Neurocomputing 333 (2019) 351–363

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Action recognition and localization with spatial and temporal contexts

Wanru Xu[a,b,*], Zhenjiang Miao[a,b], Jian Yu[c], Qiang Ji[d]

[a] Institute of Information Science, Beijing Jiaotong University, China
[b] Beijing Key Laboratory of Advanced Information Science and Network Technology, China
[c] Department of Computer Science, Beijing Jiaotong University, China
[d] Department of Electrical & Computer Engineering, Rensselaer Polytechnic Institute, USA

ARTICLE INFO

Article history:
Received 16 June 2018
Revised 27 December 2018
Accepted 3 January 2019
Available online 9 January 2019

Communicated by Dr Zhiyong Wang

Keywords:
Action recognition
Action localization
Dynamic graphical model
Spatio-temporal contexts

ABSTRACT

Locating human action in spatio-temporal domain among untrimmed videos is an important but challenging task. Recent works have shown that incorporating contextual information leads to a significant improvement in action recognition, but there is still no existing work taking full advantage of context for action localization. While the popular target-centered methods have achieved promising results, they fail to exploit contexts and capture temporal dynamics in actions. In this paper, we propose a principled dynamic model, called spatio-temporal context model (STCM), to simultaneously locate and recognize actions. The STCM integrates various kinds of contexts, including the temporal context that consists of the sequences before and after action as well as the spatial context in the surrounding of target. Meanwhile, a novel dynamic programming approach is introduced to accumulate evidences collected at a small set of candidates in order to detect the spatio-temporal location of action effectively and efficiently. We report encouraging results on the UCF-Sports and UCF-101. It demonstrates that the contextual information is not only helpful for action recognition, but also contributes to action localization.

© 2019 Elsevier B.V. All rights reserved.

1. Introduction

Nowadays, since the multimedia data grows explosively, especially for video data, automatic multimedia analysis becomes more and more important. Among them, it is more meaningful to understand "what he is doing" or "where and when he does it" for human beings. Under this case, vision-based human action recognition and localization have played an important role in multimedia content analysis and they also have contributed to other multimedia applications. Human action analysis is the most active research areas in computer vision and machine learning with many applications [1,2] such as video content-based retrieval, human-computer interaction and video-surveillance.

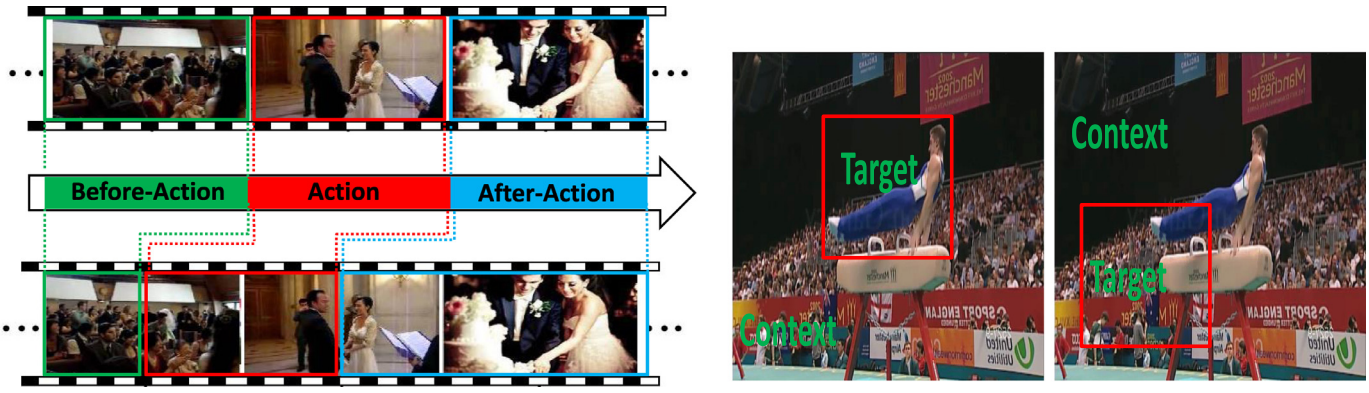
Recently, there have been a considerable amount of works focusing on action recognition [3–5], whose aim is to assign a class label for each entire video sequence. However, given a video sequence, action occur at a precise spatio-temporal extent and it is also desirable to detect the spatio-temporal location in real world. Action localization has attracted increasing attention [6,7], which involves three questions: (1) *recognition*: what the action is; (2) *temporal localization*: when it starts and ends; (3) *spatial localization*: where it happens. It is a more challenging problem than action recognition task, due to the large intra-class variations and in particular the very large spatio-temporal search space. In this paper, we propose a unified framework to model human action for both recognition and localization tasks.

Human action does not occur in a vacuum, and there also exists spatio-temporal contexts in the video sequence, which have close relevance to the action. For example, recognizing the competition venue scene of a gymnastics will provide useful evidence for the action analysis in Fig. 1. As the same, it is helpful to recognize the wedding action, if holding ceremony and cutting cake in church are recognized before and after the wedding action. Recently, there are a large number of works on discovering the relations between actions and contexts to improve the performance of action recognition. Actually, spatio-temporal contexts are not only important for action recognition, but also useful for action localization. However, there is still no existing work taking full advantage of context to help detect the accurate spatio-temporal boundary of action.

For action localization, it is challenging to detect the accurate boundary and incorporating action and its context can alleviate

* Corresponding author at: Institute of Information Science, Beijing Jiaotong University, China.
E-mail addresses: xuwanru@bjtu.edu.cn, 11112063@bjtu.edu.cn (W. Xu), zjmiao@bjtu.edu.cn (Z. Miao), jianyu@bjtu.edu.cn (J. Yu), qji@ecse.rpi.edu (Q. Ji).

https://doi.org/10.1016/j.neucom.2019.01.008
0925-2312/© 2019 Elsevier B.V. All rights reserved.



(a) Temporal localization. (b) Spatial localization.

Fig. 1. Action spatio-temporal localization. (a) The top one is the accurate temporal localization and the bottom one is inaccurate. (b) The left one is the accurate spatial localization and the right one is inaccurate. Inaccurate localization can not only affect the action itself (target), but also has negative effect on its spatio-temporal contexts.

this issue. The reason is that the boundary is just the intersection between action itself and its context, and if we change one, the other one also changes accordingly as shown in Fig. 1. As explained earlier, if the detected bounding box is smaller than the ground-truth, some part of target is lost and this region will be mistakenly considered as the part of context at the same time. Although they can provide double cues for accurate localization, there is no existing method utilizing both target and context to detect actions. In addition, most action localization methods are inspired by 2D object localization, where they usually apply an action classifier directly on a large number of spatio-temporal candidates. There are three limitations: (1) The dynamic information in action is totally ignored in many frame-level based methods [8–10], which are direct extensions of object localization approaches. Generally, they first build frame-level detector taking advantage of 2D object detection algorithm and then apply some temporal post-processing to get the final video-level detection result. (2) The search space is too large for the 3D action localization, especially considering a flexible bounding box size where it can vary across frames. The approach with flexible bounding box can be utilized to detect both the moving action and the non-moving action, while those fixed bounding box based approaches (eg. sub-volume) cannot handle action with strong moving. (3) It has to repeat the same procedure for each individual action class separately which is impractical for datasets with large numbers of action categories and it also requires a large amounts of negative samples for each of them. To address these issues, we propose a dynamic spatio-temporal context model (STCM) to simultaneously locate and recognize human actions.

Overall, the contributions of this paper are as follows: First, we integrate spatial and temporal contexts into the interpretation process by constructing a unified STCM, which can be utilized for both action recognition and action localization. Thus action and its context are incorporated to complement each other to enhance recognition capability as well as to refine each other to get more accurate boundary. Meanwhile, no negative sample is required in this paper, since the context is also considered as the positive sample. Second, we propose a video-level based method to locate actions using the trained dynamic model and take full advantage of temporal independence between frames, where the location of action in one frame is not only decided by current frame, but also affected by previous frames and previous detections. Third, we develop a dynamic programming with saliency map framework that can find more accurate action location without significantly increasing the number of candidates.

The rest of this paper is organized as follows: Section 2 focuses on related works. Section 3 provides an overall summary of our method and formulate the action recognition and localization problem. Section 4 details the structure of this novel spatio-temporal context model with the training and inference algorithms. The process of spatial and temporal localization using the STCM is given in Section 5. Section 6 reports experiment results using two human action datasets for recognition and localization tasks. Section 7 concludes this paper.

2. Related works

Depending on the techniques, human action localization approaches can be divided into three categories: proposal and classification framework, segmentation based method and deep learning based method. Meanwhile, we review some context models for action recognition.

Proposal + Classification framework. Human action localization is commonly approached by spatio-temporal proposals matching, that is the classical proposal + classification framework. The main idea is first to generate a large number of spatio-temporal candidates and then score them to find the final detection result using classifier. Traditionally, proposals are generated by a sliding window based approach [11–13] which is effective but not efficient. In [14], temporal sliding window is used to generate several candidates and then train a SVM to match them by combining motion and appearance features. Approximately normalized Fisher vector [15] is proposed to represent actions and then a sliding window scheme is applied for action localization, which yields significant improvements in the computational cost and memory of the FV. Also several efforts aim at reducing the computational cost which need to evaluate during detection process, such as the spatio-temporal branch-and-bound algorithm [11]. In [16], actions are treated as spatio-temporal patterns extracted by naive Bayes based mutual information maximization (NBMIM) and a novel search algorithm is introduced to find the optimal sub-volume in the 3D video space for detecting action efficiently. Another limitation with these sliding-window based methods is that it cannot handle the moving actions, since the detected action is usually captured by a video sub-volume. Besides the sub-volume detection, spatio-temporal action tubes can be detected using structured output

regression [17] with a max-path search. A series of spatio-temporal video tubes [18] are considered as localization candidates and generated using a greedy method by computing actionness score. Similarily, the spatio-temporal tube (ST-tube) is employed for action localization in [19] with a one order Markov model by recursively infering the action regions at consecutive frames. A novel PSDF descriptor [20] is computed for temporal action localization, which is a intuitive descriptor for action class, position and duration. For classification part, any action recognition models can be adopted as an evaluator to estimate these extracted proposals. On one hand, classification performance can be improved by improving representative power, such as a novel representation on the intrinsic shape manifold learned by graph embedding algorithm is proposed in [21]. On another hand, classification performance can be also improved by improving discriminative power, such as a probabilistic framework based on Gaussian processes [22] is proposed for providing an estimation of uncertainty.

Segmentation based method. Segmentation based method can be treated as a special proposal + classification method, which generates proposals using segmentation techniques. In [23], action proposals are 2D+t sequences of bounding boxes, called tubelets and they are generated by hierarchically merging super-voxels. A hierarchical MRF model [24] is proposed to segment human action boundaries in videos in-the-wild automatically, which bridging low-level fragments with high-level motion and appearance. The new hierarchical space-time segment [25] is considered as a representation for action recognition and localization, which can preserve their hierarchical and temporal relationships. In [26], 2D deformable part model is extended to 3D spatio-temporal DPM for action localization, where the most discriminative 3D subvolumes are selected as parts and their spatio-temporal relations are learned. Similarly, a relational model [8] is proposed for action localization, which first decomposes human action into temporal "key poses" and then further into spatial "action parts". In [27], the location is treated as a latent variable which is inferred with action recognition simultaneously using discriminative figure-centric model.

Deep learning. One line of these works is to extract spatio-temporal feature representations for building strong classifiers using deep models. In [9], they first detect frame-level proposals and score them using a combination of static and motion CNN features. Then they track proposals with high score in the video using a tracking-by-detection approach. Similarly in [28], an approach is proposed for action localization using convolutional neural networks on static and kinematic cues. Three segment-based 3D ConvNets [29] are adopted for temporal action localization in untrimmed long videos, including a proposal network, a classification network and a localization network. A novel architecture called UntrimmedNet is presented in [6] for weakly supervised action recognition and detection. It consists of a classification module and a selection module, where the first module is to learn the action model and the last one is to reason about the temporal duration of action respectively. A multi-region two-stream R-CNN model [10] is introduced to locate action in realistic videos, which starting from frame-level action detection based on faster R-CNN [30] and then linking frame-level detections to obtain video-level detections. In [7], a novel action detection pipeline is introduced, which incorporates a very deep region proposal network (RPN) like as Fast R-CNN and merges appearance and motion cues by a novel fusion strategy. However, such indirect methods are unsatisfying in terms of both computation efficiency as well as accuracy, since they fully ignore the temporal dynamic in human action. Another line of these works is the end-to-end detection. A fully end-to-end approach [31] is introduced to detect action in videos learning to directly predict the temporal boundaries of actions. The model is formulated as a recurrent neural network-based agent to interact with a video over time, where the deep reinforcement learning is used.

Context model for action recognition. Recently, the advantage of combining context for human action recognition has been fully confirmed by many works and various contextual elements have been considered including spatial context [32–34] and temporal context [35,36]. A r*CNN [32] is introduced to use more than one region to construct a strong action recognition system. A context-augmented video event recognition approach [37] is proposed to capture three levels of contexts from spatial to temporal, including image level, semantic level, and prior level. In [36], a temporal embedding is learned for complex video analysis by associating frames with the temporal context. A new active learning technique is formulated in [33] which not only exploits the informativeness of the individual action instances but also utilizes their contextual information among the actions and objects. In [38], a framework for the recognition of collective human actions is proposed, which can automatically capture relevant crowd context with a 3D Markov Random Field. In [39], a novel deep action- and context-aware sequence learning is presented for action recognition and anticipation to effectively combine both context-aware and action-aware features by a multi-stage recurrent architecture. A two-graph model [40] is constructed to represent human actions by modeling the spatial and temporal relationships among local features, and also a novel family of context-dependent graph kernels is proposed to measure similarity between graphs. A recurrent interactional context modeling scheme based on LSTM network [41] is proposed to model high order interactional context and a unified interactional feature modeling process is introduced for one-person dynamics, intra-group and inter-group interactions. Recently, it has been proved that combining target action with its context can significantly improve performance of human action recognition. However, current video-based action localization approaches are still almost target-centered which fully ignore these important contextual information. To our best knowledge, this is the first method to jointly model action and its spatio-temporal contexts for human action localization using the dynamic graphical model.

3. Overview of the approach

The graphical representation of the spatio-temporal context model proposed in this paper is depicted in Fig. 2. For the temporal domain, we integrate sufficient temporal contextual information into the interpretation process by simultaneously modeling the before-action, after-action and action itself. The before-action and after-action represent the sequences before or after the target action happens. For the spatial domain, we incorporate all information of the whole frame, including target part and context part to enhance model descriptive capability, and capture relations between human action and its surrounding context to improve model discriminative power. The target is denoted by the region within the bounding box corresponding to human, while the context is the region outside the bounding box.

The STCM is a discriminative model which directly estimates the probability of output conditioned on the observation. Actually, it is a three-layer probabilistic graphical model including input observation layer $\mathbf{x}$, intermediate hidden layer $\mathbf{h}$ and output label layer y. There are many intra-class variations in both target actions and contexts, namely an action or a context may involve many different intermediate states. Therefore, a set of hidden variables $\mathbf{h}$ are introduced to capture these variations and model complex dependencies among observations. This conditional probabilistic model can be formulated as:

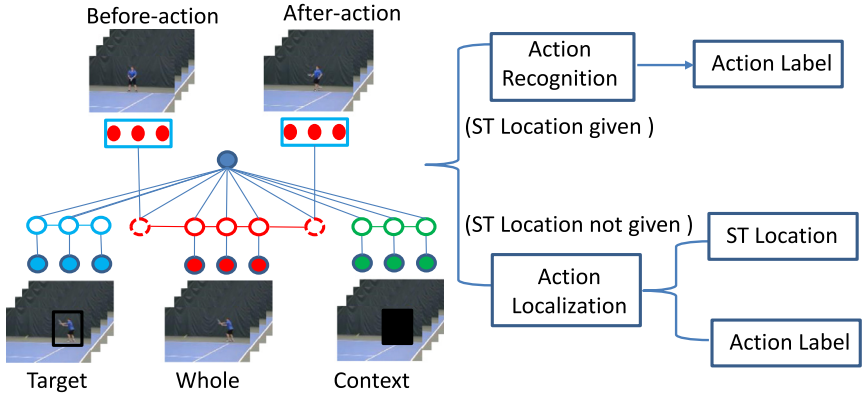


Fig. 2. The overview of our framework and the graphical representation of spatio-temporal context model (STCM), which integrates various kinds of contexts. Solid nodes represent observed variables, and hollow nodes denote as unobservable hidden variables. Compared to action recognition, action localization is a more challenging task, since it not only requires to recognize the action category, but also requires to detect the spatio-temporal location.

$$
\begin{aligned}
llP(y|\mathbf{x};\boldsymbol{\Theta}) &= \frac{P(y,\mathbf{x};\boldsymbol{\Theta})}{P(\mathbf{x};\boldsymbol{\Theta})} = \frac{P(y,\mathbf{x};\boldsymbol{\Theta})}{\sum_{\hat{y}} P(\hat{y},\mathbf{x};\boldsymbol{\Theta})} \\
&= \frac{\frac{1}{Z}\sum_{\mathbf{h}\in H}\exp(-\boldsymbol{\Theta}\cdot E(y,\mathbf{h},\mathbf{x}))}{\frac{1}{Z}\sum_{\hat{y}}\sum_{\mathbf{h}\in H}\exp(-\boldsymbol{\Theta}\cdot E(\hat{y},\mathbf{h},\mathbf{x}))} \\
&= \frac{1}{Z(\mathbf{x})}\sum_{\mathbf{h}\in H}\exp(-\boldsymbol{\Theta}\cdot E(y,\mathbf{h},\mathbf{x})).
\end{aligned} \tag{1}
$$

The partition function can be defined as:

$$
Z = \sum_{\mathbf{x},\mathbf{h},y}\exp(-\boldsymbol{\Theta}\cdot E(y,\mathbf{h},\mathbf{x})), \tag{2}
$$

playing a role of normalization to enable it to become a probability measure.

$$
Z(\mathbf{x}) = \sum_{\mathbf{h}\in H,\hat{y}\in Y}\exp(-\boldsymbol{\Theta}\cdot E(\hat{y},\mathbf{h},\mathbf{x})) \tag{3}
$$

Different from Eq. (2), Eq. (3) is also a normalization term, which introduces a different value for different sample, and later we will explain how to handle it to make learning and inference tractable. The $E(y, \mathbf{h}, \mathbf{x})$ represents the energy function, which can model various relations among variables.

As shown in Fig. 2, given a video $\mathbf{X} = \{\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_T\}$ with T frames, the goal of action localization is to recognize the class label $\hat{y}$ and detect a smooth spatio-temporal path $\hat{\mathbf{B}} = \{b_t\}_{t=t_1}^{t=t_2}$, namely a series of bounding boxes between the start frame t_1 and the end frame t_2 and each of them $b_t = \{x_1^t, y_1^t, x_2^t, y_2^t\}$ is represented by the coordinates of its two corners:

$$
(\hat{y},\hat{\mathbf{B}}) = \arg\max_{y,B} P(y,\mathbf{B}|\mathbf{X};\boldsymbol{\Theta}) = \arg\max_{y,B} P(y|\mathbf{X}(\mathbf{B});\boldsymbol{\Theta}). \tag{4}
$$

Given $\hat{\mathbf{B}}$ and $\mathbf{x} = \mathbf{X}(\hat{\mathbf{B}})$, it converts to a standard action recognition problem:

$$
\hat{y} = \arg\max_{y} P(y,\hat{\mathbf{B}}|\mathbf{X};\boldsymbol{\Theta}) = \arg\max_{y} P(y|\mathbf{X}(\hat{\mathbf{B}});\boldsymbol{\Theta}). \tag{5}
$$

Important notations in STCM are summarized as follows.

T: the number of frames in an action video.

$\mathbf{X}$: video sequence $\mathbf{X} = \{\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_T\}$, and each element denotes frame.

$\mathbf{x}$: complete observation of an action video sequence, $\mathbf{x} = \{\mathbf{x}^a, \mathbf{x}^b, \mathbf{x}^w, \mathbf{x}^t, \mathbf{x}^c\}$.

$\mathbf{x}^b, \mathbf{x}^a$: observations of before action and after action.

$\mathbf{x}^w$: observations of whole frame, $\mathbf{x}^w = \{\mathbf{x}_1^w, \mathbf{x}_2^w, \ldots, \mathbf{x}_T^w\}$.

$\mathbf{x}^t$: observations of target, $\mathbf{x}^t = \{\mathbf{x}_1^t, \mathbf{x}_2^t, \ldots, \mathbf{x}_T^t\}$.

$\mathbf{x}^c$: observations of context, $\mathbf{x}^c = \{\mathbf{x}_1^c, \mathbf{x}_2^c, \ldots, \mathbf{x}_T^c\}$.

$y, \hat{y}$: the ground-truth and predicted action label.

$\mathbf{h}$: complete hidden variables of STCM model, $\mathbf{h} = \{h^a, h^b, \mathbf{h}^w, \mathbf{h}^t, \mathbf{h}^c\}$.

h^b, h^a: hidden variables of before action and after action.

$\mathbf{h}^w$: hidden variables of whole frame, $\mathbf{h}^w = \{\mathbf{h}_1^w, \ldots, \mathbf{h}_T^w\}$.

$\mathbf{h}^t$: hidden variables of target, $\mathbf{h}^t = \{\mathbf{h}_1^t, \mathbf{h}_2^t, \ldots, \mathbf{h}_T^t\}$.

$\mathbf{h}^c$: hidden variables of context, $\mathbf{h}^c = \{\mathbf{h}_1^c, \mathbf{h}_2^c, \ldots, \mathbf{h}_T^c\}$.

$\hat{\mathbf{B}}$: the detected spatio-temporal path $\hat{\mathbf{B}} = \{b_t\}_{t=t_1}^{t=t_2}$, and each element denotes detected bounding box at each frame.

$\boldsymbol{\Theta}$: the parameters of STCM model.

4. Spatio-temporal context model

In this section, we first introduce the structure of STCM as well as all kinds of contextual information we can utilize for action recognition and localization. Then we present training and inference based on the max-margin criterion.

4.1. STCM

Actions always occur in the three-dimensional space during a period of time. In spatial domain, action is not independent from its surrounding environment; In temporal domain, action is also not isolated from its before and after event sequences. Therefore, a novel spatiao-temporal context model is proposed to fully capture the temporal context and the spatial context simultaneously. Considering the structure of STCM in Fig. 2 (Left), it can be formulated as,

$$
\begin{aligned}
P(y|\mathbf{x};\boldsymbol{\Theta}) &= \frac{1}{Z(\mathbf{x})}\sum_{\mathbf{h}\in H}\exp(-\boldsymbol{\Theta}\cdot E(y,\mathbf{h},\mathbf{x})) \\
&= \frac{1}{Z(\mathbf{x})}\sum_{\mathbf{h}\in H}\exp\{-\boldsymbol{\Theta}\cdot(E(y,\mathbf{h},\mathbf{x}^t)+E(y,\mathbf{h},\mathbf{x}^c) \\
&\quad + E(y,\mathbf{h},\mathbf{x}^w)+E(y,\mathbf{h},\mathbf{x}^b)+E(y,\mathbf{h},\mathbf{x}^a))\}.
\end{aligned} \tag{6}
$$

It consists several parts: the whole frame modeling; the before-action and after-action modeling; and the context and target modeling. The first is the most fundamental one and the last two are introduced for modeling temporal and spatial contexts, respectively.

The whole frame modeling can improve the recognition capability by integrating all the information both in target region and in context region:

$$
E(y,\mathbf{h},\mathbf{x}^w) = \sum_{t=1}^{T}E(h_t^w,\mathbf{x}_t^w) + \sum_{t=1}^{T}E(y,h_t^w) + \sum_{t=2}^{T}E(y,h_{t-1}^w,h_t^w). \tag{7}
$$

There are two unary functions to evaluate compatibilities between two variables:

$$E(h_t^w, \mathbf{x}_t^w) = -\sum_i \mathbf{x}_t^w \cdot \mathbb{1}\{h_t^w = h_i\}; \tag{8}$$

$$E(y, h_t^w) = -\sum_{i,j} \mathbb{1}\{h_t^w = h_i\} \cdot \mathbb{1}\{y = y_j\}. \tag{9}$$

A pairwise function taking some structural information into account and describes the compatibility of variables belonging to a three-wise connected clique:

$$E(y, h_{t-1}^w, h_t^w) = -\sum_{i,j,k} \mathbb{1}\{h_{t-1}^w = h_i\} \cdot \mathbb{1}\{h_t^w = h_k\} \cdot \mathbb{1}\{y = y_j\} \tag{10}$$

The before-action and after-action modeling play a role of temporal localization which are sensitive to temporal boundary:

$$E(y, \mathbf{h}, \mathbf{x}^b) = E(h^b, \mathbf{x}^b) + E(y, h^b) + E(y, h^b, h_1^w); \tag{11}$$

$$E(y, \mathbf{h}, \mathbf{x}^a) = E(h^a, \mathbf{x}^a) + E(y, h^a) + E(y, h^a, h_T^w). \tag{12}$$

where the unary and pairwise functions are calculated in the same way. These three parts $E(y, \mathbf{h}, \mathbf{x}^w)$, $E(y, \mathbf{h}, \mathbf{x}^b)$, $E(y, \mathbf{h}, \mathbf{x}^a)$ capture the complete temporal dynamics in actions, where before-action is the start node and after-action is the end node. It utilizes rich sequentially contextual information to better capture the appearance evolution and temporal structure of the full video.

The context and target modeling play a role of spatial localization which are sensitive to spatial boundary and complement each other:

$$E(y, \mathbf{h}, \mathbf{x}^c) = \sum_{t=1}^{T} E\left(h_t^c, \mathbf{x}_t^c\right) + \sum_{t=1}^{T} E(y, h_t^c) + \sum_{t=2}^{T} E\left(y, h_{t-1}^c, h_t^c\right); \tag{13}$$

$$E(y, \mathbf{h}, \mathbf{x}^t) = \sum_{t=1}^{T} E\left(h_t^t, \mathbf{x}_t^t\right) + \sum_{t=1}^{T} E\left(y, h_t^t\right) + \sum_{t=2}^{T} E\left(y, h_{t-1}^t, h_t^t\right). \tag{14}$$

The formulation is similar as the whole frame modeling and the only difference is the modeling objects of Eq. (7), Eqs. (14) and (13) are the whole frame, target part and context part respectively. Either considering context and target as a whole to model together or modeling them separately is unreasonable, where both lead to lose part of information. Therefore, our STCM is proposed to not only separate the target from the context by the context and target modeling, but also combine them to discover their latent relations by the whole frame modeling. Given a video $\mathbf{X}$ and a spatio-temporal path $\mathbf{B}$, there exists corresponding observations: $\mathbf{x}^b = \mathbf{X}(t_1)$ and $\mathbf{x}^a = \mathbf{X}(t_2)$ are the descriptors of before action and after action; $\mathbf{x}^w = \mathbf{X}(t_1, t_2)$ is extracted in the whole frame; $\mathbf{x}^t = \mathbf{X}(\mathbf{b})$ is extracted inside bounding boxes and $\mathbf{x}^c = \mathbf{X}(\sim \mathbf{b})$ is extracted outside bounding boxes.

4.2. Learning with max-margin criterion

We learn the spatio-temporal context model based on the max-margin criterion [42], which achieves significant success in machine learning especially for classification and detection tasks. From the viewpoint of probability, the margin is defined as the difference between the log-probability of the ground-truth assignment y_i and that of the "second best" assignment.

$$\begin{aligned}\delta(i, \boldsymbol{\Theta}) &= \log p(y_i|\mathbf{x_i}; \boldsymbol{\Theta}) - \max_{y \neq y_i} \log p(y|\mathbf{x_i}; \boldsymbol{\Theta}) \\ &= \log \frac{\sum_{\mathbf{h} \in H} \exp(-\boldsymbol{\Theta} \cdot E(y_i, \mathbf{h}, \mathbf{x}))}{\sum_{\mathbf{h} \in H, \hat{y} \in Y} \exp(-\boldsymbol{\Theta} \cdot E(\hat{y}, \mathbf{h}, \mathbf{x}))} \\ &\quad - \max_{y \neq y_i} \log \frac{\sum_{\mathbf{h} \in H} \exp(-\boldsymbol{\Theta} \cdot E(y, \mathbf{h}, \mathbf{x}))}{\sum_{\mathbf{h} \in H, \hat{y} \in Y} \exp(-\boldsymbol{\Theta} \cdot E(\hat{y}, \mathbf{h}, \mathbf{x}))} \\ &= \log \sum_{\mathbf{h}} \exp(-\boldsymbol{\Theta} \cdot E(y, \mathbf{h}, \mathbf{x})) - \log \sum_{\mathbf{h}, \hat{y}} \exp(-\boldsymbol{\Theta} \cdot E(\hat{y}, \mathbf{h}, \mathbf{x})) \\ &\quad - \max_{y \neq y_i} \log \sum_{\mathbf{h}} \exp(-\boldsymbol{\Theta} \cdot E(y, \mathbf{h}, \mathbf{x})) + \log \sum_{\mathbf{h}, \hat{y}} \exp(-\boldsymbol{\Theta} \cdot E(\hat{y}, \mathbf{h}, \mathbf{x})) \\ &= \log \sum_{\mathbf{h}} \exp(-\boldsymbol{\Theta} \cdot E(y_i, \mathbf{h}, \mathbf{x})) - \max_{y \neq y_i} \log \sum_{\mathbf{h}} \exp(-\boldsymbol{\Theta} \cdot E(y, \mathbf{h}, \mathbf{x})).\end{aligned} \tag{15}$$

Such that the partition function can be removed due to the same partition value for each sample. However, it is still intractable to compute because of the $\log \Sigma$, which is required to traverse all possible hidden configurations. Therefore, if we replace Σ with *max*, Eq. (15) can be simplified as:

$$\begin{aligned}\delta(i, \boldsymbol{\Theta}) &\approx \hat{\delta}(i, \boldsymbol{\Theta}) \\ &= \log \max_{\mathbf{h} \in H} \exp(-\boldsymbol{\Theta} \cdot E(y_i, \mathbf{h}, \mathbf{x})) \\ &\quad - \max_{y \neq y_i} \log \max_{\mathbf{h} \in H} \exp(-\boldsymbol{\Theta} \cdot E(y, \mathbf{h}, \mathbf{x})) \\ &= \max_{\mathbf{h} \in H} (-\boldsymbol{\Theta} \cdot E(y_i, \mathbf{h}, \mathbf{x})) - \max_{y \neq y_i, \mathbf{h} \in H} (-\boldsymbol{\Theta} \cdot E(y, \mathbf{h}, \mathbf{x})).\end{aligned} \tag{16}$$

The *log-sum-exp* function of h in Eq. (15) is so called *soft-max*, while the *max* operator in Eq. (16) is commonly called *max*-function. We introduce a "temperature" parameter followed [43,44] to smooth between *soft-max* and *max*, which motivates a more general objective function.

$$\varepsilon \log \sum_{\mathbf{h}} \exp\left(\frac{\boldsymbol{\Theta} \cdot E(y_i, \mathbf{h}, \mathbf{x})}{-\varepsilon}\right) - \max_{y \neq y_i} \varepsilon \log \sum_{\mathbf{h}} \exp\left(\frac{\boldsymbol{\Theta} \cdot E(y, \mathbf{h}, \mathbf{x})}{-\varepsilon}\right) \tag{17}$$

where ε is a temperature parameter to control how much uncertainty we want account for h. Note that this temperature parameter can be just considered as a constant scaling factor, thus it cannot change the inference result. This general function includes a number of existing methods as special cases. It reduces to the maximum likelihood framework in Eq. (15) if $\varepsilon = 1$, while $\varepsilon \to 0^+$ results in the max-margin formulation in Eq. (16). Therefore, $\varepsilon \to 0^+$ smoothly approximates the *soft-max* via the *max*-function.

This approximation has two benefits: First, it avoids exhaustively enumerating every hidden configuration; Second, it still preserves semantic information. In fact, Eq. (16) can be considered as the difference between the log-probability of the ground-truth assignment y_i with its best hidden configuration and that of the "second best" assignment, namely,

$$\hat{\delta}(i, \boldsymbol{\Theta}) = \max_{\mathbf{h} \in H} \log p(y_i, \mathbf{h}|\mathbf{x_i}; \boldsymbol{\Theta}) - \max_{y \neq y_i, \mathbf{h} \in H} \log p(y, \mathbf{h}|\mathbf{x_i}; \boldsymbol{\Theta}). \tag{18}$$

The $\max_{\mathbf{h} \in H} p(y_i, \mathbf{h}|\mathbf{x_i}; \boldsymbol{\Theta})$ can be treated as $p(y_i|\mathbf{x_i}; \boldsymbol{\Theta})$ with best hidden configuration, that maximizing with the latent variables. While $\sum_{\mathbf{h} \in H} p(y_i, \mathbf{h}|\mathbf{x_i}; \boldsymbol{\Theta})$ can be considered as combining the contributions of the various possible values by marginalizing over the latent variables.

The margin-based estimation methods usually aim to maximize the margin and increase the log-probability gap as much as possible. Because the larger is the margin, the more confident the model is to select y_i. Given training videos $D = \{\mathbf{x}_i, y_i\}_{i=1}^N$, the hinge loss function is defined as:

$$\sum_{i=1}^{N} \max(0, \Delta(y_i, y) + \max_{y \neq y_i} \log p(y|\mathbf{x_i}; \boldsymbol{\Theta}) - \log p(y_i|\mathbf{x_i}; \boldsymbol{\Theta})). \tag{19}$$

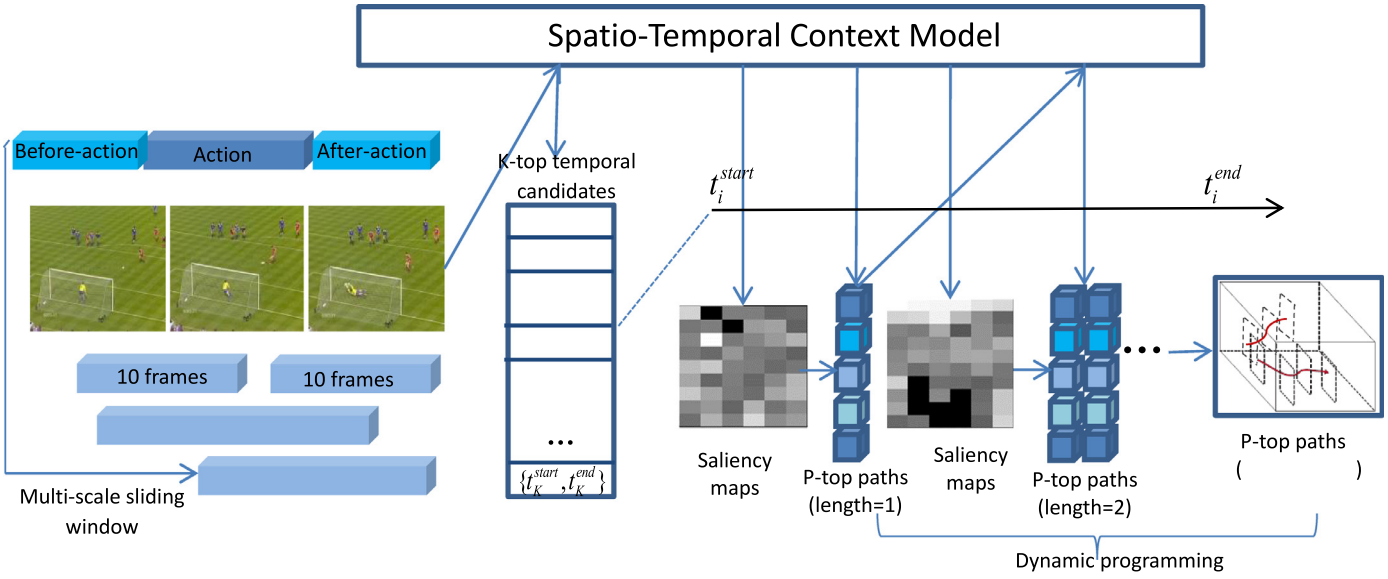


Fig. 3. The process of action spatio-temporal localization. The temporal localization and spatial localization execute in sequence and they refine each other. After get K-top temporal candidates, a dynamic programming approach is used to infer the P-top paths for each temporal candidate.

In analogy to the classical SVMs and according with Eqs. (16) and (19), the max-margin objective function is:

$$\min_{\boldsymbol{\Theta},\xi}\left\{\frac{1}{2}||\boldsymbol{\Theta}||^2+C\sum_{i=1}^{N}\xi_i\right\} \tag{20}$$

$$s.t.\ \forall i,\forall y\neq y_i,\ \xi_i\geq 0$$

$$\max_{\mathbf{h}\in H}(-\boldsymbol{\Theta}\cdot E(\mathbf{x}_i,y_i,\mathbf{h}))-\max_{\mathbf{h}\in H}(-\boldsymbol{\Theta}\cdot E(\mathbf{x}_i,y,\mathbf{h}))\geq\Delta(y_i,y)-\xi_i.$$

where C is the trade-off parameter and ξ_i is the slack variable for the $i-th$ sample to deal with the soft margin. Although there exists exponential number of constraints in Eq. (20), it can be easily solved by the Concave–Convex Procedure (CCCP) [45], which is equivalent to solving the learning problem with incomplete data using Expectation-Maximization (EM) algorithm.

5. Spatial and temporal localization

After learning the spatio-temporal context model, we detail how to use the STCM to recognize action label and detect action location simultaneously. Given a video, our goal is to recognize which action this video contains and detect when and where this action occurs, as explained in Eq. (4) above. The detected smooth spatio-temporal path $\mathbf{B}$ consists of the start frame t_1, the end frame t_2 and a series of bounding boxes $\mathbf{b}$ during this period. For the spatio-temporal localization, the search space is too large. So the temporal localization and spatial localization execute in sequence, where we fix one to detect the other, and they can refine each other as illustrated in Fig. 3.

5.1. Temporal localization with STCM

Given a new video $\mathbf{X}=\{\mathbf{X}_1,\mathbf{X}_2,\ldots,\mathbf{X}_T\}$ with T frames, the goal of temporal localization is to recognize the action $\hat{y}$ and detect the start frame t_1 and end frame t_2 of this action:

$$\begin{aligned}(\hat{y},\widehat{t_1},\widehat{t_2}) &= \arg\max_{y,t_1,t_2} P(y,t_1,t_2|\mathbf{X};\boldsymbol{\Theta}) \\ &= \arg\max_{y,t_1,t_2,\mathbf{h}} \frac{1}{Z}\exp\{-\boldsymbol{\Theta}\cdot(E(y,\mathbf{h},\mathbf{X}(t_1,t_2)) \\ &\quad +E(y,\mathbf{h},\mathbf{X}(t_1))+E(y,\mathbf{h},\mathbf{X}(t_2)))\},\end{aligned} \tag{21}$$

where only the whole frame is considered and we fix the context and target items in the STCM. The temporal search space is organized by multi-scale sliding window: $\{t_{1,i},t_{2,i}\}_{i=1}^{N}$, where $t_{1,i}>0$, $t_{2,i}<T$. Then K-top temporal candidates with highest confidence score $\{t_{1,i},t_{2,i},\hat{y}_i,s_i\}_{i=1}^{K}$ are achieved by Eq. (21) for further spatial localization, and their corresponding detection confidence scores can be calculated by $s_i=P(\hat{y}_i,t_{1,i},t_{2,i}|\mathbf{X};\boldsymbol{\Theta})$. It is benefit for detecting the real temporal boundary and avoiding the incomplete segments, since the detection score considers both action part and its temporal context part. For example, the before-action can refine the start boundary and the after-action can refine the end boundary.

5.2. Spatial localization with STCM

Then we fix the temporal localization result to detect the spatial bounding box and refine it in turn to get the final spatio-temporal localization result. Given a set of temporal candidates, the goal of spatial localization is to reevaluate the action label $\hat{y}$ and detect a series of spatial bounding boxes $\widehat{\mathbf{b}}$ among this time range:

$$(\hat{y},\widehat{\mathbf{b}})=\arg\max_{y,\mathbf{b}} P(y,\mathbf{b}|\mathbf{X}_{t_{1,i}:t_{2,i}};\boldsymbol{\Theta}). \tag{22}$$

Compared to 2D objection localization, the 3D action spatial localization is more challenging and complex. For a temporal candidate of size $M\times N\times T$, the search spaces for 3D subvolumes and 2D subwindows are only $O(M^2\times N^2\times T)$ and $O(M^2\times N^2)$, respectively. However, if we consider a flexible bounding box size where it can vary across frames, the search space will increase exponentially. Such that the exhaustive search is infeasible and Eq. (22) can not be easily inferred.

Therefore, we present a dynamic programming with saliency map framework to address this computational issue. In this paper, we propose a novel search algorithm which can locate the spatial locations in several successive frames without significantly increasing the number of candidates. The idea is to maintain a pool of the best P paths for each temporal candidate and update them at each time step. The pool of the i-th temporal candidate denotes as $Q^i=\{(\mathbf{b}^{i,p},\hat{y}^{i,p},s^{i,p}),p=1,2...P\}$, where $\mathbf{b}^{i,p}=\{b_t^{i,p},t=1,2,\ldots,t_{2,i}-t_{1,i}+1\}$ is a series of bounding boxes. For

the i-th temporal candidate, at the first frame we discover the top-P start locations $(b_1^{i,p}, \hat{y}^{i,p}, s^{i,p}), p = 1, 2, \ldots, P\}$ with highest confidence score using the multi-scale sliding window. We find the most probable location as the start of the path and infer its action label,

$$\begin{aligned}(\hat{y}^{i,p}, b_1^{i,p}) &= \arg\max_{y,b} P(y, b|\mathbf{X}_{t_{1,i}:t_{2,i}}; \boldsymbol{\Theta}) \\ &= \arg\max_{y,b,\mathbf{h}} \frac{1}{Z} exp\{-\boldsymbol{\Theta} \cdot (E(\mathbf{X}_{t_{1,i}}(\sim b), \mathbf{h}, y) \\ &\quad + E(\mathbf{X}(t_{1,i}), \mathbf{h}, y) + E(\mathbf{X}(t_{2,i}), \mathbf{h}, y) \\ &\quad + E(\mathbf{X}_{t_{1,i}}(b), \mathbf{h}, y) + E(\mathbf{X}(t_{1,i}, t_{2,i}), \mathbf{h}, y))\}\end{aligned} \tag{23}$$

then calculate its detection confidence score by $s^{i,p} = P(\hat{y}^{i,p}, b_1^{i,p}|\mathbf{X}_{t_{1,i}:t_{2,i}}, \boldsymbol{\Theta})$. During the forward search, we keep the previous paths $\{(b_{1:t-1}^{i,p}, \hat{y}^{i,p}, s^{i,p}), p = 1, 2, \ldots, P\}$ and try to find optimal current spatial locations $\{b_t^{i,p}, p = 1, 2, \ldots, P\}$ to extend paths from $t-1$ to t, that can maximize the confidence scores of the complete paths $\{(b_{1:t}^{i,p}, \hat{y}^{i,p}, s^{i,p}), p = 1, 2, \ldots, P\}$. Fixed previous detections, find the most probable location at the t-th frame and update its action label by dynamic programming,

$$\begin{aligned}\left(\hat{y}^{i,p}, b_t^{i,p}\right) &= \arg\max_{y,b} P\left(y, b_{1:t-1}^{i,p}, b|\mathbf{X}_{t_{1,i}:t_{2,i}}; \boldsymbol{\Theta}\right) + \alpha O\left(b, b_{t-1}^{i,p}\right) \\ &= \arg\max_{y,b,\mathbf{h}} \frac{1}{Z} exp\Big\{ - \boldsymbol{\Theta} \cdot \Big(E\left(\mathbf{X}_{t_{1,i}:t_{2,i}}\left(\sim b_{1:t}^{i,p}\right), \mathbf{h}, y\right) \\ &\quad + E\left(\mathbf{X}_{t_{1,i}:t_{2,i}}\left(b_{1:t}^{i,p}\right), \mathbf{h}, y\right) + E(\mathbf{X}(t_{1,i} : t_{2,i}), \mathbf{h}, y) \\ &\quad + E(\mathbf{X}(t_{1,i}), \mathbf{h}, y) + E(\mathbf{X}(t_{2,i}), \mathbf{h}, y))\Big\} + \alpha O\left(b, b_{t-1}^{i,p}\right)\end{aligned} \tag{24}$$

then recalculate its detection confidence score as $s^{i,p} = P(y, b_{1:t}^{i,p}|\mathbf{X}_{t_{1,i}:t_{2,i}}; \boldsymbol{\Theta}) + \alpha O(b_t^{i,p}, b_{t-1}^{i,p})$. Compared with Eq. (23) only considering the conditional probability of STCM as the detection confidence score, a smooth term $O(b, b_{t-1}^{i,p}) = \frac{\cap(b, b_{t-1}^{i,p})}{\cup(b, b_{t-1}^{i,p})}$ is adopted in Eq. (24), which make the spatio-temporal path smooth. Meanwhile, instead of multi-scale sliding window, a saliency map is utilized for sampling candidates. We define the map score by the previous detection score, such that we can pay more attention on those regions with high detection scores which are close to the previous paths. That is to say, we can treat previous detection as a weak detection for current frame, since human action is a relatively gradual process. Note that we don't calculate the mapping for each time step that decreases the computational cost. It can be seen that the spatial localization can not only detect the spatial location of action, but also refine the result of temporal localization. Similarly, the STCM also contributes to detecting the complete bounding box which can accurately divide the target from the context, since both target part and context part modify the spatial boundary together. After the forward search, we can achieve the P-top complete spatio-temporal paths for each temporal candidate. Then we choose the one with highest detection confidence score as the final localization result and its action label as the final recognition result. The overall action spatio-temporal localization procedure is outlined in Algorithm 1.

It's worth noting that we execute temporal localization before spatial localization and this order of localization operation can't be inversed. The reason is that our proposal is a video-level localization method for action sequence instead of frame-level method for static image, where STCM is a dynamic model and information of previous frames is already stored in hidden variables. Localization process for each frame is not independent, where the current detection is not only in terms of current frame, but also significantly affected by previous detections. If executing spatial localization before temporal localization, it means we consider the whole sequence with all T frames as current temporal candidate, then we should detect bounding box for each frame, and even for those frames without target action. Thus these correct detections would continuously accumulate errors and decrease performance for following frames.

Algorithm 1 Action spatio-temporal localization algorithm.

Input:

A video sequence $\{\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_T\}$;

Output:

Localization result $\hat{y}$ and recognition result $\hat{\mathbf{B}} = \{\hat{b}_t\}_{t=\hat{t}_1}^{t=\hat{t}_2}$;

Temporal localization:
Use multi-scale sliding window to generate N candidates for temporal localization $\{t_{1,i}, t_{2,i}\}_{i=1}^{N}$;
Achieve K-top temporal candidates with highest confidence score by STCM, $\{t_{1,i}, t_{2,i}, \hat{y}_i, s_i\}_{i=1}^{K}$ as (21);
Spatial localization:
for each $i \in [1, K]$ **do**
Discover the top-P start locations $(b_1^{i,p}, \hat{y}^{i,p}, s^{i,p}), p = 1, 2, \ldots, P\}$ with highest confidence score as (23);
for each path $p \in [1, P]$, $t \in [2 : t_{2,i} - t_{1,i} + 1]$ **do**
Use dynamic programming to find the most probable current location $b_t^{i,p}$ and update action label $\hat{y}^{i,p}$ as well as score $s^{i,p}$ for each path, as (24);
end for
end for
Find the optimal spatio-temporal path $(\hat{i}, \hat{p}) = \arg\max_{i,p}\{s^{i,p}|p = 1, 2, \ldots, P, i = 1, 2, \ldots, K\}$
$\{\hat{t}_1, \hat{t}_2\} = \{t_{1,\hat{i}}, t_{2,\hat{i}}\}$, $\hat{\mathbf{b}} = \mathbf{b}^{\hat{i},\hat{p}}$, $\hat{y} = \hat{y}^{\hat{i},\hat{p}}$;
return $\{\hat{t}_1, \hat{t}_2, \hat{\mathbf{b}}\}$ and $\hat{y}$;

6. Experimental results

6.1. Datasets and evaluation

In our experiments, we use two action datasets: UCF-Sports and UCF-101.

UCF-Sports [46]. The dataset is used for action spatial localization to detect the spatial location of action in realistic scenes. Since videos are already segmented to the short clips and bounding boxes annotations are provided for all frames as well, it does not require to detect the temporal boundary. The videos in UCF-Sports are taken under extremely challenging and uncontrolled conditions. It includes 150 clips from various sport events with 10 categories of actions: diving, golf, kicking, lifting, horse riding, running, skate boarding, swing bench, swing side and walking.

UCF-101 [47]. This large dataset is collected from YouTube for action recognition with more than 13,000 videos and 101 classes, where the spatio-temporal localization annotations are contained for a subset of 24 class labels: Basketball, BasketballDunk, Biking, CliffDiving, CricketBowling, Diving, Fencing, FloorGymnastics, GolfSwing, HorseRiding, IceDancing, LongJump, PoleVault, RopeClimbing, SalsaSpin, SkateBoarding, Skiing, Skijet, SoccerJuggling, Surfing, TennisSwing, TrampolineJumping, VolleyballSpiking, WalkingWithDog. Videos in UCF101 are relatively long and untrimmed, thus we can evaluate our spatio-temporal localization method on this dataset. In contrast to the UCF-Sports, the untrimmed nature of UCF101 makes it more realistic and challenging for localization task.

Features and parameters. Our algorithm needs a feature representation for each $\mathbf{x}^w$, $\mathbf{x}^c$, $\mathbf{x}^t$, $\mathbf{x}^b$, $\mathbf{x}^a$. Inspired by the study in [48–50], which have proved a deep architecture consisting of multiple layers with nonlinearity can improve the representation power of features, especially for recognition task, we adopt

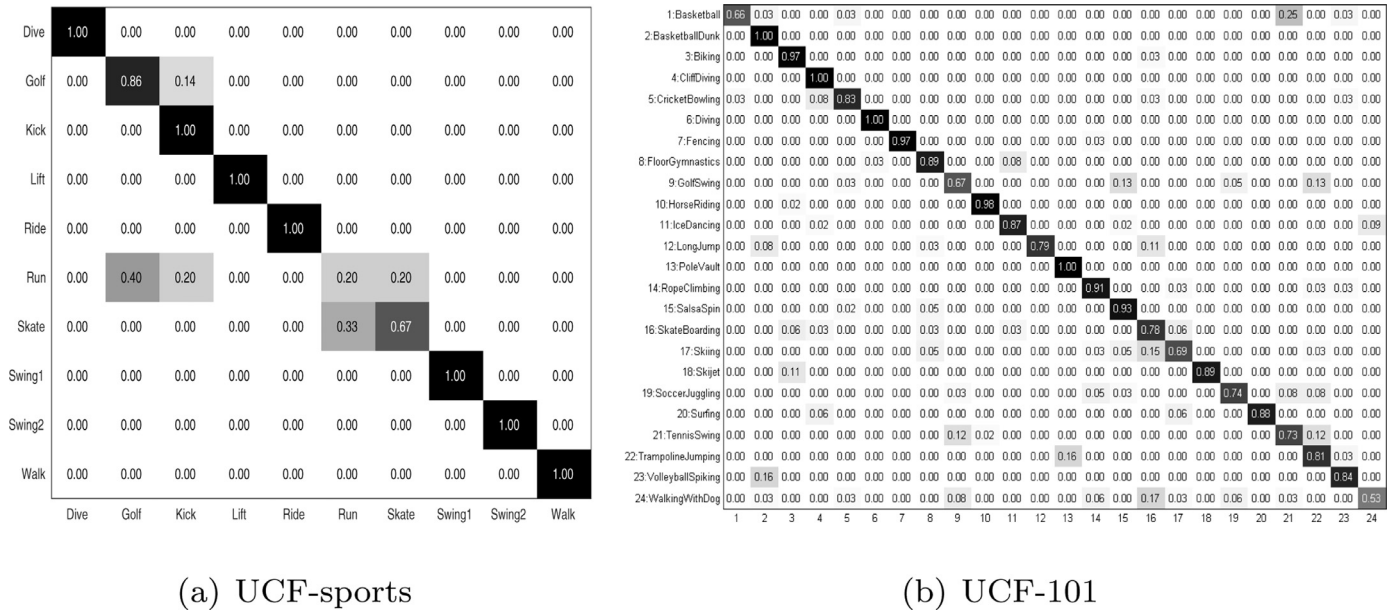


Fig. 4. The confusion matrices for action classification on the UCF-sports and UCF-101.

deep model to extract features. Each image is first warped to 224*224 pixels and then processed by a pre-trained CNN following the VGG-16 architecture [51]. For target $\mathbf{x}^t$, it is the descriptor of the inside bounding boxes, where we first crop the target region and then extract feature representation by the deep model. In contrast, the spatial context $\mathbf{x}^c$ is the descriptor of the outside bounding boxes around the target region, so we first mask the target region and then process it by the deep feature extracting model. Since the whole $\mathbf{x}^w$ contains both the target part and context part, we combine the two to represent it, namely directly extract the feature representation on the whole image. The way for computing the descriptors of temporal context $\mathbf{x}^b$ and $\mathbf{x}^a$ are same as that for calculating the whole $\mathbf{x}^w$. In this paper, the smooth coefficient α is set to 0.1 and the number of hidden states is set to 2 by cross validation. We keep 20-top temporal candidates for spatial localization and maintain a pool of best paths with size of 20 for each temporal candidate by default. The temporal search space is organized by multi-scale sliding window and we fix the range of window length to [20:10:T] with scanning stride of 5 frames, which can cover every temporal interval among video sequence and adaptively deal with action sequences with different length. For spatial search space of first frame, original sliding window size is 50×50 pixels, scanning stride is 20 pixels, and scale step is 0.5 which balances the localization precision and searching efficiency.

Evaluation metrics. A localization is considered as correct if its intersection over union (IoU) with the ground-truth is above a threshold δ. In this paper, a video prediction is considered as correct if both the predicted action label and the localization result match the ground truth. The IoU between two spatio-temporal paths is defined as the average of the IoU between bounding boxes among all overlapping frames in temporal domain. To fully evaluate our model, we fix the range of IoU threshold to [0.2,0.5] for spatial detection on the UCF-Sports, [0.2,0.8] for temporal detection on the UCF-101, and [0.05, 0.1, 0.2, 0.3] for spatio-temporal detection on the UCF-101. By default, the reported metric is the mean Average Precision (mAP) at IoU threshold δ=20% for spatial localization (UCF-Sports), δ=50% for temporal localization and δ=10% for spatio-temporal localization (UCF-101).

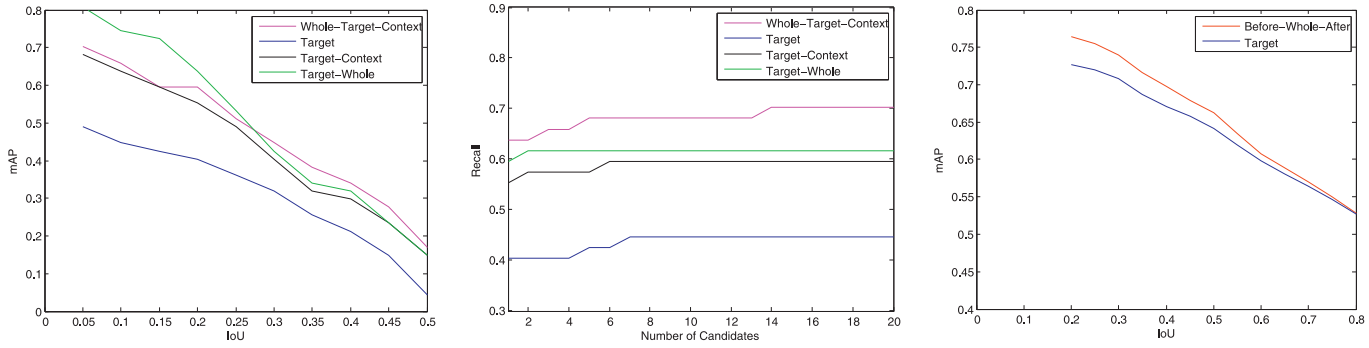
Table 1
The recognition accuracy on the UCF-sports and UCF-101.

Algorithm	UCF-sports	UCF-101
Baseline1 [52]	80%	46.95%
Baseline2 [48]	83.67%	80.20%
C3D + linear SVM [50]	–	82.30%
HSTM [53]	90.67%	–
Whole-chain [54]	82.98%	79.51%
Target-chain	85.11%	80.07 %
Context-chain	72.34%	76.08%
Before-Whole-After	–	82.39%
Whole-Target-Context	87.23%	80.84%
STCM	87.23%	85.16%

6.2. Action recognition result

Our method recognizes the action label and detects action location simultaneously. We first analyze the recognition performance of our STCM. We use the standard training and test splits for both two datasets.

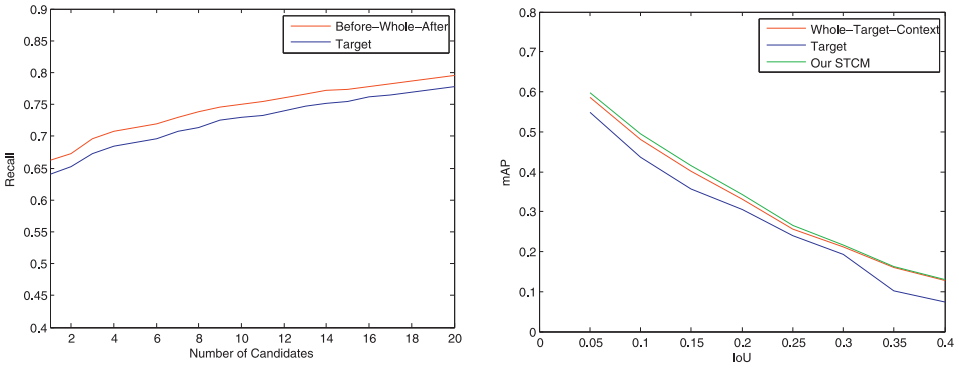
The classic approach [52] applying SVM based on the traditional spatio-temporal interested points (STIPs) with a bag-of-feature (BOF) style representation is used as the first baseline. Additional, another baseline is to directly integrate the features extracting network into a classification network to obtain the final recognition result, which is something like the spatial stream of [48]. We train the unified classification network with the similar architecture, where the only difference is that during fine-tuning we add two fully-connected layers and a classification FC layer of 24 neurons or 10 neurons (consisting of 24 UCF-101 detection classes or 10 UCF-Sports detection classes) following the features extracting network. From Table 1, it can be found that recognition performance of STCM is slightly worse than HSTM [53] on UCF-Sports, another probabilistic graphical model based method. The reason is that HSTM is a much more complex model and able to capture more fine-gained relations, e.g. the complexity of HSTM is $O(TN_y|\mathcal{E}|N_h + N_yTN_h)$ and STCM is $O(3N_yTN_h + 2N_yN_h)$, where N_h is size of hidden set, N_y is size of label set and patch edge number



(a) The mAP for spatial localization at different IoUs on UCF-sport

(b) Recall for spatial localization at different numbers of candidates on UCF-sport

(c) The mAP for temporal localization at different IoUs on UCF-101



(d) Recall for temporal localization at different numbers of candidates on UCF-101

(e) The mAP for spatial and temporal localization at different IoUs on UCF-101

Fig. 5. Action spatial and temporal localization results on UCF-sport and UCF-101, measured by mAP and Recall at different IoUs and numbers of candidates.

$|\mathcal{E}| \gg T$. Therefore, HSTM is too complex to be utilized for recognition task on big dataset (e.g. UCF-101) and also it is not suitable for localization task, which needs to handle multiple proposals.

In order to comprehensively evaluate the performance of STCM and the effectiveness of spatio-temporal contexts, we compare STCM with its sub-models with different structures on the two benchmark datasets in Table 1. The target-chain and context-chain mean only modeling the context and target region independently; Although the whole-chain is built on the whole frame, it makes no distinction between target and context, which is equivalent to [54]; The Whole-Target-Context constructs the three chains jointly, but it only contains the spatial context; The Before-Whole-After models the complete dynamic process of action, but it only contains the temporal context; The complete STCM incorporates both spatial context and temporal context. Note that videos in UCF-sports are the segmented short clips, where no before or after sequence exists, such that the STCM is equivalent to the Whole-Target-Context here. From these comparisons, the following points can be indicated: (1) Besides the target-chain, the context-chain and the whole-chain also contribute to recognition task. (2) Both spatial and temporal contexts are important for action recognition, where Before-Whole-After and Whole-Target-Context can improve the accuracy of each individual chain. (3) The spatial and temporal contexts complement each other and combining them can reach the highest recognition rate. It concludes that both spatial and temporal contexts can improve model's discriminative power and descriptive capability. (4) Our STCM can achieve comparable and even better recognition performance with state-of-the-art methods. The confusion matrixes obtained by our model on the two datasets are shown in Fig. 4, which are detailed results for action recognition of each action category.

6.3. Action localization result

Now we evaluate the performance of action localization using the STCM. The advantage of our method lies in the combination of target action and its context for action localization. Since both temporal context and spatial context are integrated into STCM, to verify the benefit of this combining, we test and analyze them independently as following.

Firstly, to evaluate the effectiveness of spatial context, we compare the performance for action spatial localization on the

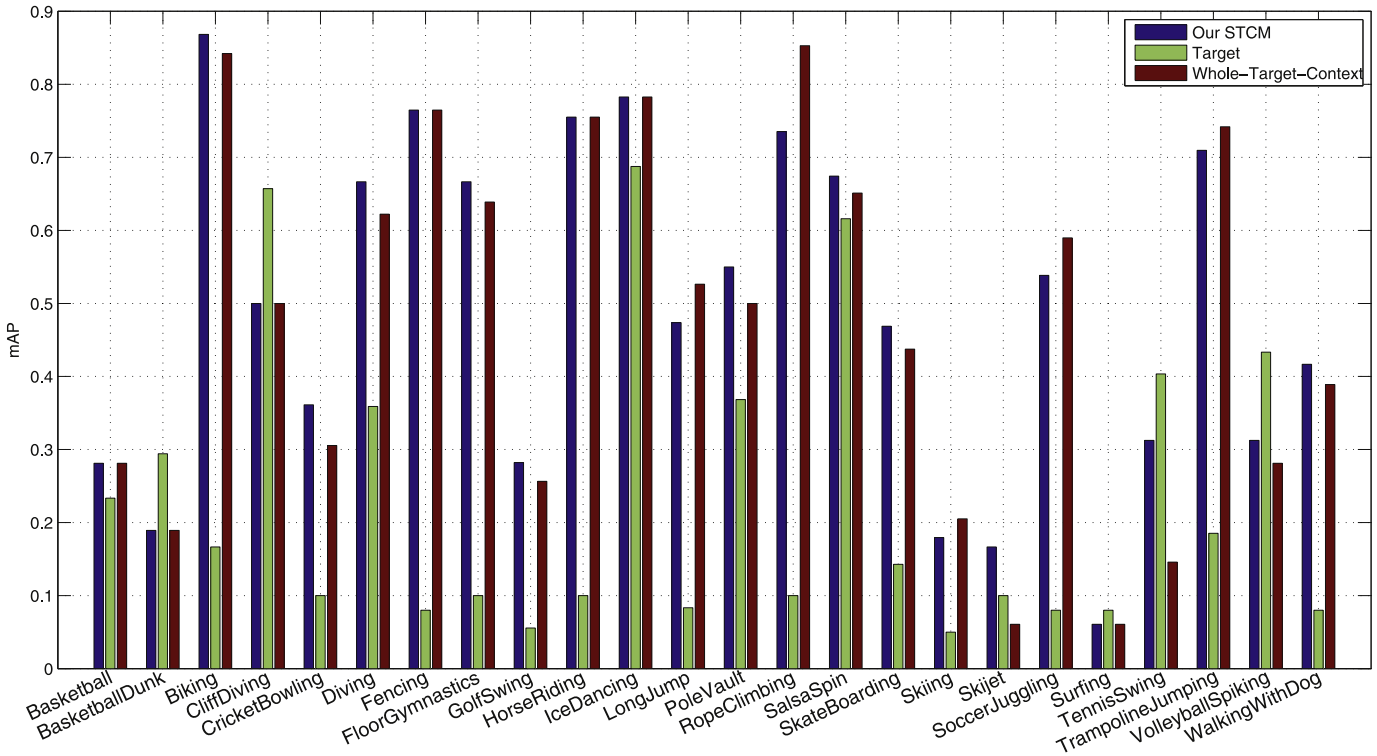


Fig. 6. The AP of each action category on the UCF-101 for target-centered method (Target), model without temporal context (Whole-Target-Context) and model with spatio-temporal context together (our STCM).

UCF-Sports: (1) Whole-Target-Context: incorporate the whole-chain, target-chain and context-chain to fully capture all kinds of spatial contextual relationships; (2) Target-Context: remove the whole-chain and only model the relations between target and context; (3) Target-Whole: reserve the target-chain and whole-chain to implicitly exploit some spatial contexts; (4)Target: remove all spatial contexts and only utilize the target-chain, which is called target-centered method. We follow the conventions, report the mAP at different IoU thresholds in Fig. 5(a). In addition, the 'Recall-candidate' measures the recall rate when varying the numbers of candidate paths as shown in Fig. 5(b). It can be seen that our method outperforms the target-centered method significantly, and it achieves promising results even with a small number of candidate paths or at a high IoU threshold. Meanwhile, it verifies both context-chain and whole-chain have contributions for capturing spatial context and locating action. Although the 'Target-Whole' can obtain higher mAP than the 'Whole-Target-Context' at low IoU, we get the best performance at high IoU when incorporating the three chain together. That is to say, the target-chain is the fundamental one which provides a coarse position extent of action, and its context contributes to refine it to get more accurate spatial boundary. Under the same condition, when using multi-scale sliding window instead of dynamic programming, it requires more than 150 proposals for each time step to obtain the similar result.

Secondly, we also evaluate the effectiveness of temporal context on action localization for detailed analysis. The compared results of temporal localization on the UCF-101 are reported in Fig. 5, where the 'Before-Whole-After' means that incorporating action itself, before-action and after-action together to fully capture temporal contextual relationships, and the 'Target' refers to the model without considering any temporal context. In the same way, we show the mAP at different IoU thresholds in Fig. 5(c) and the recall rate at different numbers of candidates in Fig. 5(d) for temporal localization. It can be found that our model outperforms the target-centered method significantly, no matter at a low or high IoU threshold. Thus we can verify that the temporal context is indeed benefit for detecting accurate temporal boundary of human actions.

Thirdly, we evaluate the performance of complete STCM for action spatio-temporal localization on the UCF-101, where we can consider the 'Whole-Target-Context' and 'Before-Whole-After' as sub-models of STCM for spatial localization task and temporal localization task respectively. The mAP when varying IoU thresholds is reported in Fig. 5(e), where STCM denotes the complete model, the 'Whole-Target-Context' means removing the temporal context from STCM, and the 'Target' refers to STCM without any spatial and temporal contexts. It demonstrates that the performance improves when incorporating more contextual information (STCM > 'Whole-Target-Context' > 'Target'), which is the consistent conclusion with spatial localization and temporal localization above. From these experimental results, we have following observations: First, it just requires a small number of candidate paths to achieve the optimal localization performance in our dynamic programming with saliency map framework, such that it saves the computational cost. Second, integrating the spatio-temporal contexts to locate action can achieve more complete boundary and get more accurate detection result.

Finally, we compare our model with other state-of-the-art action localization methods on the UCF-Sports and UCF-101. To report the performance, we use the mAP as the base metrics, which can well measure the capacity to localize accurate spatio-temporal boundary of action sequence. The localization results on UCF-Sports and UCF-101 are shown in Table 2. It is also worth noting that the mAP is reported at IoU=0.1 for spatio-temporal localization on the UCF-101, while it is reported at IoU=0.2 for only spatial localization on the UCF-Sports. On average, our approach obtains 11.57% and 5.27% performance gains than [8] and [17], respectively, and outperforms [57] substantially by 4.57% on the UCF-Sports. Despite of the challenge of locating actions both spatially and temporally, we can achieve better performance than

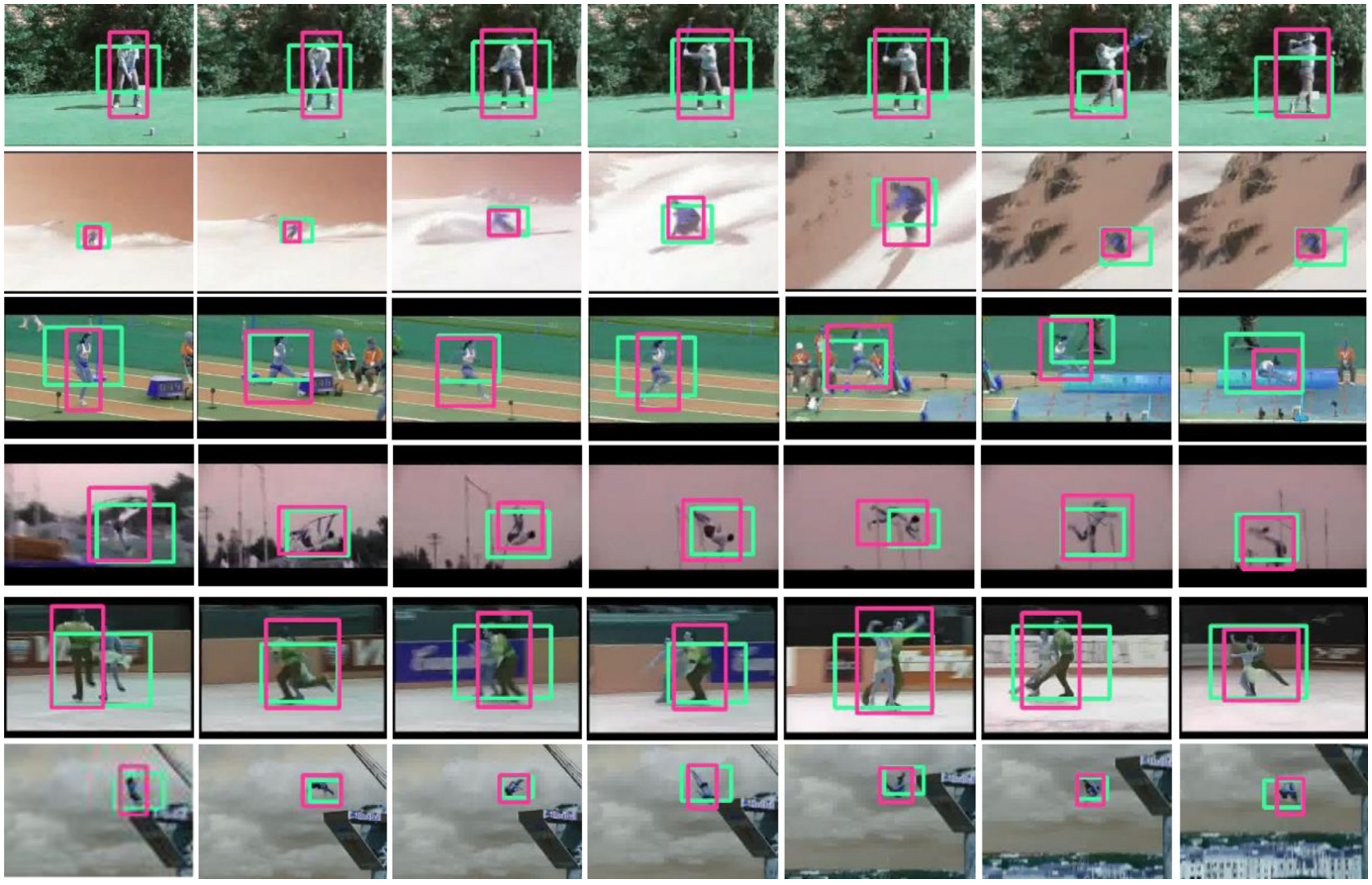


Fig. 7. Illustrative examples of our localization results (with green bounding boxes) on the UCF-101 dataset. The ground truth is marked by red bounding box. It can be seen that we can accurately locate non-moving actions as well as actions with strong moving. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 2
Action localization results measured by MAP at IoU=0.2 for spatial localization on the UCF-sports and at IoU = 0.1 for spatio-temporal localization on the UCF-101.

Algorithm	UCF-sports	UCF-101
Tran and Yuan [17]	54.30	–
Wang et al. [8]	48.0	–
Van Gemert et al. [55]	54.6	45.0
Mettes et al. [56]	54.5	34.8
Lu et al. [24]	48.1	–
Soomro et al. [57]	55	–
Gkioxari et al. [32]	58.34	–
Yu and Yuan [18]	–	42.8
Peng and Schmid [10]	–	**50.39**
Target chain	40.42	35.57
Ours	**59.57**	49.39

[18] and comparable performance with [10] using much less proposals. In order to reach the optimal performance, we keep a pool of 20 candidates in our model, while the required numbers of proposals are 10K in [18] and 256 in [10], respectively. In particular, combining spatio-temporal context leads to an improvement of 19.15% on the UCF-Sports and 13.82% on the UCF-101. Another evidence illustrating the importance of context is that R*CNN [32] is superior to [57] by exploiting contextual cues, which both depend on RCNN framework. Instead of considering the most informative secondary region as context [32], we find a more suitable way to define context by dynamically considering the spatial context as the outside of detected bounding box and obtain more accurate localization result on UCF-Sports. It clearly demonstrates the superior ability of our method in accurate action localization and the benefit of explicitly modeling context. Additionally, since R*CNN [32] is a frame-level method only designed for spatial localization, it cannot be used for spatio-temporal localization on UCF-101.

The AP of action localization on the UCF-101 is presented in Fig. 6, which contains the per class results for target-centered method, model without temporal context and our STCM. Compared to other works [18], which causes much performance decrease for some actions with strongly moving, our STCM can relatively accurately locate these moving actions. Since they detect each frame independently and fail to exploit temporal dynamics in action sequences. Take some moving actions as examples, we show some comparison results between [18] and ours: 'Diving' (22% *vs* 67%), 'VolleyballSpiking' (lower than 10% *vs* higher than 30%), 'CliffDiving' (21% *vs* 50%), and 'PoleVault' (31% *vs* 55%). Some examples of action localization results for different categories on the UCF-101 are illustrated in Fig. 7.

7. Conclusion

In this paper, we propose a unified framework to simultaneously recognize action label and detect action spatio-temporal location. The STCM is proposed to associate target action with its context and model their underlying relationships. Moreover, a novel dynamic programming approach is utilized to infer the best spatio-temporal path, which not only preserves the dynamic property of actions, but also reduces the searching space. By introducing various kinds of contextual information into action representation, we achieve better performance in action recognition task

and obtain more accurate boundary than those target-centered methods in action localization task.

Acknowledgment

This work is supported by the NSFC 61672089, 61273274, 61572064 and National Key Technology R&D Program of China 2012BAH01F03.

References

[1] G. Yu, N.A. Goussies, J. Yuan, Z. Liu, Fast action detection via discriminative random forest voting and top-k subvolume search, IEEE Trans. Multimed. 13 (3) (2011) 507–517.
[2] Z. Zhou, F. Shi, W. Wu, Learning spatial and temporal extents of human actions for action detection, IEEE Trans. Multimed. 17 (4) (2015) 512–525.
[3] W. Xu, Z. Miao, X.P. Zhang, et al., A hierarchical spatio-temporal model for human activity recognition, IEEE Trans. Multimed. 19 (7) (2017) 1494–1509.
[4] H. Wang, C. Schmid, Action recognition with improved trajectories, in: Proceedings of the IEEE International Conference on Computer Vision, 2013, pp. 3551–3558.
[5] L. Wang, Y. Qiao, X. Tang, Action recognition with trajectory-pooled deep-convolutional descriptors, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 4305–4314.
[6] L. Wang, Y. Xiong, D. Lin, et al., Untrimmednets for weakly supervised action recognition and detection, IEEE Conf. on Computer Vision and Pattern Recognition, 2017 2.
[7] S. Saha, G. Singh, M. Sapienza, et al., Deep learning for detecting multiple space-time action tubes in videos, Pattern Recognit. (2015).
[8] L. Wang, Y. Qiao, X. Tang, Video action detection with relational dynamic-poselets, in: Proceedings of European Conference on Computer Vision, Springer, 2014, pp. 565–580.
[9] P. Weinzaepfel, Z. Harchaoui, C. Schmid, Learning to track for spatio-temporal action localization, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 3164–3172.
[10] X. Peng, C. Schmid, Multi-region two-stream r-cnn for action detection, in: Proceedings of European Conference on Computer Vision, Springer, 2016, pp. 744–759.
[11] J. Yuan, Z. Liu, Y. Wu, Discriminative video pattern search for efficient action detection, IEEE Trans. Pattern Anal. Mach. Intell. 33 (9) (2011) 1728–1743.
[12] P. Siva, T. Xiang, Action detection in crowd., in: Proceedings of British Machine Vision Conference, BMVC, 2010, pp. 1–11.
[13] A. Gaidon, Z. Harchaoui, C. Schmid, Temporal localization of actions with actoms, IEEE Trans. Pattern Anal. Mach. Intell. 35 (11) (2013) 2782–2795.
[14] L. Wang, Y. Qiao, X. Tang, Action Recognition and Detection by Combining Motion and Appearance Features, THUMOS14 Action Recognition Challenge 1 (2014) 2.
[15] D. Oneata, J. Verbeek, C. Schmid, Efficient action localization with approximately normalized fisher vectors, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 2545–2552.
[16] J. Yuan, Z. Liu, Y. Wu, Discriminative subvolume search for efficient action detection, in: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, CVPR, IEEE, 2009, pp. 2442–2449.
[17] D. Tran, J. Yuan, Max-margin structured output regression for spatio-temporal action localization, in: Proceedings of Advances in Neural Information Processing Systems, 2012, pp. 350–358.
[18] G. Yu, J. Yuan, Fast action proposals for human action detection and search, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 1302–1311.
[19] T. Wang, S. Wang, X. Ding, Detecting human action as the spatio-temporal tube of maximum mutual information, IEEE Trans. Circuits Syst. Video Technol. 24 (2) (2014) 277–290.
[20] J. Yuan, B. Ni, X. Yang, A.A. Kassim, Temporal action localization with pyramid of score distribution features, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 3093–3102.
[21] F. Zheng, L. Shao, Z. Song, A set of co-occurrence matrices on the intrinsic manifold of human silhouettes for action recognition, in: Proceedings of the ACM International Conference on Image and Video Retrieval, ACM, 2010, pp. 454–461.
[22] L. Liu, L. Shao, F. Zheng, X. Li, Realistic action recognition via sparsely-constructed gaussian processes, Pattern Recognit. 47 (12) (2014) 3819–3827.
[23] M. Jain, J. Van Gemert, H. Jégou, P. Bouthemy, C.G. Snoek, Action localization with tubelets from motion, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 740–747.
[24] J. Lu, J.J. Corso, et al., Human action segmentation with hierarchical supervoxel consistency, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 3762–3771.
[25] S. Ma, J. Zhang, N. Ikizler-Cinbis, S. Sclaroff, Action recognition and localization by hierarchical space-time segments, in: Proceedings of the IEEE International Conference on Computer Vision, 2013, pp. 2744–2751.
[26] Y. Tian, R. Sukthankar, M. Shah, Spatiotemporal deformable part models for action detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2013, pp. 2642–2649.
[27] T. Lan, Y. Wang, G. Mori, Discriminative figure-centric models for joint action localization and recognition, in: Proceedings of the 2011 IEEE International Conference on Computer Vision, ICCV, IEEE, 2011, pp. 2003–2010.
[28] G. Gkioxari, J. Malik, Finding action tubes, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 759–768.
[29] Z. Shou, D. Wang, S.-F. Chang, Temporal action localization in untrimmed videos via multi-stage cnns, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 1049–1058.
[30] R. Girshick, Fast r-cnn, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 1440–1448.
[31] S. Yeung, O. Russakovsky, G. Mori, L. Fei-Fei, End-to-end learning of action detection from frame glimpses in videos, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 2678–2687.
[32] G. Gkioxari, R. Girshick, J. Malik, Contextual action recognition with r* cnn, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 1080–1088.
[33] M. Hasan, A.K. Roy-Chowdhury, Context aware active learning of activity recognition models, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 4543–4551.
[34] F.C. Heilbron, A. Thabet, J.C. Niebles, B. Ghanem, Camera motion and surrounding scene appearance as context for action recognition, in: Proceedings of Asian Conference on Computer Vision, Springer, 2014, pp. 583–597.
[35] Y. Zhang, Y. Zhang, E. Swears, N. Larios, Z. Wang, Q. Ji, Modeling temporal interactions with interval temporal Bayesian networks for complex activity recognition, IEEE Trans. Pattern Anal. Mach. Intell. 35 (10) (2013) 2468–2483.
[36] V. Ramanathan, K. Tang, G. Mori, L. Fei-Fei, Learning temporal embeddings for complex video analysis, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 4471–4479.
[37] X. Wang, Q. Ji, Hierarchical context modeling for video event recognition, IEEE Trans. Pattern Anal. Mach. Intell. 39 (9) (2017) 1770–1782.
[38] W. Choi, K. Shahid, S. Savarese, Learning context for collective activity recognition, in: Proceedings of 2011 IEEE Conference on Computer Vision and Pattern Recognition, CVPR, IEEE, 2011, pp. 3273–3280.
[39] M.S. Aliakbarian, F. Saleh, B. Fernando, M. Salzmann, L. Petersson, L. Andersson, Deep Action-and Context-aware Sequence Learning for Activity Recognition and Anticipation, arXiv:1611.05520 (2016).
[40] B. Wu, C. Yuan, W. Hu, Human action recognition based on context-dependent graph kernels, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 2609–2616.
[41] M. Wang, B. Ni, X. Yang, Recurrent Modeling of Interaction Context for Collective Activity Recognition.
[42] D. Koller, N. Friedman, Probabilistic Graphical Models: Principles and Techniques, MIT Press, 2009.
[43] A.G. Schwing, T. Hazan, M. Pollefeys, R. Urtasun, Efficient structured prediction with latent variables for general graphical models, in: Proceedings of International Conference on Machine Learning, 2012, pp. 1659–1666.
[44] W. Ping, Q. Liu, A. Ihler, Marginal Structured svm with Hidden Variables, Eprint Arxiv (2014) 190–198.
[45] A.L. Yuille, A. Rangarajan, A. Yuille, The concave-convex procedure (cccp), in: Proceedings of Advances in Neural Information Processing Systems, 2, 2002, pp. 1033–1040.
[46] K. Soomro, A.R. Zamir, Action recognition in realistic sports videos, in: Proceedings of Computer Vision in Sports, Springer, 2014, pp. 181–208.
[47] K. Soomro, A.R. Zamir, M. Shah, UCF101: A dataset of 101 human actions classes from videos in the wild, Computer Science (2012).
[48] K. Simonyan, A. Zisserman, Two-stream convolutional networks for action recognition in videos, in: Proceedings of Advances in Neural Information Processing Systems, 2014, pp. 568–576.
[49] J. Yue-Hei Ng, M. Hausknecht, S. Vijayanarasimhan, O. Vinyals, R. Monga, G. Toderici, Beyond short snippets: deep networks for video classification, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 4694–4702.
[50] D. Tran, L. Bourdev, R. Fergus, L. Torresani, M. Paluri, Learning spatiotemporal features with 3d convolutional networks, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 4489–4497.
[51] K. Simonyan, A. Zisserman, Very Deep Convolzutional Networks for Large-scale Image Recognition, arXiv:1409.1556 (2014).
[52] H. Wang, M.M. Ullah, A. Klaser, I. Laptev, C. Schmid, Evaluation of local spatio-temporal features for action recognition, in: Proceedings of British Machine Vision Conference, BMVC, BMVA Press, 2009, p. 124.1.
[53] W. Xu, Z. Miao, X.P. Zhang, Y. Tian, A hierarchical spatio-temporal model for human activity recognition, IEEE Trans. Multimed. 19 (7) (2017) 1494–1509.
[54] Y. Wang, G. Mori, Max-margin hidden conditional random fields for human action recognition, in: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, CVPR, IEEE, 2009, pp. 872–879.
[55] J.C. van Gemert, M. Jain, E. Gati, C.G. Snoek, et al., Apt: action localization proposals from dense trajectories., in: Proceedings of British Machine Vision Conference, BMVC, 2, 2015, p. 4.
[56] P. Mettes, J.C. van Gemert, C.G. Snoek, Spot on: action localization from pointly-supervised proposals, in: Proceedings of European Conference on Computer Vision, Springer, 2016, pp. 437–453.
[57] K. Soomro, H. Idrees, M. Shah, Action localization in videos through context walk, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 3280–3288.

Wanru Xu received the B.S. degrees in biomedical engineering and the Ph.D. degree in signal and information processing from Beijing Jiaotong University, Beijing, China, in 2011 and 2017, respectively. She is currently a Post-Doctoral Researcher with the School of Computer and Information Technology, Beijing Jiaotong University, Beijing, China. Her current research interests include computer vision, machine learning and pattern recognition.

Zhenjiang Miao (M'11) received the B.E. degree from Tsinghua University, Beijing, China, in 1987, and the M.E. and Ph.D. degrees from Northern Jiaotong University, Beijing, in 1990 and 1994, respectively. From 1995 to 1998, he was a Post-Doctoral Fellow with the École Nationale Supérieure d'Electrotechnique, d'Electronique, d'Informatique, d'Hydraulique et des Télécommunications, Institut National Polytechnique de Toulouse, Toulouse, France, and was a Researcher with the Institute National de la Recherche Agronomique, Sophia Antipolis, France. From 1998 to 2004, he was with the Institute of Information Technology, National Research Council Canada, Nortel Networks, Ottawa, Canada. He joined Beijing Jiaotong University, Beijing, in 2004. He is currently a Professor, Director of the Media Computing Center, Beijing Jiaotong University, and Director of the Institute for Digital Culture Research, Center for Ethnic & Folk Literature & Art Development, Ministry Of Culture, P.R. China. His current research interests include image and video processing, multimedia processing, and intelligent human–machine interaction.

Jian Yu received the B.S. and M.S. degrees in mathematics and the Ph.D. degree in applied mathematics from Peking University, Beijing, China, in 1991, 1994, and 2000, respectively. He is currently a Professor with Beijing Jiaotong University, Beijing, and the Director of the Beijing Key Laboratory of Traffic Data Analysis and Mining. His current research interests include machine learning, image processing, and pattern recognition.

Qiang Ji received the Ph.D. degree from the University of Washington. He is currently a Professor in the Department of Electrical, Computer, and Systems engineering, RPI. From January, 2009 to August, 2010, he served as a program director of the National Science Foundation, managing NSF's machine learning and computer vision programs. Prior to joining RPI in 2001, he was an assistant professor in the Department of Computer Science, University of Nevada, Reno. He also held research and visiting positions in the Beckman Institute, University of Illinois at Urbana-Champaign, the Robotics Institute, Carnegie Mellon University, and the US Air Force Research Laboratory. He is a fellow of the IEEE and the IAPR.