (2018)
2. Awad, G., et al.: TRECVID 2017: evaluating ad-hoc and instance video search, events detection, video captioning, and hyperlinking. In: TREC Video Retrieval Evaluation (TRECVID) (2017)
3. Awad, G., et al.: TRECVID 2020: a comprehensive campaign for evaluating video retrieval tasks across multiple application domains. In: Proceedings of TRECVID 2020 (2020)
4. Awad, G., et al.: Evaluating multiple video understanding and retrieval tasks at TRECVID 2021. In: Proceedings of TRECVID 2021 (2021)
5. Awad, G., et al.: Trecvid 2019: an evaluation campaign to benchmark video activity detection, video captioning and matching, and video search retrieval. In: Proceedings of TRECVID 2019 (2019)
6. Awad, G., et al.: TRECVID 2016: evaluating video search, video event detection, localization, and hyperlinking. In: TREC Video Retrieval Evaluation (TRECVID) (2016)
7. Bojanowski, P., Bach, F., Laptev, I., Ponce, J., Schmid, C., Sivic, J.: Finding actors and actions in movies. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV), pp. 2280–2287 (2013). https://doi.org/10.1109/ICCV.2013.283
8. Chao, Y.W., Liu, Y., Liu, X., Zeng, H., Deng, J.: Learning to detect human-object interactions. In: 2018 IEEE Winter Conference on Applications of Computer Vision (WACV), pp. 381–389 (2018). https://doi.org/10.1109/WACV.2018.00048
9. Chen, L., Yang, H., Xu, Q., Gao, Z.: Harmonious attention network for person re-identification via complementarity between groups and individuals. Neurocomputing **453**, 766–776 (2021). https://doi.org/10.1016/j.neucom.2020.07.118
10. Deng, J., Guo, J., Ververas, E., Kotsia, I., Zafeiriou, S.: Retinaface: single-shot multi-level face localisation in the wild. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2020). https://doi.org/10.1109/CVPR42600.2020.00525
11. Deng, J., Guo, J., Xue, N., Zafeiriou, S.: Arcface: additive angular margin loss for deep face recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2019). https://doi.org/10.1109/CVPR.2019.00482
12. Gu, C., et al.: AVA: a video dataset of spatio-temporally localized atomic visual actions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 6047–6056 (2018). https://doi.org/10.1109/CVPR.2018.00633
13. Haq, I.U., Muhammad, K., Ullah, A., Baik, S.W.: Deepstar: Detecting starring characters in movies. IEEE Access **7**, 9265–9272 (2019). https://doi.org/10.1109/ACCESS.2018.2890560

14. Jiang, L., et al.: Whu-nercms at trecvid 2019: Instance search task. In: Proceedings of TRECVID Workshop (2019). https://www-nlpir.nist.gov/projects/tvpubs/tv19.papers/whu_nercms.pdf
15. Jiang, W., Wu, Y., Jing, C., Yu, T., Jia, Y.: Unsupervised deep quantization for object instance search. Neurocomputing **362**, 60–71 (2019). https://doi.org/10.1016/j.neucom.2019.06.088
16. Klinkigt, M., et al.: Nii hitachi uit at trecvid 2019. In: Proceedings of TRECVID Workshop (2019). https://www-nlpir.nist.gov/projects/tvpubs/tv19.papers/nii_hitachi_uit.pdf
17. Laptev, I., Perez, P.: Retrieving actions in movies. In: 2007 IEEE 11th International Conference on Computer Vision (ICCV), pp. 1–8 (2007). https://doi.org/10.1109/ICCV.2007.4409105
18. Le, D.D., et al.: Nii_uit at trecvid 2020. In: Proceedings of TRECVID Workshop (2020). https://www-nlpir.nist.gov/projects/tvpubs/tv20.papers/nii_uit.pdf
19. Li, Y.L., et al.: Transferable interactiveness knowledge for human-object interaction detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3585–3594 (2019). https://doi.org/10.1109/CVPR.2019.00370
20. Liao, Y., Liu, S., Wang, F., Chen, Y., Qian, C., Feng, J.: PPDM: Parallel point detection and matching for real-time human-object interaction detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 482–490 (2020). https://doi.org/10.1109/CVPR42600.2020.00056
21. Meng, J., Yuan, J., Yang, J., Wang, G., Tan, Y.P.: Object instance search in videos via spatio-temporal trajectory discovery. IEEE Trans. Multimedia **18**(1), 116–127 (2016). https://doi.org/10.1109/TMM.2015.2500734
22. Peng, Y., et al.: PKU-ICST at TRECVID 2019: Instance search task. In: Proceedings of TRECVID Workshop (2019). https://www-nlpir.nist.gov/projects/tvpubs/tv19.papers/pku-icst.pdf
23. Peng, Y., Ye, Z., Zhang, J., Sun, H.: PKU WICT at TRECVID 2020: Instance search task. In: Proceedings of TRECVID Workshop (2020). https://www-nlpir.nist.gov/projects/tvpubs/tv20.papers/pku-wict.pdf
24. Qi, S., Wang, W., Jia, B., Shen, J., Zhu, S.C.: Learning human-object interactions by graph parsing neural networks. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 401–417 (2018). https://doi.org/10.1007/978-3-030-01240-3_25
25. Kumar, N., Du, V., Doja, M.N., Shambharkar, P., Nimesh, U.K.: Automatic Face Recognition and Finding Occurrence of Actors in Movies. In: Ranganathan, G., Chen, J., Rocha, Álvaro. (eds.) Inventive Communication and Computational Technologies. LNNS, vol. 145, pp. 115–129. Springer, Singapore (2021). https://doi.org/10.1007/978-981-15-7345-3_10
26. Stoian, A., Ferecatu, M., Benois-Pineau, J., Crucianu, M.: Fast action localization in large-scale video archives. In: IEEE Trans. Cir. and Sys. for Video Technol. **26**(10), 1917–1930 (2016). https://doi.org/10.1109/TCSVT.2015.2475835
27. Tang, J., Xia, J., Mu, X., Pang, B., Lu, C.: Asynchronous Interaction Aggregation for Action Detection. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) ECCV 2020. LNCS, vol. 12360, pp. 71–87. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58555-6_5

28. Ulutan, O., Rallapalli, S., Srivatsa, M., Torres, C., Manjunath, B.S.: Actor conditioned attention maps for video action detection. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), pp. 527–536 (2020). https://doi.org/10.1109/WACV45572.2020.9093617
29. Wang, X., Liu, W., Chen, J., Wang, X., Yan, C., Mei, T.: Listen, look, and find the one: robust person search with multimodality index. ACM Trans. Multimedia Comput. Commun. Appl. 16(2) (2020). https://doi.org/10.1145/3380549
30. Wu, C.Y., Feichtenhofer, C., Fan, H., He, K., Krahenbuhl, P., Girshick, R.: Long-term feature banks for detailed video understanding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 284–293 (2019). https://doi.org/10.1109/CVPR.2019.00037
31. Yang, F., Yan, K., Lu, S., Jia, H., Xie, D., Yu, Z., Guo, X., Huang, F., Gao, W.: Part-aware progressive unsupervised domain adaptation for person re-identification. IEEE Trans. Multimedia **23**, 1681–1695 (2021). https://doi.org/10.1109/TMM.2020.3001522
32. Yang, J., Kang'an Chen, Y.N., Fan, X., Liang, C.: WHU-NERCMS at TRECVID 2020: Instance search task. In: Proceedings of TRECVID Workshop (2020). https://www-nlpir.nist.gov/projects/tvpubs/tv20.papers/whu_nercms.pdf
33. Yang, S., Luo, P., Loy, C.C., Tang, X.: Wider face: A face detection benchmark. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 5525–5533 (2016). https://doi.org/10.1109/CVPR.2016.596
34. Yu, F., Wang, D., Shelhamer, E., Darrell, T.: Deep layer aggregation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2403–2412 (2018). https://doi.org/10.1109/CVPR.2018.00255
35. Zhang, W., Wei, Z., Huang, L., Xie, K., Qin, Q.: Adaptive attention-aware network for unsupervised person re-identification. Neurocomputing **411**, 20–31 (2020). https://doi.org/10.1016/j.neucom.2020.05.094