FAST ACTION LOCALIZATION BASED ON SPATIO-TEMPORAL PATH SEARCH

Qingtian Wu[1] *Huiwen Guo*[1] *Xinyu Wu*[1,2,*] *Yimin Zhou*[1] *Nannan Li*[3]

[1]Shenzhen College of Advanced Technology, University of Chinese Academy of Sciences
[2]Department of Mechanical and Automation, the Chinese University of Hong Kong
[3]Peking University Shenzhen Graduate School, Shenzhen, P.R.China

ABSTRACT

In this paper, a method is proposed to search for spatio-temporal path for action localization in unconstrained videos. We mainly focus on two requirements, i.e., accurate human extraction and speeding generation of action proposal. The approach first generates human proposals at the frame level, then scores them based on two complementary parts, i.e., posteriori probability evaluated via a fine-tuned Faster-RCNN and template-matching similarity based on the spatio-temporal continuity. Finally, the generation of action proposal is formulated as a Max-Path discovery problem, coupled with dynamic programming to find an optimal path with maximum score. Experiments on UCF-Sports are performed to verify that the proposed method can achieve fast high-quality action proposal and link the missed-detection proposals in successive frames together to form a complete action.

Index Terms— Action localization, Faster-RCNN, Spatio-temporal path, Dynamic programming.

1. INTRODUCTION

Video analysis is quite significant in a wide range of advanced technology applications, e.g., automatic driving system, human-machine interaction and smart homes, especially video surveillance and monitoring (VSAM) in crowded scene. The wide applications and domanial challenges make it become a hot research field in the computer vision.

Action recognition [1] [2] is a common task in video analysis, the purpose of which is to assign a class label to a video clap. Moreover, action detection [3] [4] is an advanced task for it not only should identify the action type, but also localize the actor(s) spatially and temporally, i.e., action localization. Recent progress in action localization has advanced the frontiers of this problem in these aspects, e.g., feature representations [5–10], models [11–13], general schemas [14–20].

Various features have been studied recently. Motivated by the success of CNN (Convolutional Neural Network), CNN features have gained satisfying performance with combining appearance features and motion cues [5] [6]. Other features such as bag-of-words [7] [8] on spatio-temporal descriptors have gained comparable performance to CNN representation. Dense trajectories are adopted to generate action proposal with satesfying performance [9] [10].

Figure-centric model has been applied to action localization. A spatio-temporal model is developed in [11] via a figure-centric visual word representation, where the location of the subject is treated as a latent variable. In [12], humans and objects are detected and modeled by their interaction. In [13], a weakly supervised model based on multiple instance learning is proposed to slide the spatial-temporal subvolumes for action detection.

An traditional schema for action localization is handled by the sliding window approach. To avoid the exhaustive search of the target subvolume, spatio-temporal branch-and-bound search [15] has been proposed to detect human actions in videos. A linear complexity search algorithm has been proposed in [16]. The segmentation-and-merging stragety is also widely used for generating action tubes [17] [18]. In [19], Nan proposes to generate action proposal by using actionness estimation and evaluate discriminative scores at local patches, then the search step is considered as a maximum score problem. Hough voting based approach can be applied for action localization with acceptable performance [20].

The purpose of the paper is to generate high-quality action proposal with spatial extent and temporal continuity. Not only can human proposal extraction operate with high accuracy, it can also generate action proposal with high speed so that future work such as fast action recognition or semantic understanding will be continued. The remainder of the paper is organized as follows. After describing the proposed method in Section 2, Section 3 presents the experiment results on UCF-Sports and make comparisons to analyse the performance of the proposed method. Conclusion is given in Section 4.

2. ACTION LOCALIZATION

In Fig. 1, it shows the overview of the proposed method which takes video clip as input and generates the spatio-temporal path for action localization. Our method mainly consists of three stages which are described as follows.

This work presented in the paper is partially supported by National Natural Science Foundation of China (61473277). $*$ indicates the corresponding author (email:{qt.wu, hw.guo, xy.wu, ym.zhou}@siat.ac.cn)

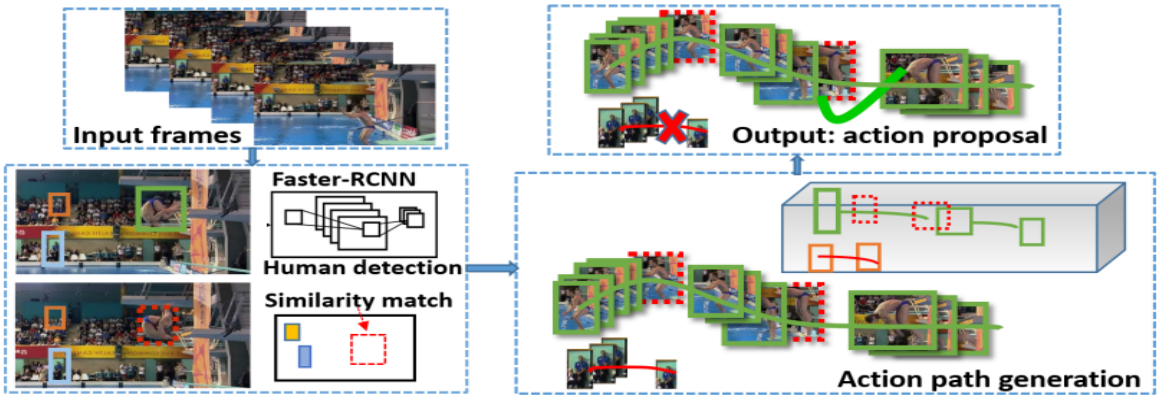


Fig. 1: Overview of the proposed approach for action localization

2.1. Frame-level Human Proposals via the Faster-RCNN

Human proposal which constitutes the ground work of action proposal, includes two parts: human detected by the Faster-RCNN (region-based CNN) and similarity matching based on the spatio-temporal continuity. In the sliding-window based methods, human detection can be achieved by scanning each frame in a fixed step, accompanied with huge computational cost and lots of redundancies. Since Faster-RCNN is regarded as an end-to-end deep-learning network from region proposal to recognition, a RPN (Region Proposal Network) is introduced to reduce the proposal set to regions that are most likely to contain an object or a person. We adopt the novel network instead of the time-consuming sliding window approach to detect human for it can achieve a satisfying detection accuracy and nearly real-time speed with the GPU (Graphic Processing Unit) acceleration.

As for training the Faster-RCNN, a VGG-16 model pre-trained on the ILSVRC dataset [21] is employed for its parameter initialization. However, due to the rapidly-changing appearance of the actor in action like diving and lifting, the network cannot detect them well. To meet the demand of multi-pose human detection, a more effective network is fine-tuned by adding extra data through Nan's approach [19], i.e., rotating the human-label data from PASCAL 2007 in a fixed step at 45° from 0° to 360°. A comparison of human detection results from the initial Faster-RCNN and our fine-tuned model is shown in Fig. 2. It can be seen that there is a missing detection in the middle column from the initial network while there are more overlapping in the bounding boxes with the ground truth from our fine-tuned model.

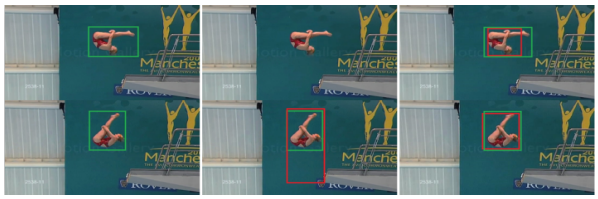


Fig. 2: Comparison of human detection results from the initial Faster R-CNN (*middle*) and our fine-tuned model (*right*). The bounding boxes in green and red are the ground truth (*left*) and the human detection results respectively.

2.2. Scoring Proposals for Tracking

Since missed detections are inevitable in detection task or there are even several missed detections in successive frames, it is difficult to address these missed proposals and link them together to a complete action or tell whether the action comes to an end. To solve the problem, we adopt the template-matching method based on correlation coefficient to trigger the missed detection proposals in successive frames due to spatio-temporal continuity. An evaluation function is designed to score each proposal for searching action path. The function which considers the confidence coefficient of a bounding box $H_d(\cdot)$ detected by the Faster-RCNN and the correlation coefficient of a bounding box $H_m(\cdot)$ matched by the template-matching method, is defined as:

$$S(b_t^i) = \alpha \times \sigma(H_d(b_t^i)) + \beta \times \sigma(H_m(b_t^i)) \qquad (1)$$

where b_t^i is the i^{th} detection bounding box on the t^{th} frame; α and β are the constraints to activate the human estimation or similarity matching, i.e., $\alpha + \beta = 1$; $\sigma(\cdot) = 1/(1 + e^{-x})$ normalizes the likelihood into [0, 1]. It should be noted that there is a trade-off between the optimal matching and time consumption, so a near-optimal matching result is acceptable when the $M(b_t^i) \geq \eta$, where η is the defined threshold. An example of scoring metric for tracking is shown in Fig. 3.

2.3. Problem Formulation

Given an unconstrained video clip, the purpose is to generate a generic action proposal which can represent the action well and determine the beginning and ending of a complete action automatically. The action proposal $P = \{b_{t_s}^i, b_{t_s+1}^i, \cdots, b_{t_e}^i\}$ corresponds to a trajectory linking the bounding box b_t^i from

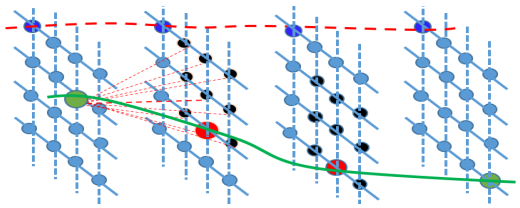


Fig. 3: Example of the scoring metric for tracking in 4 frames. Each node represents a discriminative region: the blue one is a human proposal representing an audience; the green one is the actor detected by the Faster-RCNN; the red one is the matching proposal by 3×3 local-neighbor searching.

the t_s-th frame to the t_e-th frame. As an action is characterized by a spatio-temporal path, the path generation can be formulated as Max-Path coverage problem:

$$\begin{aligned} \max_{\mathbf{P}\in\Phi} & \sum_{b_t\in\cup p_i} S(b_t) + \lambda \times \sum_{b_t\in\cup p_i} M(b_t) \\ s.t. \quad & \mathbf{P} \leq N \\ & \mathbf{O}(p_i, p_j) \leq \delta_p, \forall p_i, p_j \in \mathbf{P}, i \neq j \end{aligned} \tag{2}$$

where Φ is a set of action path candidates; $S(b_t)$ is the discriminative actionness score of a bounding box b_t in Eq.(1); $M(b_t)$ is the motion score of b_t; λ is the coefficient for balancing the human detection score and motion score, here $\lambda = 0$ (The reason will be explained later); δ_p is the defined threshold. The first constraint sets the maximum number of paths contained in $\mathbf{P}$ while the second one is to avoid redundant path generation. The overlapping of the two paths is defined as:

$$\mathbf{O}(p_i, p_j) = \frac{\sum_{\max(t_s^i, t_s^j) \leq t \leq \min(t_e^i, t_e^j)} o(b_t^i, b_t^j)}{max(t_e^i, t_e^j) - min(t_s^i, t_s^j)} \tag{3}$$

where $o(b_t^i, b_t^j)$ is defined as $\frac{\cap(b_t^i, b_t^j)}{\cup(b_t^i, b_t^j)}$, indicating the IoU (Intersection over Union) of the two bounding boxes b_t^i and b_t^j.

2.4. Action Path Generation for Localization

To solve the Max-Path discovery problem with low complexity, the method used in [16] is applied with message propagation mechanism. The proposed algorithm includes two steps: forward propagation to record the best path and backward tracking to guarantee the beginning and ending of an action.

A solution for the Max-Path search problem is given,

$$\Omega(b_t^i) = \begin{cases} -\infty, & t > e \\ S(b_t^i), & t = e \\ \max_{b\in I_t} \Omega(b_{t-1}^i) + S(b_t^i), & t < e \end{cases} \tag{4}$$

where $\Omega(b_t^i)$ represents the accumulated scores of the best path starting from the t^{th} frame to the final location (b^i, e). Apart from obtaining the matrix Ω in linear form with the dynamic programming (which has been proven in [16]), the algorithm also stores all possible action paths. When Ω is solved, the algorithm can trace back the best path with all possible starting b_s^i and ending locations b_e^i. Here, the following constraints are satisfied, i.e., $S(b_s^i) > \delta_i$ and $S(b_t^i) < \delta_o$, where δ is the defined threshold.

As the pseudo-code shown in Algorithm 1, it takes bounding box score $S(b_t)$ as the input, and generates the best action path P^* with the ending location l^* . The localization of the best path can be directly obtained by searching the values stored in P and tracing back until reaching a null box.

Algorithm 1: Forward Message Propagation

Input: $S(b_t)$: the local score of the bounding box;
Output: $\Omega(b_e^i)$: the accumulated scores of the optimal path leading to (b^i, e);
$P^*(b_e^i)$: the optimal path record for tracking back;
S^*: the accumulated score of the best path;
l^*: the ending location of the best path;

```
Initialize : Ω(b^i, 1) = S(b_1), P*(b_e^i) = null, ∀(b^i, e);
S* = −∞, l* = null;
for t = 2 → n do
    for b ∈ I_t do
        b_i ← argmax Ω(b_{t−1});
        if Ω(b_{t−1}^i) > 0 then
            Ω(b_t^i) ← Ω(b_{t−1}^i) + S(b_t^i);
            P*(b_t^i) ← (b^i, t − 1);
        else
            Ω(b_t^i) ← S(b_t^i);
        if Ω(b_t^i) > S* then
            S* ← Ω(b_t^i);
            l* ← (b^i, t);
return Ω, P, S*, l*;
```

3. EXPERIMENTS

We experimented on the UCF-Sports. Comparisons with the state of the art are evaluated to verity the efficiency of the proposed method. All experiments are run on a computer with 16GB memory, a 3.5GHz CPU and an Tesla-K80.

3.1. Datasets and Evaluation

UCF-Sports The dataset contains 150 short videos belonging to 10 action categories. Since all frames are annotated with bounding boxes and videos are truncated to contain a single action, it has been widely used for action localization.

Evaluation metrics We evaluate the efficiency of the proposed method by referring to the metric in [6] [19]. The mean IoU between the action proposal p and the ground-truth g is defined as $IoU(p, g) = \frac{1}{|\mathbf{C}|} \sum_{t\in\mathbf{C}} o(p_t, g_t)$, where $|\mathbf{C}|$ is the frame set including p and g; p_t and g_t are the detection box and the ground-truth on the t^{th} frame; $o(\cdot)$ is predefined in

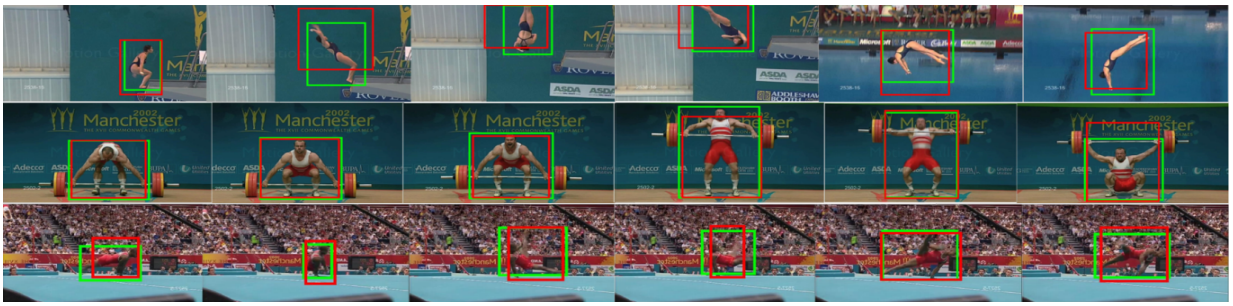


Fig. 4: Examples of the action generation results on UCF-Sports. The bounding boxes in green and red are the ground truth and the action proposal respectively.

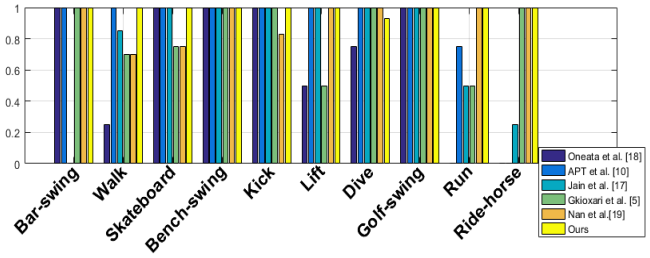
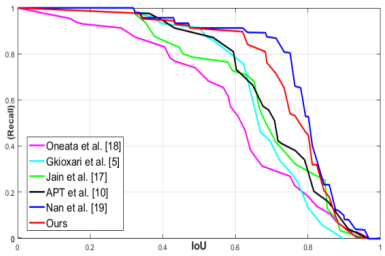


Fig. 5: Recall performance on each action types on UCF-Sports compared with the state of the art.



Method	Recall (%)	Speed (s/f)
[18]	68.09	-
[17]	78.72	-
[5]	87.23	-
[10]	89.36	-
[19]	91.49	1.69
Ours	**90.23**	**0.3**

(Table 1)

Fig. 6: UCF-Sports: Recall(%) per IoUs threshold (*Left*); Recall performance comparison when $\eta(IoU) = 0.5$ (*right*).

Eq. 3. An action proposal is considered positive if the $IoU(\boldsymbol{P}, \boldsymbol{G}) \geq \eta$, where η is the predefined threshold, here, $\eta = 0.5$.

3.2. Experimental Results and Analysis

Recall performance on each action class on UCF-Sports is compared with different methods, as shown in Fig. 5, from which it can be seen that the proposed method can achieve conspicuous recall performance on almost all the action types except for Diving. Since the body posture of the actor changes dramatically during the diving, it is effective to combine appearance features and motion cues for human estimation in [19]. We adopt template matching based on spatio-temporal continuity to address these missed-detection cases and finally link them together to a complete action. Fig. 4 presents several examples of more accurate action localization via our fine-tuned Faster-RCNN. We also evaluate the recall performance under different IoU thresholds η ranging from 0 to 1, as shown in Fig. 6 from which it can be concluded that when η is above 0.6, the recall of other methods indicate a rapid downward trend except for Nan's method [19] and ours.

Real-time capability is important in practice, e.g., the VSAM system. Since there is a trade-off between the recall performance and action extraction speed, relatively high recall performance is acceptable when the extraction speed is nearly real-time. Table 1 presents the recall(%) performance comparison when $\eta = 0.5$. The proposed approach gains a recall of 90.23% more than 0.87% higher than that in [10]. Although the method [19] achieve state-of- the-art performance, it takes them 1.69 seconds per frame (s/f) on average while our method spend 0.3 s/f, almost 5 times faster than that of Nan's method when $\lambda = 0$. The break-down of time consumption in our method is assessed as: it takes almost 0.2 second to generate human proposal on each frame by the Faster-RCNN and about 10 millisecond to run template matching when missed detection occurs; the average time spent on forward-backward searching is nearly 0.09 second. In summary, the total time spent on each frame is 0.3 second.

4. CONCLUSION

A novel approach for action localization has been presented in the paper. Effective human proposal is extracted via human detection with a fine-tuned Faster-RCNN and template matching. The action path searching is formulated as a Max-Path discovery problem and solved via dynamic programming. Experiments on UCF-Sports have shown that the proposed approach can generate more accurate action proposal and link missed-detection proposal in successive frames together to a complete action. Time-tag recursive neural network will be discussed for fast action recognition in the future.

Acknowledgement This work also thanks to Key Laboratory of Human-Machine Intelligence Synergic Systems, Shenzhen Institutes of Advanced Technology (SIAT), Chinese Academy of Sciences (CAS) & Guangdong Provincial Key Laboratory of Robotics and Intelligent System, SIAT, CAS.

5. REFERENCES

[1] Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei, "Large-scale video classification with convolutional neural networks," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2014, pp. 1725–1732.

[2] Ronald Poppe, "A survey on vision-based human action recognition," *Image and vision computing*, vol. 28, no. 6, pp. 976–990, 2010.

[3] Bharat Singh, Tim K. Marks, Michael Jones, Oncel Tuzel, and Ming Shao, "A multi-stream bi-directional recurrent neural network for fine-grained action detection," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 1961–1970.

[4] Serena Yeung, Olga Russakovsky, Greg Mori, and Fei Fei Li, "End-to-end learning of action detection from frame glimpses in videos," *Computer Science*, 2015.

[5] Georgia Gkioxari and Jitendra Malik, "Finding action tubes," pp. 759–768, 2015.

[6] Philippe Weinzaepfel, Zaid Harchaoui, and Cordelia Schmid, "Learning to track for spatio-temporal action localization," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 3164–3172.

[7] Ivan Laptev, Marcin Marszalek, Cordelia Schmid, and Benjamin Rozenfeld, "Learning realistic human actions from movies," in *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*. IEEE, 2008, pp. 1–8.

[8] Heng Wang, Alexander Kläser, Cordelia Schmid, and Cheng-Lin Liu, "Dense trajectories and motion boundary descriptors for action recognition," *International journal of computer vision*, vol. 103, no. 1, pp. 60–79, 2013.

[9] Heng Wang, Dan Oneata, Jakob Verbeek, and Cordelia Schmid, "A robust and efficient video representation for action recognition," *International Journal of Computer Vision*, pp. 1–20, 2015.

[10] Jan C Van Gemert, Mihir Jain, Ella Gati, and Cees G M Snoek, "Apt: Action localization proposals from dense trajectories," in *BMVC*, 2015.

[11] Tian Lan, Yang Wang, and Greg Mori, "Discriminative figure-centric models for joint action localization and recognition," in *2011 International Conference on Computer Vision*. IEEE, 2011, pp. 2003–2010.

[12] Alessandro Prest, Vittorio Ferrari, and Cordelia Schmid, "Explicit modeling of human-object interactions in realistic videos," *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 4, pp. 835–848, 2013.

[13] Parthipan Siva and Tao Xiang, "Weakly supervised action detection.," in *BMVC*, 2011, vol. 2, p. 6.

[14] Ivan Laptev and Patrick Perez, "Retrieving actions in movies," in *IEEE International Conference on Computer Vision, ICCV 2007, Rio De Janeiro, Brazil, October*, 2007, pp. 1–8.

[15] Junsong Yuan, Zicheng Liu, and Ying Wu, "Discriminative video pattern search for efficient action detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 9, pp. 1728–1743, 2011.

[16] Du Tran, Junsong Yuan, and David Forsyth, "Video event detection: From subvolume localization to spatiotemporal path search," *IEEE transactions on pattern analysis and machine intelligence*, vol. 36, no. 2, pp. 404–416, 2014.

[17] M Jain, J Van Gemert, H Jegou, and P Bouthemy, "Action localization with tubelets from motion," in *Computer Vision and Pattern Recognition*, 2014, pp. 740–747.

[18] Oneata Dan, Jerome Revaud, Jakob Verbeek, and Cordelia Schmid, *Spatio-temporal Object Detection Proposals*, Springer International Publishing, 2014.

[19] Nannan Li, Dan Xu, Zhenqiang Ying, Zhihao Li, and Ge Li, "Searching action proposals via spatial actionness estimation and temporal path inference and tracking," *arXiv preprint arXiv:1608.06495*, 2016.

[20] Gang Yu, Junsong Yuan, and Zicheng Liu, "Propagative hough voting for human activity recognition," in *European Conference on Computer Vision*. Springer, 2012, pp. 693–706.

[21] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, and Michael Bernstein, "Imagenet large scale visual recognition challenge," *International Journal of Computer Vision*, vol. 115, no. 3, pp. 211–252, 2015.