BLP - BOUNDARY LIKELIHOOD PINPOINTING NETWORKS FOR ACCURATE TEMPORAL ACTION LOCALIZATION

Weijie Kong[1], Nannan Li[1], Shan Liu[3], Thomas Li[2], Ge Li[1]

[1]School of Electronic and Computer Engineering, Peking University, Shenzhen, China.
[2] Advanced Institute of Information Technology, Peking University, Hangzhou, China.
[3] Media Lab, Tencent, Shenzhen, China.

ABSTRACT

Despite tremendous progress achieved in temporal action detection, state-of-the-art methods still suffer from the sharp performance deterioration when localizing the starting and ending temporal action boundaries. Although most methods apply boundary regression paradigm to tackle this problem, we argue that the direct regression lacks detailed enough information to yield accurate temporal boundaries. In this paper, we propose a novel Boundary Likelihood Pinpointing (BLP) network to alleviate this deficiency of boundary regression and improve the localization accuracy. Given a loosely localized search interval that contains an action instance, BLP casts the problem of localizing temporal boundaries as that of assigning probabilities on each equally divided unit of this interval. These generated probabilities provide useful information regarding the boundary location of the action inside this search interval. Based on these probabilities, we introduce a boundary pinpointing paradigm to pinpoint the accurate boundaries under a simple probabilistic framework. Compared with other C3D feature based detectors, extensive experiments demonstrate that BLP significantly improves the localization performance of recent state-of-the-art detectors, and achieves competitive detection mAP on both THUMOS' 14 and ActivityNet datasets, particularly when the evaluation tIoU is high.

Index Terms— Temporal Action Detection, Temporal Action Localization, Boundary Pinpointing, Boundary Regression

1. INTRODUCTION

Recently, as an essential but challenging task in the large research scope of video analysis, temporal action detection in untrimmed videos has drawn tremendous attention from the research community [1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11]. Given a long untrimmed video consisting of multiple action instances and complex background contents, temporal action detection aims at solving two problems: (1) recognizing the action categories of actions contained in the video; (2) localizing temporal intervals (starting and ending boundaries) where actions of interest occur. Currently, temporal action detection has been applied to multiple practical applications, such as video surveillance, human-robot interaction and intelligent home care.

For temporal action detection, how to accurately localize the starting and ending boundaries of a complex action instance is a
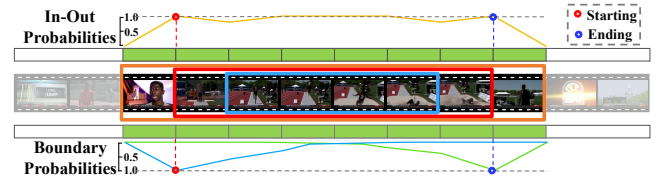
Fig. 1. **Localization process of boundary pinpointing paradigm.** A search interval is extended from an action proposal by a factor γ, and equally divided into M units. To localize the precise temporal boundary, we assign *in-out* or *boundary* probabilities to each equally divided unit of the search interval for being on the inside of an action temporal boundary or being the starting or ending boundary of the action instance. By maximizing the likelihood of each probability, we pinpoint the predicted temporal boundaries (red box).

challenging problem, since an action instance can happen at arbitrary temporal location with uncertain duration in a video of diverse length. As addressed in [12], *the localization error is the most common and the most impactful error that hampers the detection performance of existing state-of-the-art approaches.* Preferentially fixing localization errors can significantly boost the detection average-mAP. Therefore, to achieve high temporal localization accuracy, most recently detection methods [3, 4, 6, 7, 13, 14, 15] apply boundary regression paradigm to refine the boundaries given a proposal. However, we argue that trying to directly regress the action boundary temporally constitutes a difficult learning task and hardly yield accurate enough boundaries.

To alleviate this deficiency and focus on the need for improving the localization accuracy of current detection methods, we propose a novel Boundary Likelihood Pinpointing (BLP) network. The *main contribution* of BLP is that we cast the problem of localizing temporal boundaries as that of assigning probabilities on each equally divided unit of a search interval. Specifically, instead of using boundary regression, we propose a novel *boundary pinpointing* paradigm to perform accurate temporal action localization, which is implemented with three steps (see Fig. 1). First, given a loosely localized action proposal within a video, we obtain a larger search interval via extending the proposal boundaries by a factor γ and equally divide it to M units. Second, we assign one or more discrete probabilities to each unit indicating whether the unit is inside of the temporal span of action ground truth or being the starting or ending boundary of the action instance. Finally, we pinpoint the boundaries by simply maximizing the likelihood for estimating the optimal boundaries under these probabilities. Since these probabilities provide far more detailed and useful boundary information, they

would encourage the model to yield more accurate boundaries than the regression models, that just predict 2 temporal boundary coordinates. We evaluate BLP model on two challenging datasets: THUMOS'14 [16] and ActivityNet [17]. Extensive experiments demonstrate that BLP model can obtain detection results with more precise boundaries than direct regression. Integrating our BLP model with existing action classifier into detection framework leads to competitive detection mAP on both datasets, especially when the evaluation tIoU is high. ***Specifically, our detection framework achieves 34.5% (tIoU = 0.7) and 15.5× (tIoU = 0.95) relative gain over the mAP of the state-of-the-arts on THUMOS'14 and ActivityNet respectively.***

Relation to prior work. Recently, an immense amount of deep models [18, 19, 20, 21, 22] have been proposed in action recognition, among which the Two-Stream [18] and C3D [19] models are deployed in most existing methods. Due to the explosive growth of untrimmed video data, another challenging task called temporal action detection has been put to the center of attention. Currently, many approaches [6, 4, 3, 7, 14, 15] adopt a "detection by classification" framework, in which boundary regression has been widely employed for adjusting the temporal boundaries and boosting the localization accuracy. Differently from the aforementioned work, we propose the *boundary pinpointing* paradigm. This paradigm estimates the optimal boundaries by maximizing likelihood under predefined probabilities. where these probabilities can provide more useful information than dirct regression. Our idea stems from a novel *object localization methodology* called LocNet [23], which proposed to revise the horizontal and vertical object boundaries of the given proposal using border probabilities. Inspired by this work, BSN [24] also adopted similar boundary probabilities for *temporal action proposal*. However, BSN generates proposal boundaries by simply selecting temporal locations with high starting and ending probabilities separately. Our method localizes temporal boundaries using maximizing likelihood estimation under these probabilities, which can provide a more accurate measure of confidence for delimiting the boundaries on any point in time.

2. PROPOSED METHOD

2.1. Temporal Action Detection Pipeline

To begin with, we provide a brief overview of the temporal action detection pipeline. ***Our detection pipeline contains two major modules: an action classification and localization network***.

Formally, an action proposal is represented as $\psi = (B_s, B_e)$, where B_s and B_e are the starting (s) and ending (e) boundary coordinates of the segment ψ separately. Given a set of action proposals $\Psi = \{\psi_n\}_{n=1}^{N_p}$ generated by either sliding temporal window or other temporal action proposal methods [25, 14, 24], the ***action classification network*** anticipates action categories by predicting a set of classification scores $\{\{s_{nj}|c_j\}_{j=1}^{N_c}\}_{n=1}^{N_p}$. The score s_{nj} represents how likely the n-th temporal proposal is recognized to be the j-th action category c_j. Meanwhile, for each loosely localized proposal, the ***action localization network*** localizes boundaries where actions start and end temporally. It generates a new set of action segments that have more compact boundaries enclosing the actions inside the proposal. To eliminate the redundant segments, an extra Non-Maximum-Suppression (NMS) [26] operation is applied to obtain the final segments with accurate boundaries $\Psi' = \{\psi_n' = (B_{ns}', B_{ne}')\}_{n=1}^{N_d}, (N_d \leq N_p)$. Details on the localization process will be discussed in Sec. 2.2. N_c, N_d and N_p are the number of action categories, final results and proposals respectively.

2.2. Boundary Likelihood Pinpointing Network

The purpose of our work is to improve the localization accuracy of the detection pipeline. Currently, most existing detection methods [3, 4, 6, 7, 13, 14, 15] accomplish this by directly regressing two boundary coordinates, which lacks detailed enough information to yield accurate boundaries. Thus, we propose a novel **Boundary Likelihood Pinpointing (BLP)** network as our localization network.

BLP accepts selected proposal segments and outputs conditional probabilities indicating the boundary location. Given a proposal segment $\psi = (B_s, B_e)$, BLP first extends it by a factor γ to create a search interval $I = (\frac{(1+\lambda)B_s+(1-\lambda)B_e}{2}, \frac{(1-\lambda)B_s+(1+\lambda)B_e}{2})$ and equally divided it into M units. Then, BLP predicts one or more discrete probabilities for each unit to indicating whether the unit is inside of the temporal span of action ground truth or being the starting or ending boundary of the action instance. These probabilities provide more detailed information for precise boundary inference than direct boundary regression, which is detailed in Sec. 2.2.1. During inference, we propose a novel boundary pinpointing paradigm. Based on the probabilities generated by BLP, we can pinpoint the action boundaries by simply maximizing the likelihood for estimating the optimal boundaries. This paradigm is detailed in Sec. 2.2.2.

2.2.1. Boundary Likelihood Predictions

For each unit i within a search interval I, BLP predicts one or more conditional probabilities $p^{I,c} = \{p(i|I,c)\}_{i=1}^M$ corresponding to a specific category c. Here we design two types of probabilities.

In-Out probabilities: We define the *in-out* probabilities $p_{io} = \{p_{io}(i)\}_{i=1}^M$ to represent the likelihood of unit i being inside the temporal span of an action instance of category c. Ideally, given a ground truth segment $\psi_{gt} = (B_s^{gt}, B_e^{gt})$, the *in-out* probabilities p_{io} should be equal to the following target probabilities $T = \{T_{io}\}$.

$$\forall i \in \{1,...,M\}, T_{io}(i) = \begin{cases} 1, & \text{if unit } i \in [B_s^{gt}, B_e^{gt}] \\ 0, & \text{otherwise} \end{cases}$$

Boundary probabilities: $p_s = \{p_s(i)\}_{i=1}^M$ and $p_e = \{p_e(i)\}_{i=1}^M$ represent two independent probabilities of unit i being the starting and ending boundaries of an action instance for category c. Given a ground truth ψ_{gt}, the output *boundary* probabilities $p_{bd} = p_s, p_e$ should ideally equal to target probabilities $T = \{T_s, T_e\}$, where $l \in \{s, e\}$.

$$\forall i \in \{1,...,M\}, T_l(i) = \begin{cases} 1, & \text{if } B^{gt} \in \text{unit } i \\ 0, & \text{otherwise} \end{cases}$$

2.2.2. Inference by Boundary Pinpointing

Given aforementioned probabilities of I, we propose a novel boundary pinpointing paradigm to inference the temporal boundaries $\psi' = (B_s', B_e')$ of the action inside I. This process is implemented by adopting one of the following two BLP localization models.

In-Out localization model: Maximizes the likelihood of *in-out* elements of temporal boundary ψ':

$$L_{in-out}(\psi') = \prod_{i \in \{B_s,...,B_e\}} p_{io}(i) \prod_{i \notin \{B_s,...,B_e\}} \widetilde{p_{io}}(i), \quad (1)$$

where $\widetilde{p_{io}}(i) = 1 - p_{io}(i)$. The first term in the right hand of the equation represents each unit of ψ' to be inside a ground truth interval and the second term represents the likelihood of units that are not part of ψ' to be outside a ground truth interval.

Boundary localization model: Maximizes the likelihood of *boundary* elements of boundary ψ':

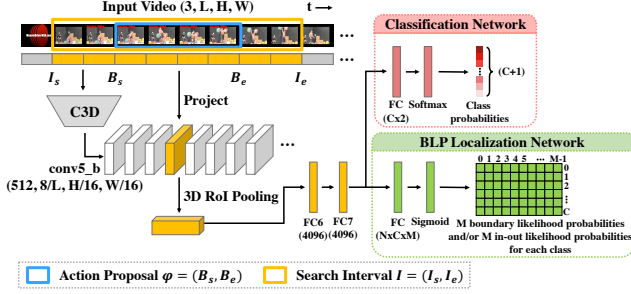$$L_{boundary}(\psi') = p_s(B_s') \cdot p_e(B_e'). \quad (2)$$

Fig. 2. Temporal action detection network architecture with BLP localization model.

2.3. Action Detection Network Architecture

The architecture of the detection network is shown in Fig. 2. Given a video sequence $\mathbf{V} \in \mathbb{R}^{3 \times L \times H \times W}$ consists of L frames and a set of action proposals $\mathbf{\Psi} = \{\psi_n\}_{n=1}^{N_p}$, the network outputs category-specific action segments with accurate temporal boundaries.

BLP localization network architecture. BLP network aims to predict aforementioned *in-out* or *boundary* probabilities for each proposal. To begin with, a deep shared C3D model [19] is utilized to process the input video $\mathbf{V}$ to extract rich spatio-temporal feature hierarchies, and outputs a shared feature map $\mathbf{F_{conv5b}} \in \mathbb{R}^{512 \times \frac{L}{8} \times \frac{H}{16} \times \frac{W}{16}}$. Then, given a search interval I extended by an action proposal, we map it on $\mathbf{F_{conv5b}}$ and use a 3D RoI pooling layer [6] to extract fixed-size feature maps (of size $512 \times 1 \times 4 \times 4$) from activation that inside I. The resulting feature maps can be fed forward into two fully connected (fc) layers of C3D and an extra fc layer to yields a 1-dimension feature vector with length $N \times M \times C$, where $N = 1, 2$ for *in-out* and *boundary* possibilities respectively, M is the number of divided units of I and C is the number of action categories. Finally, in order to output the category-specific conditional probabilities, the 1-dimension feature vector is reshaped and fed into a sigmoid layer to obtain the final conditional probability matrix with dimension $N \times M \times C$.

Action classification network architecture. For a given proposal, action classification network anticipates action categories by predicting a set of softmax scores for $(C + 1)$ categories (including "background"). To this end, the fc7 features are fed into another fc layer and an extra softmax layer to output $(C+1)$ class probabilities.

2.4. Optimization

We train the detection network by optimizing classification and localization networks jointly. The multi-task objective function is:

$$\mathcal{L}oss = \frac{1}{N_{cls}} \sum_i \mathcal{L}_{cls}(\theta_1 | a_i, a_i^*) + \frac{\lambda}{N_{loc}} \sum_j \mathcal{L}_{loc}(\theta_2 | p_{(\cdot)j}, T_{(\cdot)j}, c_j) \tag{3}$$

where N_{cls} and N_{loc} stand for batch size and number of proposal segments, respectively, and λ is the trade-off parameter and set empirically. The i and j are the indexes for action proposals, and $\theta.$ are the network parameters. For the classification network, $\mathcal{L}_{cls}$ is a standard multi-class cross-entropy loss, where a_i and a_i^* are the predicted class probability and the ground truth, respectively; while for the localization network, L_{loc} adopts a binary logistic regression loss conditioned on a specific class c, where $p_{(\cdot)j} = \{p_{(io)j}, p_{(db)j}\}$ represent evaluation probabilities of *in-out* or *boundary* for each segment, and $T_{(\cdot)j} = \{T_{(io)j}, T_{(bd)j}\}$ are the corresponding target

probabilities. Specifically, in the case of *in-out*, the loss $\mathcal{L}_{loc}$ is given by:

$$\mathcal{L}_{loc} = \sum_{j=1}^{M} T_{(io)j} log(p_{(io)j}) + \widetilde{T}_{(io)j} log(\widetilde{p}_{(io)j}), \tag{4}$$

for the *boundary* case, it is:

$$\mathcal{L}_{loc} = \sum_{bd \in \{s,e\}} \sum_{j=1}^{M} \beta^+ T_{(bd)j} log(p_{(bd)j}) + \beta^- \widetilde{T}_{(bd)j} log(\widetilde{p}_{(bd)j}), \tag{5}$$

where $\widetilde{p}_{(\cdot)j} = 1 - p_{(\cdot)j}$, and $\widetilde{T}_{(\cdot)j} = 1 - T_{(\cdot)j}$. In equation (5), we adapt the trade-off parameters $\beta^- = 0.5M/(M-1)$ and $\beta^+ = (M+1) \cdot \beta^-$ as in [23] to balance the two terms of *boundary* and *non-boundary* elements.

3. EXPERIMENTS

In this section, we evaluate proposed BLP network on two prevailing datasets: THUMOS'14 [16] and ActivityNet *v1.3* [17]. **Baseline Model:** We take R-C3D [6] as our baseline, since it's a regression-based temporal action detection method. To detect actions, we integrate examined BLP localization model and R-C3D classification model into one holistic detection framework. For a fair comparison, we train and test our detection network with the same classification network and the same proposal set generated by R-C3D for all experiments. The whole BLP model is implemented on Caffe [27].

3.1. Datasets and Experimental Details

THUMOS'14. THUMOS'14 contains 20 different sport activities, with 200 videos for training and 213 videos for testing. **Evaluation metrics.** We report the mean Average Precision (mAP) of each action category at tIoU thresholds with [0.1:0.1:0.7], and the mAP at tIoU=0.5 is used for the final comparison with other methods. **Implementation details.** The weights of C3D model are pre-trained on Sport-1M and finetuned on UCF101. The λ in loss function (3) is set to be 20. Other implementation details are the same as in [6].

ActivityNet. ActivityNet *v1.3* contains 19,994 videos with 200 classes and is divided into three sets: training, validation, testing with a ratio of 2:1:1. **Evaluation metrics.** We report the mAP at tIoU={0.5, 0.75, 0.95}, and the average of mAPs with tIoU thresholds [0.5:0.05:0.95] is used for comparison. **Implementation details.** The C3D model is initialized with the pre-trained Sport-1M weights finetuned on ActivityNet training videos. We train the BLP with a learning rate fixed at 10^{-4} for first 10 epochs and decreased to 10^{-5} for the last 5 epochs. The λ is set to be 250.

3.2. Ablation Experiments

In this section, we explore the best hyper-parameter settings for BLP.

How many units should a search interval be divided into? Given a video search interval, we divide it to M units. To explore the influence of M, we examine three *In-Out* models with $M = \{16, 32, 48\}$ (the extension factor $\gamma = 1.8$). As shown in Table 1, the detection performance with *In-Out* model achieves the best performance when $M = 32$. We analyze that with finer resolution ($M = 48$), each unit contains fewer features to determine whether the unit is inside an action of interest. Conversely, with coarse resolution ($M = 16$), each unit spans a longer time interval, therefore the temporal boundary localization may be ambiguous and less precise. The same analysis can be applied to *Boundary* models. *As a result, we choose M = 32 for the following experiments.*

How long should a proposal be extended to? A search interval is obtained by extending a temporal segment by a factor γ. Our intuitive assumption is that with larger γ, the BLP model will comprehend and leverage more surrounding temporal context. To

explore the impact of γ, we investigate six *In-Out* and *Boundary* models with $\gamma = \{1.0, 1.6, 1.8, 2.0, 2.4, 3.0\}$ ($M = 32$). As shown in Table 2, we observe that when $\gamma = 2.0$, two models achieves the peak performance, while the worst performance occurs when no context is considered ($\gamma = 1.0$). However, including redundant context ($\gamma > 2.0$) also leads to the deterioration of performance. ***Thus, we choose $\gamma = 2.0$ for the following experiments.***

Table 1. Ablation experiment results on hyper-parameter M for *In-Out* and *Boundary* models ($\gamma = 1.8$, %mAP@tIou).

tIoU	0.1	0.2	0.3	0.4	0.5
M=16	54.8	52.7	47.9	39.4	31.2
M=32	**54.9**	**52.9**	**48.5**	**40.3**	**31.6**
M=48	53.3	51.1	47.1	39.7	29.6

Table 2. Ablation experiment results on hyper-parameter γ for *In-Out* model ($M = 32$, %mAP@tIou=0.5).

γ	1.0	1.6	1.8	2.0	2.4	3.0
In-Out	30.5	31.3	31.6	**32.1**	31.8	31.7
Boundary	29.3	32.4	32.2	**32.5**	31.9	31.9

3.3. Action Localization Effectiveness Analysis

In this section, we compare the localization performance of proposed *In-Out* and *Boundary* models with the regression-based model R-C3D on THUMOS'14 testing set and ActivityNet validation set. As shown in Fig. 3, to evaluate the localization performance of examined model, we report the class-specific recall (averaging per class recalls) as a function of the tIoU thresholds with [0.05:0.05:1.0] for the final detection results generated by the corresponding detection pipeline. We also report the average recall (AR) for each model in the legend. The higher AR indicates the model can yield the more accurate temporal boundaries. ***Fig. 3 shows that two proposed models achieve remarkably higher recall than the baseline, and surpass the AR of the baseline by average 5.9% and 4.8% on both datasets respectively.*** We argue that *in-out* and *boundary* probabilities help the BLP model to yield more accurate boundaries to have larger overlap with ground truth instances. This demonstrates the effectiveness and superior localization performance of BLP model.
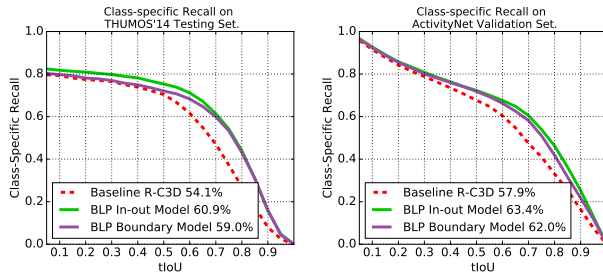

Fig. 3. Localization performance comparison on THUMOS' 14 and ActivityNet under matric: Class-Specific Recall@tIoU. For comparison, the average recall (AR) are reported in the legend.

3.4. Action Detection Performance Analysis

The detection performance is highly related to the choice of feature extractor. Since the Two-Stream features [18, 20] or other improved 3D ConvNet features [21] are more discriminative and superior than the vanilla C3D features deployed in our model, ***here we only compare with the state-of-the-art methods that adopt vanilla C3D as their feature extractors for a fair comparison.***

THUMOS'14. The comparison results on THUMOS'14 testing set are summarized in Table 3. We can observe that: (1) The

detection framework with proposed *In-Out* and *Boundary* localization models outperform the baseline R-C3D [6] by **3.2%** and **3.6%** respectively, which demonstrates that our proposed boundary pinpointing paradigm can truly boost the localization performance and yield much more accurate temporal boundaries than boundary regression. (2) When it comes to the other well-developed regression-based detection methods [14, 4], the detection performance of our emerging probability-based localization method is remarkably superior. (3) Our detector shows superior mAP over state-of-the-art methods SS-TAD [8] across a wide range of tIoU thresholds. Especially when tIoU = 0.7, we outperform SS-TAD by **35.4%** relatively. These results confirm that our well-designed probabilities can provide more useful boundary information for accurate localization.

Table 3. Temporal action detection results on THUMOS'14 testing set (%mAP@tIoU). Here we only list C3D feature based methods.

Detection Method	0.7	0.6	0.5	0.4	0.3	0.2	0.1
SCNN [1]	5.3	10.3	19.0	28.7	36.3	43.5	47.7
CBR-C3D [4]	7.9	13.8	22.7	30.1	37.7	44.3	48.2
CDC [2]	7.9	13.1	23.3	29.4	40.1	-	-
TURN + S-CNN [14]	-	-	25.6	34.9	44.1	50.9	54.0
SS-TAD [8]	9.6	-	29.2	-	45.7	-	-
R-C3D (Baseline) [6]	-	-	28.9	35.6	44.8	51.5	54.5
R-C3D + In-Out	**12.6**	**23.0**	**32.1**	**41.1**	**49.2**	**53.9**	**56.2**
R-C3D + Boundary	**13.0**	**22.3**	**32.5**	**41.3**	**48.5**	**53.0**	**54.7**

ActivityNet *v1.3*. The comparison results on the ActivityNet *v1.3* testing set are shown in Table 4. The results show that after using BLP models to refine temporal boundaries, we gain obvious improvement over the baseline R-C3D [6] in terms of all range of tIoU and the average mAP. Meanwhile, ***compared with the state-of-the-art method CDC [2], our method shows competitive performance and get $15.5\times$ relative gain when the tIoU is high (tIoU=0.95).*** This indicates that after the refinement, the segments have more precise boundaries and have larger overlap with ground truth instances.

Table 4. Temporal action detection results on ActivityNet *v1.3* testing set (%mAP@tIoU). We only list C3D feature based methods.

Detection Method	0.95	0.75	0.5	Average
Wang *et al.* [28]	0.06	2.88	42.48	14.62
CDC [2]	0.20	**25.70**	**43.00**	**22.90**
R-C3D (Baseline) [6]	1.69	11.47	26.45	13.33
R-C3D + In-Out	**2.70**	14.90	27.18	15.62
R-C3D + Boundary	**3.30**	16.31	28.10	16.78

4. CONCLUSION

In this paper, we propose a novel Boundary Likelihood Pinpointing (BLP) network for accurate temporal action localization. Specifically, instead of using boundary regression, we propose a substitution paradigm called boundary pinpointing. The localization process starts by assigning conditional probabilities to each equally divided unit of a search interval. These probabilities provide a measurement of confidence for each unit being within an action instance or being at the two boundaries. We can exploit these probabilities to accurately pinpoint the temporal boundaries under a simple probabilistic framework. Extensive experiments demonstrate that effectiveness of BLP localization model. Integrating our BLP model with existing action classifier into detection pipeline, competitive detection performance is achieved and we get 34.5% (tIoU = 0.7) and $15.5\times$ (tIoU = 0.95) relative gain over the mAP of state-of-the-art detectors on THUMOS'14 and ActivityNet respectively.

5. REFERENCES

[1] Zheng Shou, Dongang Wang, and Shih-Fu 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.

[2] Zheng Shou, Jonathan Chan, Alireza Zareian, Kazuyuki Miyazawa, and Shih-Fu Chang, "Cdc: convolutional-de-convolutional networks for precise temporal action localization in untrimmed videos," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2017, pp. 1417–1426.

[3] Yue Zhao, Yuanjun Xiong, Limin Wang, Zhirong Wu, Xiaoou Tang, and Dahua Lin, "Temporal action detection with structured segment networks," in *The IEEE International Conference on Computer Vision (ICCV)*, 2017, vol. 8.

[4] Jiyang Gao, Zhenheng Yang, and Ram Nevatia, "Cascaded boundary regression for temporal action detection," in *Proceedings of the British Machine Vision Conference (BMVC)*, 2017.

[5] Xiyang Dai, Bharat Singh, Guyue Zhang, Larry S Davis, and Yan Qiu Chen, "Temporal context network for activity localization in videos," in *2017 IEEE International Conference on Computer Vision (ICCV)*. IEEE, 2017, pp. 5727–5736.

[6] Huijuan Xu, Abir Das, and Kate Saenko, "R-c3d: Region convolutional 3d network for temporal activity detection," in *The IEEE International Conference on Computer Vision (ICCV)*, 2017, vol. 6, p. 8.

[7] Tianwei Lin, Xu Zhao, and Zheng Shou, "Single shot temporal action detection," in *Proceedings of the 2017 ACM on Multimedia Conference*. ACM, 2017, pp. 988–996.

[8] Shyamal Buch, Victor Escorcia, Bernard Ghanem, Li Fei-Fei, and Juan Carlos Niebles, "End-to-end, single-stream temporal action detection in untrimmed videos," in *Proceedings of the British Machine Vision Conference (BMVC)*, 2017.

[9] Ke Yang, Peng Qiao, Dongsheng Li, Shaohe Lv, and Yong Dou, "Exploring Temporal Preservation Networks for Precise Temporal Action Localization," *arXiv.org*, Aug. 2017.

[10] F Caba Heilbron, Wayner Barrios, Victor Escorcia, and Bernard Ghanem, "Scc: Semantic context cascade for efficient action detection," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, vol. 2.

[11] Colin Lea, Michael D Flynn, Rene Vidal, Austin Reiter, and Gregory D Hager, "Temporal Convolutional Networks for Action Segmentation and Detection," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2017, pp. 1003–1012, IEEE.

[12] Humam Alwassel, Fabian Caba Heilbron, Victor Escorcia, and Bernard Ghanem, "Diagnosing error in temporal action detectors," in *The European Conference on Computer Vision (ECCV)*, September.

[13] Tianwei Lin, Xu Zhao, and Zheng Shou, "Temporal convolution based action proposal: Submission to activitynet 2017," *arXiv preprint arXiv:1707.06750*, 2017.

[14] Jiyang Gao, Zhenheng Yang, Kan Chen, Chen Sun, and Ram Nevatia, "Turn tap: Temporal unit regression network for temporal action proposals," in *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[15] Yu-Wei Chao, Sudheendra Vijayanarasimhan, Bryan Seybold, David A Ross, Jia Deng, and Rahul Sukthankar, "Rethinking the faster r-cnn architecture for temporal action localization," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 1130–1139.

[16] Y.-G. Jiang, J. Liu, A. Roshan Zamir, G. Toderici, I. Laptev, M. Shah, and R. Sukthankar, "THUMOS challenge: Action recognition with a large number of classes," http://crcv.ucf.edu/THUMOS14/, 2014.

[17] Bernard Ghanem Fabian Caba Heilbron, Victor Escorcia and Juan Carlos Niebles, "Activitynet: A large-scale video benchmark for human activity understanding," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 961–970.

[18] Karen Simonyan and Andrew Zisserman, "Two-stream convolutional networks for action recognition in videos," in *Advances in neural information processing systems*, 2014, pp. 568–576.

[19] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri, "Learning spatiotemporal features with 3d convolutional networks," in *Computer Vision (ICCV), 2015 IEEE International Conference on*. IEEE, 2015, pp. 4489–4497.

[20] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool, "Temporal segment networks: Towards good practices for deep action recognition," in *European Conference on Computer Vision*. Springer, 2016, pp. 20–36.

[21] Joao Carreira and Andrew Zisserman, "Quo vadis, action recognition? a new model and the kinetics dataset," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2017, pp. 4724–4733.

[22] Chao-Yuan Wu, Manzil Zaheer, Hexiang Hu, R Manmatha, Alexander J Smola, and Philipp Krähenbühl, "Compressed video action recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 6026–6035.

[23] Spyros Gidaris and Nikos Komodakis, "Locnet: Improving localization accuracy for object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 789–798.

[24] Tianwei Lin, Xu Zhao, Haisheng Su, Chongjing Wang, and Ming Yang, "Bsn: Boundary sensitive network for temporal action proposal generation," in *European Conference on Computer Vision*, 2018.

[25] Victor Escorcia, Fabian Caba Heilbron, Juan Carlos Niebles, and Bernard Ghanem, "Daps: Deep action proposals for action understanding," in *European Conference on Computer Vision*. Springer, 2016, pp. 768–784.

[26] Navaneeth Bodla, Bharat Singh, Rama Chellappa, and Larry S. Davis, "Soft-nms – improving object detection with one line of code," 2017.

[27] Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell, "Caffe: Convolutional architecture for fast feature embedding," in *Proceedings of the 22nd ACM international conference on Multimedia*. ACM, 2014, pp. 675–678.

[28] R. Wang and D. Tao, "Uts at activitynet 2016," *AcitivityNet Large Scale Activity Recognition Challenge*, 2016.