Learning a strong detector for action localization in videos

Yongqiang Zhang^a, Mingli Ding^{a,**}, Yancheng Bai^{b,**}, Dandan Liu^c, Bernard Ghanem^d

^aSchool of Instrument Science and Engineering, Harbin Institute of Technology (HIT), Harbin 150001, China

^bInstitute of Software, Chinese Academy of Sciences, Beijing 100190, China

^cSchool of Foreign Language Studies, Harbin Huade University (HD), Harbin 150025, China

^dVisual Computing Center, King Abdullah University of Science and Technology (KAUST), Thuwal 23955-6900, Saudi Arabia

ABSTRACT

We address the problem of spatio-temporal action localization in videos in this paper. Current state-of-the-art methods for this challenging task rely on an object detector to localize actors at frame-level firstly, and then link or track the detections across time. Most of these methods commonly pay more attention to leveraging the temporal context of videos for action detection while ignoring the importance of the object detector itself. In this paper, we prove the importance of the object detector in the pipeline of action localization, and propose a strong object detector for better action localization in videos, which is based on the single shot multibox detector (SSD) framework. Different from SSD, we introduce an anchor refine branch at the end of the backbone network to refine the input anchors, and add a batch normalization layer before concatenating the intermediate feature maps at frame-level and after stacking feature maps at clip-level. The proposed strong detector have two contributions: (1) reducing the phenomenon of missing target objects at frame-level; (2) generating deformable anchor cuboids for modeling temporal dynamic actions. Extensive experiments on UCF-Sports, J-HMDB and UCF-101 validate our claims, and we outperform the previous state-of-the-art methods by a large margin in terms of frame-mAP and video-mAP, especially at a higher overlap threshold.

Keywords: *Frame-level object detection, deformable anchor cuboid, action localization.*

© 2019 Elsevier Ltd. All rights reserved.

1. Introduction

Action localization in videos is one of the most challenging problems in video understanding (Zhang and Ma, 2014; Chen et al., 2017; Khelifi and Bouridane, 2017; Ma et al., 2017; Hergetz et al., 2017; Yuan et al., 2017; Shiao et al., 2017; Afonso et al., 2019), which aims to classify the category of actions and localize actions in both time and space. In other words, given a video, we should not only recognize the category of actions (Peng et al., 2018; Zhao and Peng, 2017) but also localize the start time and end time of actions and further detect the corresponding bounding-boxes of subjects (actors). It has been widely studied in the past few decades due to various high-level real-world applications, like video surveillance (Giancola et al., 2018; Oh et al., 2011; Hu et al., 2004) and video caption-

ing (Yao et al., 2015; Venugopalan et al., 2015), which are much more challenging due to the significant variations in occlusion, viewpoint, background, quick motion, and quality of video data. The solutions to this challenging task have developed from traditional handcrafted feature based approaches (Laptev and Pérez, 2007; Tian et al., 2013; Yuan et al., 2009) to deep learning based methods (Gkioxari and Malik, 2015; Li et al., 2018; Kalogeiton et al., 2017; Alwassel et al., 2018; Caba Heilbron et al., 2018), and significant progress has been made recently (Peng and Schmid, 2016; Saha et al., 2016; Wang et al., 2016; Yang et al., 2017) with the success of deep convolutional neural networks (DCNNs) (He et al., 2016; Krizhevsky et al., 2012; Szegedy et al., 2015).

Object detection is a fundamental problem and it is the basic technology of some advanced tasks, like video Captioning (Zhang and Peng, 2019b,a), action detection (Li et al., 2018). In addition to the general object detection methods based on RGB images, there are also some methods (Jian et al., 2014, 2018a,b) based on saliency detection to highlight and

^{**}Corresponding author.

e-mail: mingli.ding.hit@gmail.com (Mingli Ding),
yancheng.bai.1987@gmail.com (Yancheng Bai)

separate salient objects from complex background. Inspired by the advancement of object detection methods (Ren et al., 2017; He et al., 2017; Liu et al., 2016), most of state-of-the-art action localization methods (Saha et al., 2016; Weinzaepfel et al., 2015; Singh et al.; Hou et al., 2017; Kalogeiton et al., 2017) employ object detectors to detect human actions in videos. More concretely, these methods first detect proposals at frame-level (Peng and Schmid, 2016) or clip-level (Hou et al., 2017), and then link or track the fragmentary detections (subjects/actors) over time to generate spatio-temporal tubes. Although superior performance has been achieved, such methods pay more attention to exploring the temporal relations across frames or clips and just employ an off-the-shelf object detector to localize actors in videos, ignoring the importance of the object detector itself. As for the methods of detecting subjects at frame-level, a weaker object detector may result in missing target objects in some frames when scenarios are complex (*e.g.* in occlusion, under quick motion, with complicated background, *etc.*), which will cut off a tube in the link procedure and even lead to a failed action detection. In terms of the methods of detecting actors at clip-level, an unsuitable detector is not able to regress the prior defined anchors since they usually use rigid 3D action cuboid proposals (*i.e.* the bounding-boxes in a clip is fixed even though the spatial positions and scales of actors might vary significantly along time) to recognize actions, which failed to model the flexibility of actions.

To address the aforementioned issue, in this paper, we learn a strong detector for action spatio-temporal action detection in videos. Our baseline detector is based on SSD framework in (Kalogeiton et al., 2017). Different from the baseline detector, we concatenate some intermediate convolutional layers (*i.e.* *conv4_3*, *conv5_2* and *fc7* layers in our implementation) as original SSD detector (Liu et al., 2016) for detecting the objects with different scales. More importantly, we add a normalization layer (Ioffe and Szegedy, 2015) before concatenating feature maps of intermediate convolutional layers at frame-level, and also add a normalization layer after stacking feature maps of each frame in an input sequence (*i.e.* concatenating feature maps at clip-level) for reducing internal covariate shift of the designed network, as shown in Figure 1. Meanwhile, the introduced normalization layers allow us to use much higher learning rates and to be less careful about initialization to train the action detection network. Moreover, we introduce an anchor refine branch at the backbone network of the detector to refine prior defined anchors with different scales and aspect ratios. The added anchor refine branch can generate accurate bounding-boxes for actors at each frame in videos. For an input clip, the per-frame accurate bounding-boxes across the input sequence can be regarded as a deformable region tube proposal, instead of the rigid 3D action cuboid proposals as in (Kalogeiton et al., 2017). Based on the deformable region tube proposals, we can model the flexibility of actions. Finally, a tube localization network is followed to perform action classification and location regression.

Our experiments demonstrate that the proposed strong detector has two main advantages: (1) it can greatly reduce the phenomenon of missing target objects at frame-level; (2) it can gen-

erate deformable region tube proposals for modeling temporal dynamics of actions in videos. Our proposed method achieves state-of-the-art frame-mAP and video-mAP performance on UCF-Sports (Rodriguez et al., 2008), J-HMDB (Jhuang et al., 2013) and UCF-101 (Soomro et al., 2012) benchmarks, in particular at a higher overlap threshold.

To sum up, we make the following three main contributions in this paper: **(1)** A strong object detector is proposed to perform robust actor detection at frame-level. Meanwhile, a stabilized action localization network is designed for easier training. **(2)** We introduce an anchor refine branch at the end of the actor detection backbone network to generate deformable region tube proposals, which is flexible for modeling actions in real-world scenarios. **(3)** Finally, extensive experiments on widely used UCF-Sports, J-HMDB and UCF-101 datasets demonstrate that the object detector is very important in the pipeline of action localization, and our proposed method outperforms previous state-of-the-art action localization methods by a large margin.

The rest of this paper is organized as follows. We review the related works about action localization in Section 2. In Section 3, we provide the detailed information of the strong object detector and the stabilized action localization network. In Section 4, extensive experiments on widely used spatio-temporal action localization benchmarks are conducted to verify the effectiveness of the proposed method. Finally, we conclude the paper and provide the further work in Section 5.

2. Related Work

Since lots of methods for action localization are inspired by the methodologies of object detection in 2D images, we first review the related works about object detection, and then introduce the related works for action spatio-temporal localization.

Object detection. Object detection aims to localize the spatial location of each instance and classify the category of each target object in 2D images, and it has been widely studied in the past few decades and huge improvement has been achieved with the advent of DCNNs (He et al., 2016; Krizhevsky et al., 2012; Szegedy et al., 2015; Zhang et al., 2018a,b; Bai et al., 2018b; Zhang et al., 2019a,b) when compared with traditional methods (Felzenszwalb et al., 2010; Fidler et al., 2013; Wang et al., 2013; Zhang et al., 2017). R-CNN (Girshick et al., 2014) first casts the task of object detection as a region classification problem, and it can be viewed as a milestone of object detection. Then, Fast-RCNN (Girshick, 2015) is introduced to improve the efficiency by sharing the computation cost of the proposal generation and classification processes. By training the proposal generation branch and detection branches (*i.e.* classification and regression) in an end-to-end manner, Faster-RCNN (Ren et al., 2017) achieves a satisfactory result in both performance and efficiency. Over time, some RCNN-based follow-up works (Dai et al., 2016; Singh et al., 2017; He et al., 2017; Lin et al., 2017) are proposed to boost the detection performance. In order to speed up the detection efficiency, some one-stage object detectors are proposed, such as YOLO (Redmon et al., 2016), SSD (Liu et al., 2016) and some subsequent methods (Fu et al., 2017; Li and Zhou, 2017; Zheng

et al., 2018; Redmon and Farhadi, 2017, 2018), which directly classify image regions or anchors into specific categories and regress bounding-boxes in a dense manner without a step of generating proposals.

Another direction for object detection is based on saliency detection. For instance, (Jian et al., 2014) proposes a visual-attention-aware model for the task of salient-object detection. (Jian et al., 2018b) integrates the holistic center-directional map with the principal local color contrast map to build a computational model for saliency detection.

In this paper, we adopt the SSD framework with high efficiency to design our frame-level detector for action localization. However, compare to the general object detector in (Kalogeiton et al., 2017), we learn a strong detector by introducing the skip connection and adding some batch normalization layers in the designed deep architecture. Moreover, we reinforce the object detector in the pipeline of action localization by introducing an anchor refine branch for generating deformable region tube proposals to model flexible actions in videos. Meanwhile, different from the saliency detection methods (Jian et al., 2014, 2018a,b), we make use of two-stream data (*i.e.* RGB frames and optical flow) to perform frame-level object detection.

Action localization. Initial methods treat spatio-temporal action localization as a proposal classification problem (Escorcia et al., 2016), and they first generate hundreds of action proposals for each video and then perform classification based these action proposals. Currently, most of approaches rely on object detectors to detect actors at frame-level or clip-level, and then link detections to form action tubes. Here, we will go through these two directions (*i.e.* link detection at frame-level, and link detection at clip-level). In the methods of linking detections at frame-level, (Gkioxari and Malik, 2015) generates region proposals by selective search (Uijlings et al., 2013), and then links the proposals across frame to form action tubes. (Weinzaepfel et al., 2015) proposes a tracking-by-detection method to track proposals for action detection. (Saha et al., 2016) introduces a two-pass dynamic linking approach to construct spatio-temporal action tubes based on the appearance and motion detection boxes. (Peng and Schmid, 2016) extracts proposals from a two-stream network and then classifies and regresses them with fused features. (Singh et al.) proposes an online linking algorithm, and a real-time action localization method using SSD is introduced. To further model the temporal information, some action localization methods are proposed by linking detections at clip-level. For instance, (Kalogeiton et al., 2017) inputs a sequence of frames and outputs some anchors cuboids, which are further classified and regressed for the purpose of action detection. (Hou et al., 2017) first extends 2D Region-of-Interest pooling to 3D Tube-of-Interest (ToI) pooling, and then performs action detection with 3D convolution. For incorporating temporal contextual information, (Li et al., 2018) proposes a recurrent tubelet proposal and recognition (RTPR) networks for the task of action detection in videos.

In this paper, we follow the method in (Kalogeiton et al., 2017) to localize actions. However, thanks to the proposed strong detector, we can reducing the phenomenon of missing target objects, and can generate deformable region tube propos-

als for more accurate action localization instead of the fixed 3D action cuboid proposals as in (Kalogeiton et al., 2017).

3. Proposed method

In this section, we present a strong detector for spatio-temporal action localization in videos. An overview of our pipeline is shown in Figure 1. To detect actors in videos at different scales, we concatenate some intermediate convolutional feature maps. In order to make the network training process easier and more stable, we introduce some normalization layers nearby the concatenation operation. More importantly, we add an anchor refine branch at the end of detection backbone network for generating deformable actor-based proposals, which can not only capture actors in a flexible manner but also can encode the spatio-temporal information. Finally, some action detection layers behind the detection network are designed to further classify and regress the candidate tube proposals, which can be trained with object detection network in an end-to-end way in our implementation.

Anchor refine branch. Some SSD-based action detection methods directly classify the prior defined anchors and regress the location of actions. For example, (Kalogeiton et al., 2017) first generates some fixed anchor cuboids based on the frame-level anchors with different prior defined scales and aspect ratios, which may result in a failed detection when facing some quick motions. To address this problem, we add an anchor refine branch (ARB) at end of the detection backbone network, as shown in Figure 1 (a). ARB is designed to refine the imputed frame-level anchors, and it can predict an approximate location for each actor in videos. For each frame in a video clip, ARB outputs a set of region proposals with a confidence score to represent how likely it belongs to an actor. Following a dynamic programming algorithm in (Li et al., 2018), we generate deformable action cuboids by linking these generated regions. The generated deformable anchor cuboids, as shown in Figure 1 (b), can model the flexibility of actions when the spatial positions and scales of actors vary significantly along time. All convolutional layers except the last layer in the ARB are shared with the detection backbone network, so it will not increase so much computation cost. The detailed information of ARB is the same as region proposal network (RPN) in (Ren et al., 2017), and the loss function of ARB is defined as:

$$Loss_{ARB} = \frac{1}{N_{cls}} \sum_i L_{cls}(p_i, p_i^*) + \lambda \frac{1}{N_{reg}} \sum_i p_i^* L_{reg}(t_i, t_i^*), \quad (1)$$

where i is the index of anchors in each frame, p_i is the probability of anchor i being an actor, and ground truth label p_i^* equals 1 if the anchor is positive otherwise 0. t_i is the parameterized coordinates of predicted bounding-boxes, t_i^* is the corresponding regression targets. The classification loss L_{cls} is a *log* loss, and the regression loss is smooth *L1* loss. For more detailed information about classification and regression losses, please refer to (Ren et al., 2017).

Skip connections. Skip connections are widely used in the task of object detection (Liu et al., 2016; Lin et al., 2017; Bai et al., 2018a) in 2D images for detecting target objects with different scales, especially for some small sized objects. However,

Fig. 1. Overview of our proposed method. For a sequence of frames (*i.e.* a clip), a VGG backbone network is used to extract feature maps, which are stacked to predict categories and to refine the coordinates of anchor cuboids. An anchor refine branch (ARB) is introduced to generate deformable anchor cuboids as in (b). To reduce the internal covariate shift in the network, we add some batch normalization (BN) layers before each concatenating operation and classification & regression layers. *cat.b1* and *cat.b2* are two concat feature maps. (b) is the generated deformable anchor cuboid. (c) is the achieved action tubelet in the videos. Best seen on the computer, in color and zoomed in.

this trick is usually overlooked in the areas of action detection and action localization such as (Kalogeiton et al., 2017). To detect those small target objects in videos, in this paper, we concatenate the feature maps of *conv4_3*, which includes rich high-resolution but semantically weak features, with *conv5_3* and *fc7* layer, called *cat.b1* in Figure 1. Meanwhile, in order to detect subjects with different scales, we also concatenate the feature maps of *fc7* and *conv6_2*, termed as *cat.b2* in Figure 1. The ablation study experiments in section 4 will show the effectiveness of introducing skip connection to the frame-level detection network in the pipeline of the action localization.

Batch normalization layers. As shown in Figure 1, the proposed action localization network is a deep architecture, which includes nine convolutional layers, two fully connected layers, and some concatenated and stacked layers. Even though stochastic gradient is simple and effective, it requires careful tuning of the hyper-parameters of the designed network, especially for the learning rate used in optimizer, as well as the initial values for each layer. To overcome these issues, we add a batch normalization layer after the intermediate convolutional layers, which are concatenated to detect different sized objects at frame-level. At the same time, we also add a batch normalization layer before classification and regression layers. The added batch normalization layers can greatly reduce internal covariate shift of the network. More importantly, the batch normalization layers allow us to use much higher learning rates and be less careful about initialization to train the proposed action localization network.

Action detection network. Our action detection network takes cuboid proposal representations as input and then performs action classification and spatio location regression. In particular, two sibling layers are designed at the end of the de-

tection network to predict action classes and actor positions respectively. The first layer outputs a probability set for each cuboid proposal over $C + 1$ categories, *i.e.* $p = (p_0, p_1, \dots, p_C)$ where C denotes the number of action categories, while the second sibling layer outputs the bounding-boxes offsets for the t_{th} cuboid proposal, *i.e.* $\delta_t = (\delta x_t, \delta y_t, \delta w_t, \delta h_t)$.

Tubelets generation. We would like to note that we are not focus on the linking algorithm, so we just use the frame linking algorithm of (Li et al., 2018) to generate action tubelets. Specifically, for each inputted K ($K=6$) frames, we extract only the N ($N = 10$) highest scored tubelets for each class, which are selected by a non-maximum suppression (NMS) processing at a threshold 0.3.

Linking tubelets. Once achieving action tubelets, we follow the link algorithm of (Kalogeiton et al., 2017) to build the final spatio-temporal action tubes. Concretely, in the first frame of a video, a link is started for each of the N tubelets. To extend a link L , we select the tubelet candidate t that meets the following three conditions: (1) t is not selected by other links, (2) t has the highest score, and (3) the overlap between a link and a tubelet verifies $ov(L, t) > \theta, \theta = 0.2$. A link will be stopped when these conditions are not met for more than $K-1$ consecutive frames.

We claim that this post-processing link algorithm is not optimal, which may result in some failed results. May be a good idea is to use a network to learn the interconnections of these action tubelets and automatically generate action tubes for a video, and we put this work in the future.

4. Experiments

In this section, we experimentally demonstrate the importance of the object detector in the pipeline of action localiza-

Fig. 2. Examples of target objects detected by the baseline detector and our proposed detector respectively. The solid yellow bounding-box denotes the detections and the dotted red bounding-box represents the missing target objects. Best seen on the computer, in color and zoomed in.

tion, and validate the effectiveness of our proposed method on three public action detection benchmarks, *i.e.* UCF-Sports (Rodriguez et al., 2008), J-HMDB (Jhuang et al., 2013) and UCF-101 (Soomro et al., 2012). First, we present the implementation details of the proposed method. Then, some ablation study experiments are conducted to investigate the influence of each component in our proposed method. Finally, we compare our method with previous state-of-the-art methods.

4.1. Datasets and evaluation metrics

Dataset. *UCF – Sports* (Rodriguez et al., 2008) includes 150 videos from 10 sports categories. The videos are trimmed, and we use the train/test split of (Rodriguez et al., 2008) to train and evaluate our model. *J – HMDB* (Jhuang et al., 2013) includes 928 videos from 21 categories such as brush hair and swing baseball. The videos are trimmed, and we use the standard splits to train our model and report average results over three splits defined in (Jhuang et al., 2013). *UCF–101* (Soomro et al., 2012) includes 3207 videos from 24 sports categories, in which spatio-temporal annotations are provided. The videos are not trimmed, and we follow (Gkioxari and Malik, 2015; Peng and Schmid, 2016; Saha et al., 2016) to use split 1 to train and test our model.

Evaluation metrics. We adopt *frame – mAP* and *video – mAP* as our metrics to evaluate the trained model, which can evaluate the quality of the detections independence of the link algorithm. Specifically, a frame or an action tube is correct when IoU with a ground-truth bound box or tube is larger than a threshold (*e.g.* 0.5), at the same time, the action label is predicted correctly.

4.2. Implementation details

Following the standard implementation methods for action localization (Peng and Schmid, 2016; Saha et al., 2016; Weinzaepfel et al., 2015), we adopt a two-stream (*i.e.* RGB and flow) detector to localize actors in videos. An appearance detector and a motion detector are trained respectively, which take a sequence of K RGB frames and K consecutive flow images as

inputs. We set K to 6 in our paper by cross-validation experiments. We employ pre-trained VGG (Szegedy et al., 2015) as our backbone network, and further add some extra convolutional layers for action classification and localization. In ARB, we assign positive labels to anchors with an IoU larger than 0.7 and consider other anchors as negatives, and other details are the same as RPN in (Ren et al., 2017). The learning rate is set to 0.001, and the maximum training iteration is 40,000/120,000/600,000 for UCF-Sports/J-HMDB/UCF-101 dataset, respectively. All experiments are conducted on an NVIDIA P100 GPU.

4.3. Ablation studies

To verify the importance of the object detector in the task of action localization, we first compare our proposed method with the baseline method (Kalogeiton et al., 2017), in which a simplified SSD is employed as the object detector. Then, we demonstrate the positive influence of anchor refine branch (ARB) in the detection network by comparing the frame-mAP performance with/without this branch. Finally, some ablation study experiments are conducted to prove the effectiveness of each component (*i.e.* skip connection and batch normalization) in our proposed method by cumulatively adding each of them to the baseline method. Unless otherwise stated, all the ablation study experiments are conducted on UCF-Sports dataset and the frame-mAP with 0.5 IoU threshold is reported.

Importance of the detector in the task of action localization. We validate the importance of the detector in the task of action localization by conducting an ablation experiment between our proposed method and a baseline method where a weaker object detector is used to detect target subjects at frame-level in videos. Table 1 (the 1st and 4th rows) indicates that our proposed method improves the frame-mAP by 2.1% absolutely. We thank the high-quality action tubelets predicted by the proposed strong object detector for this huge improvement, which confirms the positive influence of the strong object detector in the pipeline of the action localization. Figure 2 shows that our proposed method can indeed reduce the phenomenon of missing target objects at frame-level when comparing with the baseline detector.

Effectiveness of the anchor refine branch. To validate the contribution of the introduced anchor refine branch, we conduct an experiment by adding the anchor refine branch to the backbone network of the baseline method. From Table 1 (the 1st and 2nd rows), we can see that the frame-mAP improves by 1.0% absolutely (*i.e.* from 87.7% to 88.7%). In particular, the performance of some flexible actions such as kicking increases by more than 10%. The reason is that the anchor refine branch can generate deformable anchor cuboids, which can model those hard actions for better classification and localization.

Effectiveness of the skip connections. To validate the effectiveness of the skip connection in our detection network, we also conduct an ablation experiment by adding this trick to the baseline method. From Table 1 (the 2nd and 3rd rows), we can observe that the frame-mAP increases to 89.1% from 88.7%. Upon closely inspecting the per-class performance, we find that the performance of some classes increases dramatically. For example, the frame-mAP performance of walk increases by more

Table 1. Ablation study results (frame-mAP %) on the UCF-Sports dataset, where “BN” denotes the batch normalization layers.

Method	ARB	Skip Connection	BN	Diving	Golf	Kicking	Lifting	Riding	Run	SkateBoarding	Swing1	Swing2	Walk	mAP
Baseline	✗	✗	✗	99.4	89.3	67.5	100.0	100.0	88.8	65.5	98.4	99.5	68.4	87.7
Ours	✓	✗	✗	98.8	84.4	78.0	99.9	100.0	92.5	63.8	100.0	99.9	69.2	88.7
	✓	✓	✗	100.0	88.0	69.3	99.5	100.0	91.8	62.7	99.8	100.0	79.6	89.1
	✓	✓	✓	100.0	89.2	72.3	99.3	100.0	92.6	65.3	99.6	100.0	79.8	89.8

Table 2. Comparison of frame-mAP to state-of-the-art methods when the detection threshold equal 0.5 on UCF-Sports, J-HMDB and UCF-101 datasets respectively.

Methods	UCF-Sports	J-HMDB	UCF-101
Gkioxari <i>et. al</i> (Gkioxari and Malik, 2015)	68.1	36.2	-
Weinzaepfel <i>et. al</i> (Weinzaepfel <i>et al.</i> , 2015)	71.9	45.8	35.8
Peng <i>et. al</i> (Peng and Schmid, 2016)	82.3	56.9	64.8
Kalogeiton <i>et. al</i> (Kalogeiton <i>et al.</i> , 2017)	87.7	65.7	67.1
Hou <i>et. al</i> (Hou <i>et al.</i> , 2017)	86.7	61.3	-
Ours	89.8	68.2	69.0

than 10% absolutely, and the reason is that the scale of actors in these classes varies significantly. Clearly, this validates the claim that the skip connection operation can detect the target subjects with different scales, which is common in real-world scenarios.

Effectiveness of the batch normalization layers. We validate the contribution of batch normalization layers by designing a comparison experiment with/without them in the proposed methods. From Table 1 (the 3rd and 4th rows), we can see that the frame-mAP improves by 0.7%, which demonstrates the effectiveness of the batch normalization layers. We would like to stress that the performance of almost all classes has a slight improvement by adding batch normalization layers. As mentioned above the reason is that the batch normalization layers can reduce internal covariate shift between each convolutional layer and they can make network training procedure more stable. It is essential to use batch normalization in the complex networks, so we include it in our network for action localization.

4.4. Comparison with state-of-the-art methods

In this section, we compare our proposed method with some previous state-of-the-art action localization methods (Gkioxari and Malik, 2015; Weinzaepfel *et al.*, 2015; Peng and Schmid, 2016; Saha *et al.*, 2016; Singh *et al.*; Kalogeiton *et al.*, 2017; Hou *et al.*, 2017; Zheng *et al.*, 2018) on UCF-Sports, J-HMDB and UCF-101 datasets, respectively.

Frame-mAP. Table 2 shows the comparison results in terms of frame-mAP when the detection threshold equals 0.5 on these three datasets. From Table 2, we can see that our method achieves the best result when comparing with other methods. Specifically, we achieve the highest frame-mAP performance on the UCF-Sports, outperforming the second best method by 2.1% absolutely. Similarly, the performance of our method surpasses the previous state-of-the-art method by 2.5% and 1.9% on J-HMDB and UCF-101 datasets respectively. The boosted performance, especially on the untrimmed dataset UCF-101, validates the effectiveness of the proposed method for action localization, as well as highlights the importance of the object detector in the pipeline of action localization.

Video-mAP. Table 3 reports video-mAP performance of our method and some previous state-of-the-art methods at different

Table 3. Comparison of video-mAP to state-of-the-art methods at different detection threshold (*i.e.* 0.5 and 0.75) on UCF-Sports, J-HMDB and UCF-101 datasets respectively.

Methods	UCF-Sports		J-HMDB		UCF-101	
	0.5	0.75	0.5	0.75	0.5	0.75
(Gkioxari and Malik, 2015)	75.8	-	53.3	-	-	-
(Weinzaepfel <i>et al.</i> , 2015)	90.5	-	60.7	-	-	-
(Peng and Schmid, 2016)	94.8	47.3	70.6	48.2	35.9	1.6
(Saha <i>et al.</i> , 2016)	-	-	71.5	43.3	35.9	7.9
(Singh <i>et al.</i>)	-	-	72.0	44.5	46.3	15.0
(Kalogeiton <i>et al.</i> , 2017)	92.7	78.4	73.7	52.1	51.4	22.7
(Hou <i>et al.</i> , 2017)	95.2	-	76.9	-	-	-
(He <i>et al.</i> , 2018)	95.7	-	77.0	-	-	-
Ours	96.6	82.6	77.2	56.6	53.8	26.0

detection IoU thresholds (*i.e.* 0.5 and 0.75) on these widely used benchmarks. From Table 3, we can observe that we also achieve the highest performance on all datasets when the IoU threshold is set to 0.5, outperforming the state-of-the-art method by 0.9%/0.2%/2.4% absolutely on UCF-Sport, J-HMDB and UCF-101 respectively. In particular, our method is more efficient when IoU threshold comes to a high value 0.75. As shown in Table 3, when detection IoU threshold equals 0.75, our video-mAP performance surpasses the second best method by 4.2%/4.5%/3.3% on UCF-Sport, J-HMDB and UCF-101 respectively, which further demonstrates the effectiveness of our proposed. We summarize the reasons for these improvement as following: (1) the skip connection of some intermediate feature maps can detect those hard target actors at frame-level (*e.g.* the scale of several objects varies greatly in one frame). (2) The ARB in the network can generate deformable anchor cuboids, which can model the flexibility of actions (*i.e.* the spatial positions and scales of humans might vary significantly along time). (3) The introduced batch normalization layers can reduce the internal covariate shift between each layer, and can make training procedure more easier and stable. Based on the above comparison results and analysis, we can confirm that our proposed method is useful in the task of action localization, and meanwhile it can exactly be embedded in some existing action detection methods.

4.5. Efficiency analysis

We compare the runtime of our proposed method the frame-based SSD approach of (Singh *et al.*), the frame-based Faster R-CNN methods (Peng and Schmid, 2016; Saha *et al.*, 2016), and our baseline detector (Kalogeiton *et al.*, 2017). The methods of (Peng and Schmid, 2016; Saha *et al.*, 2016) run at 4fps and the method of (Singh *et al.*) runs at 25-30fps. The baseline ACT-detector also runs at 25-30fps. Compare with the baseline detector (Kalogeiton *et al.*, 2017), in our proposed method, the ARB shares the convolutional layers with the VGG backbone network, and the only additional computation cost is the introduced skip connection and the batch normalization layers. We report our runtime on a single P100 GPU, and our proposed

method runs at 22-28fps, which demonstrates the efficiency of our proposed method.

5. Conclusion and future work

In this paper, we verify a claim that the object detector is very important in the task of action localization. Meanwhile, a strong object detector is introduced in the pipeline of action localization, where skip connection and batch normalization are used. The proposed strong detector can detect actors with different scales in videos, as well as make training procedure more stable. Moreover, an anchor refine branch is introduced at the end of the detection backbone network, which can generate deformable anchor cuboids instead of fixed anchor cuboid proposals as previous methods. With the deformable anchor cuboids, we can locate some challenging actions where the spatial positions and scales of actors might vary significantly along time. Extensive experiments on UCF-Sports, J-HMDB and UCF-101 validate the effectiveness of our proposed method, and we achieve the highest performance in terms of frame-mAP and video-mAP on these three datasets, outperforming the previous state-of-the-art methods by a large margin, especially at a higher detection IoU threshold.

In the future, we plan to propose an online link network to form action tubes (*i.e.* a convolutional neural network is designed to link action tubelets predicted from the detection network for the purpose of forming action tubes in videos), which can be trained with the detection network in an end-to-end manner.

Acknowledgment

This work was supported by Natural Science Foundation of China, Grant No. 61603372.

References

- Afonso, M., Zhang, F., Bull, D.R., 2019. Video compression based on spatio-temporal resolution adaptation. *IEEE Transactions on Circuits and Systems for Video Technology* 29, 275–280.
- Alwassel, H., Heilbron, F.C., Ghanem, B., 2018. Action search: Spotting actions in videos and its application to temporal action localization, in: *European Conference on Computer Vision*, Springer. pp. 253–269.
- Bai, Y., Zhang, Y., Ding, M., Ghanem, B., 2018a. Finding tiny faces in the wild with generative adversarial network. *CVPR*. IEEE.
- Bai, Y., Zhang, Y., Ding, M., Ghanem, B., 2018b. Sod-mtgan: Small object detection via multi-task generative adversarial network, in: *European Conference on Computer Vision*, Springer. pp. 210–226.
- Caba Heilbron, F., Lee, J.Y., Jin, H., Ghanem, B., 2018. What do i annotate next? an empirical study of active learning for action localization, in: *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 199–216.
- Chen, Y., Li, D., Zhang, J.Q., 2017. Complementary color wavelet: a novel tool for the color image/video analysis and processing. *IEEE Transactions on Circuits and Systems for Video Technology*.
- Dai, J., Li, Y., He, K., Sun, J., 2016. R-fcn: Object detection via region-based fully convolutional networks, in: *NIPs*, pp. 379–387.
- Escorcia, V., Heilbron, F.C., Niebles, J.C., Ghanem, B., 2016. Daps: Deep action proposals for action understanding, in: *European Conference on Computer Vision*, Springer. pp. 768–784.
- Felzenszwalb, P.F., Girshick, R.B., McAllester, D., Ramanan, D., 2010. Object detection with discriminatively trained part-based models. *IEEE transactions on pattern analysis and machine intelligence* 32, 1627–1645.
- Fidler, S., Mottaghi, R., Yuille, A., Urtasun, R., 2013. Bottom-up segmentation for top-down detection, in: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3294–3301.
- Fu, C.Y., Liu, W., Ranga, A., Tyagi, A., Berg, A.C., 2017. Dssd: Deconvolutional single shot detector. preprint arXiv:1701.06659.
- Giancola, S., Amine, M., Dghaily, T., Ghanem, B., 2018. Soccernet: A scalable dataset for action spotting in soccer videos. arXiv preprint arXiv:1804.04527.
- Girshick, R., 2015. Fast r-cnn, in: *IEEE ICCV*, pp. 1440–1448.
- Girshick, R., Donahue, J., Darrell, T., Malik, J., 2014. Rich feature hierarchies for accurate object detection and semantic segmentation, in: *IEEE CVPR*, pp. 580–587.
- Gkioxari, G., Malik, J., 2015. Finding action tubes, in: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 759–768.
- He, J., Deng, Z., Ibrahim, M.S., Mori, G., 2018. Generic tubelet proposals for action localization, in: *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, IEEE. pp. 343–351.
- He, K., Gkioxari, G., Dollár, P., Girshick, R., 2017. Mask r-cnn, in: *IEEE ICCV*, IEEE. pp. 2980–2988.
- He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition, in: *IEEE CVPR*, pp. 770–778.
- Herglotz, C., Heindel, A., Kaup, A., 2017. Decoding-energy-rate-distortion optimization for video coding. *IEEE Transactions on Circuits and Systems for Video Technology*.
- Hou, R., Chen, C., Shah, M., 2017. Tube convolutional neural network (t-cnn) for action detection in videos, in: *Proceedings of the IEEE International Conference on Computer Vision*, pp. 5822–5831.
- Hu, W., Tan, T., Wang, L., Maybank, S., 2004. A survey on visual surveillance of object motion and behaviors. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)* 34, 334–352.
- Ioffe, S., Szegedy, C., 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift. arXiv preprint arXiv:1502.03167.
- Jhuang, H., Gall, J., Zuffi, S., Schmid, C., Black, M.J., 2013. Towards understanding action recognition, in: *Proceedings of the IEEE international conference on computer vision*, pp. 3192–3199.
- Jian, M., Lam, K.M., Dong, J., Shen, L., 2014. Visual-patch-attention-aware saliency detection. *IEEE transactions on cybernetics* 45, 1575–1586.
- Jian, M., Qi, Q., Dong, J., Yin, Y., Lam, K.M., 2018a. Integrating qdwd with pattern distinctness and local contrast for underwater saliency detection. *Journal of visual communication and image representation* 53, 31–41.
- Jian, M., Zhang, W., Yu, H., Cui, C., Nie, X., Zhang, H., Yin, Y., 2018b. Saliency detection based on directional patches extraction and principal local color contrast. *Journal of Visual Communication and Image Representation* 57, 1–11.
- Kalogeiton, V., Weinzaepfel, P., Ferrari, V., Schmid, C., 2017. Action tubelet detector for spatio-temporal action localization, in: *2017 IEEE International Conference on Computer Vision (ICCV)*, IEEE. pp. 4415–4423.
- Khelifi, F., Bouridane, A., 2017. Perceptual video hashing for content identification and authentication. *IEEE Transactions on Circuits and Systems for Video Technology*.
- Krizhevsky, A., Sutskever, I., Hinton, G.E., 2012. Imagenet classification with deep convolutional neural networks, in: *NIPs*, pp. 1097–1105.
- Laptev, I., Pérez, P., 2007. Retrieving actions in movies, in: *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*, IEEE. pp. 1–8.
- Li, D., Qiu, Z., Dai, Q., Yao, T., Mei, T., 2018. Recurrent tubelet proposal and recognition networks for action detection, in: *European Conference on Computer Vision*, Springer. pp. 306–322.
- Li, Z., Zhou, F., 2017. Fssd: Feature fusion single shot multibox detector. preprint arXiv:1712.00960.
- Lin, T.Y., Dollár, P., Girshick, R.B., He, K., Hariharan, B., Belongie, S.J., 2017. Feature pyramid networks for object detection., in: *CVPR*, p. 3.
- Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.Y., Berg, A.C., 2016. Ssd: Single shot multibox detector, in: *ECCV*, Springer. pp. 21–37.
- Ma, C., Liu, D., Peng, X., Wu, F., Li, H., Wang, T., 2017. Reference clip for inter prediction in video coding. *IEEE Transactions on Circuits and Systems for Video Technology*.
- Oh, S., Hoogs, A., Perera, A., Cuntoor, N., Chen, C.C., Lee, J.T., Mukherjee, S., Aggarwal, J., Lee, H., Davis, L., et al., 2011. A large-scale benchmark dataset for event recognition in surveillance video, in: *Computer vision and pattern recognition (CVPR), 2011 IEEE conference on*, IEEE. pp. 3153–3160.

- Peng, X., Schmid, C., 2016. Multi-region two-stream r-cnn for action detection, in: European Conference on Computer Vision, Springer. pp. 744–759.
- Peng, Y., Zhao, Y., Zhang, J., 2018. Two-stream collaborative learning with spatial-temporal attention for video classification. *IEEE Transactions on Circuits and Systems for Video Technology* 29, 773–786.
- Redmon, J., Divvala, S., Girshick, R., Farhadi, A., 2016. You only look once: Unified, real-time object detection, in: *IEEE CVPR*, pp. 779–788.
- Redmon, J., Farhadi, A., 2017. Yolo9000: better, faster, stronger. *arXiv preprint* .
- Redmon, J., Farhadi, A., 2018. Yolov3: An incremental improvement. *preprint arXiv:1804.02767* .
- Ren, S., He, K., Girshick, R., Sun, J., 2017. Faster r-cnn: towards real-time object detection with region proposal networks. *IEEE TPAMI* , 1137–1149.
- Rodriguez, M.D., Ahmed, J., Shah, M., 2008. Action mach a spatio-temporal maximum average correlation height filter for action recognition, in: *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, IEEE. pp. 1–8.
- Saha, S., Singh, G., Sapienza, M., Torr, P.H., Cuzzolin, F., 2016. Deep learning for detecting multiple space-time action tubes in videos. *arXiv preprint arXiv:1608.01529* .
- Shiau, Y.H., Kuo, Y.T., Chen, P.Y., Hsu, F.Y., 2017. Vlsi design of an efficient flicker-free video defogging method for real-time applications. *IEEE Transactions on Circuits and Systems for Video Technology* .
- Singh, B., Li, H., Sharma, A., Davis, L.S., 2017. R-fcn-3000 at 30fps: Decoupling detection and classification. *preprint arXiv:1712.01802* .
- Singh, G., Saha, S., Sapienza, M., Torr, P., Cuzzolin, F., . Online real time multiple spatiotemporal action localisation and prediction on a single platform. *arXiv preprint* .
- Soomro, K., Zamir, A.R., Shah, M., 2012. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402* .
- Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A., 2015. Going deeper with convolutions, in: *IEEE CVPR*, pp. 1–9.
- Tian, Y., Sukthankar, R., Shah, M., 2013. Spatiotemporal deformable part models for action detection, in: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2642–2649.
- Uijlings, J.R., Van De Sande, K.E., Gevers, T., Smeulders, A.W., 2013. Selective search for object recognition. *International journal of computer vision* 104, 154–171.
- Venugopalan, S., Rohrbach, M., Donahue, J., Mooney, R., Darrell, T., Saenko, K., 2015. Sequence to sequence-video to text, in: *Proceedings of the IEEE international conference on computer vision*, pp. 4534–4542.
- Wang, L., Qiao, Y., Tang, X., Van Gool, L., 2016. Actionness estimation using hybrid fully convolutional networks, in: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2708–2717.
- Wang, X., Yang, M., Zhu, S., Lin, Y., 2013. Regionlets for generic object detection, in: *Proceedings of the IEEE international conference on computer vision*, pp. 17–24.
- Weinzaepfel, P., Harchaoui, Z., Schmid, C., 2015. Learning to track for spatio-temporal action localization, in: *Proceedings of the IEEE international conference on computer vision*, pp. 3164–3172.
- Yang, Z., Gao, J., Nevatia, R., 2017. Spatio-temporal action detection with cascade proposal and location anticipation. *arXiv preprint arXiv:1708.00042* .
- Yao, L., Torabi, A., Cho, K., Ballas, N., Pal, C., Larochelle, H., Courville, A., 2015. Describing videos by exploiting temporal structure, in: *Proceedings of the IEEE international conference on computer vision*, pp. 4507–4515.
- Yuan, J., Liu, Z., Wu, Y., 2009. Discriminative subvolume search for efficient action detection, in: *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, IEEE. pp. 2442–2449.
- Yuan, Y., Mei, T., Cui, P., Zhu, W., 2017. Video summarization by learning deep side semantic embedding. *IEEE Transactions on Circuits and Systems for Video Technology* .
- Zhang, H., Ma, Z., 2014. Fast intra mode decision for high efficiency video coding (hevc). *IEEE Transactions on circuits and systems for video technology* 24, 660–668.
- Zhang, J., Peng, Y., 2019a. Hierarchical vision-language alignment for video captioning, in: *International Conference on Multimedia Modeling*, Springer. pp. 42–54.
- Zhang, J., Peng, Y., 2019b. Object-aware aggregation with bidirectional temporal graph for video captioning, in: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 8327–8336.
- Zhang, Y., Bai, Y., Ding, M., Li, Y., Ghanem, B., 2018a. W2f: A weakly-supervised to fully-supervised framework for object detection, in: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 928–936.
- Zhang, Y., Bai, Y., Ding, M., Li, Y., Ghanem, B., 2018b. Weakly-supervised object detection via mining pseudo ground truth bounding-boxes. *Pattern Recognition* 84, 68–81.
- Zhang, Y., Ding, M., Bai, Y., Ghanem, B., 2019a. Detecting small faces in the wild based on generative adversarial network and contextual information. *Pattern Recognition* .
- Zhang, Y., Ding, M., Bai, Y., Xu, M., Ghanem, B., 2019b. Beyond weakly-supervised: Pseudo ground truths mining for missing bounding-boxes object detection. *IEEE Transactions on Circuits and Systems for Video Technology* .
- Zhang, Y., Ding, M., Fu, W., Li, Y., 2017. Reading recognition of pointer meter based on pattern recognition and dynamic three-points on a line, in: *Ninth International Conference on Machine Vision (ICMV 2016)*, International Society for Optics and Photonics. p. 103410K.
- Zhao, Y., Peng, Y., 2017. Saliency-guided video classification via adaptively weighted learning, in: *2017 IEEE International Conference on Multimedia and Expo (ICME)*, IEEE. pp. 847–852.
- Zheng, L., Fu, C., Zhao, Y., 2018. Extend the shallow part of single shot multi-box detector via convolutional neural network. *preprint arXiv:1801.05918* .