This paper is a submitted and accepted version for publication in:
IEEE Signal Processing Letters.

Murtaza, F., Yousaf, M.H., Velastin, S. A., Qian, Y. (En prensa). End-to-End Temporal Action Detection using Bag of Discriminant Snippets (BoDS). *IEEE Signal Processing Letters.*

DOI:

©2018 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works, for resale or redistribution to servers or lists, or reuse of any copyrighted component of this work in other works.

End-to-End Temporal Action Detection using Bag of Discriminant Snippets (BoDS)

Fiza Murtaza, Muhammad Haroon Yousaf, *Member, IEEE,* Sergio A. Velastin, *Senior Member, IEEE,* and Yu Qian

***Abstract*—Detecting human actions in long untrimmed videos is a challenging problem. Existing temporal action detection methods have difficulties in finding the precise starting and ending time of the actions in untrimmed videos. In this letter, we propose a temporal action detection framework based on a Bag of Discriminant Snippets (BoDS) that can detect multiple actions in an end-to-end manner. BoDS is based on the observation that multiple actions and the background classes have similar snippets, which cause incorrect classification of action regions and imprecise boundaries. We solve this issue by finding the key-snippets from the training data of each class and compute their discriminative power which is used in BoDS encoding. During testing of an untrimmed video, we find the BoDS representation for multiple candidate proposals and find their class label based on a majority voting scheme. We test BoDS on the Thumos14 and ActivityNet datasets and obtain state-of-the-art results. For the sports subset of ActivityNet dataset, we obtain a mean Average Precision (mAP) value of 29% at 0.7 temporal intersection over union (tIoU) threshold. For the Thumos14 dataset, we obtain a significant gain in terms of mAP i.e., improving from 20.8% to 31.6% at tIoU=0.7.**

***Index Terms*—Temporal Action Detection, 3D-Convolutional network (C3D), untrimmed videos, Thumos14, ActivityNet, temporal action proposals.**

I. INTRODUCTION

WITH the ubiquity of camera devices, a large volume of untrimmed video data is being recorded which contains multiple human action plus background actions. Given an untrimmed video, the task of the Temporal Action Detection (TAD) is to answer, “when does an action of interest start and end?”. All other background scenes and activities present in the untrimmed video, other than actions of interest, are referred as the background class. TAD has emerged as an important topic in the research community due to its numerous applications in video analysis, surveillance and many others [1–4]. In contrast to conventional human action recognition [5–9] which only recognizes the action category in manually trimmed videos, TAD is also expected to output the starting and ending time of the actions of interest present in untrimmed videos.

Over the last few years, convolutional neural networks (CNNs) have led to improved accuracy of action recognition [5–8]. However, TAD methods [2, 10–14] still need improvement. In [10], Pyramid of Score Distribution Feature (PSDF) based TAD approach is proposed. PSDF is computationally complex as it captures the motion information at multiple resolutions. A Structured Segment Network (SSN) is proposed in [15] which utilizes a structured temporal pyramid for modeling human activities in untrimmed videos. Temporal Actionness Grouping (TAG) is proposed in [16] to generate multiple action proposals. TAG is dependent upon two thresholds to filter the action from the background and incomplete regions which makes it less practical in real time scenarios. Convolutional De-Convolutional Networks (CDC) [17] perform dense prediction at each frame to find temporal boundaries. However, CDC is dependent on other proposal generation methods, e.g. Segment-CNN (SCNN)[4], to produce an initial set of proposals. Temporal Unit Regression Network (TURN) [18] and Cascaded Boundary Regression (CBR) [19] perform the temporal boundary regression for boundary refinement but they struggle to produce good results at high tIoU thresholds.

These existing TAD methods have some shortcomings. First, they tend not to exploit the discriminative power of the snippets, therefore, they fail to discriminate one action from other actions and from the background class. This subsequently produces invalid detections as there are many small clips of δ frames, known as snippets, which are similar in different action classes as well as in the background class. We claim that overcoming this problem requires incorporating the discriminative power of the snippets during the encoding process. Second, most existing methods [4, 15–19] require a two-stage paradigm, i.e. proposal and classification. This requires multiple passes through testing data for these two stages, therefore, it is difficult to use these methods in an end-to-end manner for TAD.

In this work, we propose an effective TAD method to address such shortcomings. Specifically, we adopt an end-to-end paradigm which can directly detect multiple actions in untrimmed videos while rejecting the non-action sections i.e. the background sections, in a single pass. We propose a bag of discriminant snippets (BoDS) encoding method which incorporates the discriminating power of the key-snippets in terms of weights. This encoding scheme is integrated with a 3D-Convolutional network (C3D) [6] representation, however, it can be used with any CNN as well as with handcrafted

Manuscript submitted September 14, 2018, revised December 27, 2018. This work was supported by the ASR&TD, University of Engineering and Technology (UET) Taxila, Pakistan. The work of S. A. Velastin was supported by the Universidad Carlos III de Madrid, the European Unions Seventh Framework Program for research, technological development, and demonstration under Grant 600371, el Ministerio de Economia y Competitividad (COFUND2013-51509), and Banco Santander.

F. Murtaza and M. H. Yousaf are with the Department of Computer Engineering, UET Taxila, Pakistan. (e-mail:fizamurtaza@yahoo.com; haroon.yousaf@uettaxila.edu.pk).

S. A. Velastin is with Cortexica Vision Systems Ltd., London, SE1 9LQ, UK (e-mail: sergio.velastin@ieee.org) and a visiting professor at Queen Mary University of London, UK and Universidad Carlos III de Madrid, Spain.

Yu Qian is with Cortexica Vision Systems Ltd., London, SE1 9LQ, UK.(e-mail: yu.qian@cortexica.com).

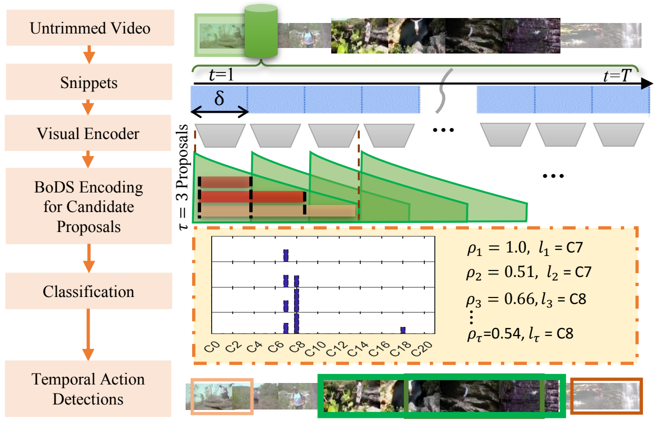

Untrimmed Video
Snippets
Visual Encoder
BoDS Encoding for Candidate Proposals
Classification
Temporal Action Detections
t=1
t=T
δ
τ = 3 Proposals
ρ_1 = 1.0, l_1 = C7
ρ_2 = 0.51, l_2 = C7
ρ_3 = 0.66, l_3 = C8
ρ_τ=0.54, l_τ = C8
C0 C2 C4 C6 C8 C10 C12 C14 C16 C18 C20

Fig. 1: Block Diagram of the proposed BoDS-TAD framework.

features.

The proposed BoDS temporal detection (BoDS-TAD) framework has the following contributions: 1) It proposes an effective encoding model, i.e. BoDS, that has the capability of discriminating different actions from the background actions. 2) It does not require a separate stage for proposal generation to eliminate background sections. It can unroll over long untrimmed testing videos in an end to end manner to encode and classify the proposals of multiple durations into either targeted actions or background class. 3) Our proposed method achieves state-of-the-art TAD performance on standard datasets.

II. Proposed BoDS-TAD

In this section, we provide the details of the proposed BoDS-TAD method, of which Fig. 1 shows an overview. We represent a video as $V = \{f_n\}_{n=1}^{N}$ where f_n is the n-th frame and N is the total number of frames. Each video V is associated with a set of ground truth annotations $\mathcal{G} = \{(g_m, g_m^{'}, l_m)\}_{m=1}^{G}$, where G is the total number of action instances in V and g_m, $g_m^{'}$ and l_m represent, respectively, starting time, ending time and the action category of the occurrence m. $l_m \in \{1, \cdots, C\}$, where C represents the number of targeted action classes. All remaining regions are treated as background with class label $C + 1$. In this work, the goal is to simultaneously detect the temporal boundaries of the action instances and their class labels in the untrimmed videos.

A. Visual Encoder

For feature extraction, we choose C3D [6] as it has been effectively used by other TAD methods [4, 11, 12] to capture the visual as well as the motion information over non-overlapping snippets of δ frames. We divide each video into $T = N \,/\, \delta$ snippets. Each snippet, at time step t, is represented using C3D based feature representation as $\{s_t\}_{t=1}^{T}$. As a standard practice, δ is set to 16 frames for C3D features [4, 6, 11, 12]. We use the publicly available C3D model that is pretrained on Sports1M dataset [6], the output of the $fc6$ is used as snippet-level features.

B. Learning to Extract Key-snippets

To extract the key-snippets, we compile all snippets-level features from videos belonging to the class i into a snippet matrix as $X_i \in \mathbb{R}^{D \times n_i}$ where n_i represents the total number of snippets in the training set of class i and D represents the dimension of the snippet-level feature vector. We extract K key-snippets from each class by performing the class-specific clustering over the snippet matrix X_i of each class using K-means clustering with Euclidean distance. We finally obtain a total of $K \times (C + 1)$ key-snippets $\{S_j\}_{j=1}^{K \times (C+1)}$.

In the next step, we compute the relative importance of each key-snippet by obtaining their weights w_j in accordance to their ability to differentiate between the different action and background classes. From all of the $C+1$ classes, each snippet s_l is assigned to the nearest key-snippet S_j, such that $||s_l - S_j||$ is minimum. For each key-snippet S_j, we record the correct and false assignments of snippets in terms of within-class q_j and out-of-class q_j' assignments respectively. Within-class assignment q_j represents the number of times S_j is matched with the snippets of its own class whereas the out-of-class assignment q_j' represent the number of times it matched with the snippets of other classes. These assignments will be used to find the discriminative importance of each key-snippet S_j described by weights w_j, as given below:

$$w_j = \frac{q_j}{q_j + q_j'} \qquad \forall\; j \in [1 : K \times (C + 1)]. \tag{1}$$

From Eq. 1 it follows that if S_j is not assigned to any of the snippet of its own class then $w_j = 0$ or if it is only assigned to the snippets of its own class then $w_j = 1$ else $0 < w_j < 1$. Visualization of some key-snippets is given in the supplementary material.

C. BoDS Encoding for Untrimmed Videos

We integrate the discriminative power of key-snippets in the encoding process to find the initial BoDS representation for the untrimmed videos. Each snippet-level feature s_t at time t is compared with $K \times (C + 1)$ key-snippets and its nearest key-snippet is obtained as:

$$j = \underset{1 \le r \le K \times (C+1)}{\arg\min} \; ||s_t - S_r||. \tag{2}$$

Then each snippet at time step t, votes for the nearest key-snippet using hard-voting scheme [20] as given by:

$$v_t = [0, \cdots, w_j, \cdots, 0] \tag{3}$$

where v_t is the $K \times (C + 1)$ dimensional vector initially having all zeros except at the j-th index where we vote the weight w_j of its nearest matching key-snippet S_j. As the key-snippets which are common to more than one class will have less weights, they will contribute less in the final decision.

D. Candidate Temporal Proposals

Next, we aggregate multiple snippets to find the candidate temporal proposals of variable durations which are likely to contain action regions. At each time step t, we produce a left aligned proposal set P_t, having τ proposals of duration

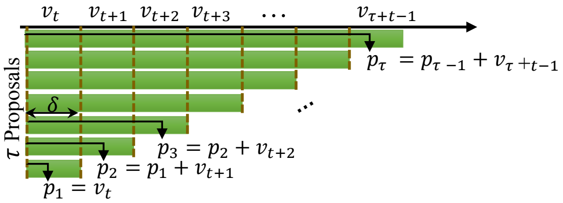


Fig. 2: Generation of BoDS representation for τ candidate proposals at time step t.

$1\delta, 2\delta, \cdots, \delta\tau$ frames respectively as shown in Fig. 2. Each candidate proposal set is represented as $P_t = \{p_h, o_h, o'_h\}_{h=1}^{\tau}$, where $o_h = t$ and $o'_h = h + t - 1$ are respectively the starting and ending time of the h-th candidate proposal. The aggregated feature vector p_h, considered as the final BoDS representation for the h-th proposal, is calculated using sum pooling as given by:

$$p_h = p_{h-1} + v_{t+h-1}, \qquad \forall\ h \in [1:\tau], p_0 = 0 \tag{4}$$

where v is calculated using Eq. 3.

E. Classifying Candidate Proposals

Once the $K \times (C+1)$ dimensional vector p_i is calculated for each proposal i, the next task is to classify it into one of the $C+1$ classes. For each proposal i, we accumulate the weighted votes for each class c, M_i^c, as given by:

$$M_i^c = \sum_{r=(c-1)\times K+1}^{c\times K} p_i(r), \qquad \forall c \in [1:C+1]. \tag{5}$$

Eq. 5 indicates that when $c = 1$, it accumulates the votes for first K key-snippets as $1:K$ elements of p_i belong to the first class. In this way, we find the total weighted votes for all $C+1$ classes. Finally, a majority voting scheme is used to classify the given proposal i into one of the $C+1$ classes as:

$$l_i = \underset{1\leq c\leq C+1}{\arg\max}\ M_i^c \tag{6}$$

where l_i is the label of the class having maximum votes.

We then assign the probability to each the proposal i based on its maximum class probability as calculated by:

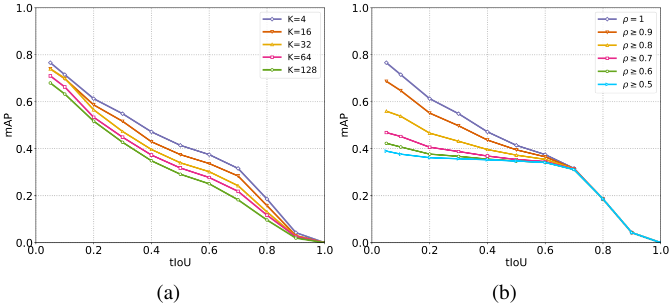
$$\rho_i = \frac{\max_{1\leq c\leq C+1} M_i^c}{\sum_{c=1}^{C+1} M_i^c}. \tag{7}$$

This indicates that those proposals having most of the snippets assigned to a single class will have higher probability values. Later in Section III-C, we see the effect of selecting proposals based on their corresponding probabilities ρ_i.

III. Experimental Results

A. Datasets and Evaluation Measure

Two untrimmed video datasets are used for the evaluation of the proposed method: Thumos14 [21] and ActivityNet [22]. Thumos14 contains untrimmed videos from 20 sports actions compiled from YouTube. As a standard practice [21], we use 200 untrimmed validation videos for training and 213 untrimmed test videos for testing purpose. For ActivityNet, we use training and validation splits, respectively, for training and testing purpose. To compare with previous work [3, 18, 22], we perform the experiments on the 'sports' subset of ActivityNet v1.3 containg 21 sports actions. In the supplementary material, we also report results on other subsets of ActivityNet v1.3 containing 'household' 'personal care' 'eating and drinking' 'socializing and leisure' and 'sports and exercises'. For performance evaluation, we use mean average precision (mAP) calculated at different temporal intersection over union (tIoU) thresholds using the publicly available evaluation toolkit [21].



(a) (b)

Fig. 3: Evaluation of the effect of the BoDS parameters i.e. K (a) and ρ (b) on a testing set of THUMOS-14 dataset.

B. Implementation Details

For setting the value of τ in Eq. 4, we utilize the maximum length of the actions from the training data. For Thumos14 and ActivityNet, the maximum action duration is about 1024 and 1600 frames respectively. Therefore we set $\tau = 64$ (1024/16) and $\tau = 100$ (1600/16) snippets for both datasets respectively. This resulted in proposals of all possible durations which may overlap in time. As a common practice [2, 4, 12], we perform non-maximal suppression (NMS) on the proposals selected using Eq. 7 (as discussed in Section III-C), with 0.7 overlap threshold, to remove the highly overlapping proposals.

C. Evaluating BoDS parameters

We evaluate the impact of the number of key-snippets, i.e. K, on the BoDS performance, assessed with $K \in \{4, 16, 32, 64, 128\}$. The results in Fig. 3(a) show that mAP obtained for different values of K have small differences. Using only few key-snippets, e.g. $K = 4$, our method correctly classifies the candidate proposals for tIoU thresholds between 0.05 to 1. Similarly, using a large number of key-snippets, $K = 128$, may lead to the wrong assignment of snippets as the distance between cluster centroids will be smaller. Based on this observation, we choose $K = 4$ key-snippets for the rest of the experiments for both datasets.

We also evaluate the performance of the proposed method by retrieving different proposals based upon their probability ρ as shown in Fig. 3(b). Using $\rho = 1$ removes the proposals of long duration which overlap with the temporal span of more than one action or the background instances. Therefore it resulted in higher mAP value than all other probability

TABLE I: Comparison of TAD performance in terms of mAP(%) @ different tIoU thresholds.

tIoU threshold	0.3	0.4	0.5	0.6	0.7
Thumos14					
FG [3] (2016)	36.0	-	17.1	-	-
PSDF [10] (2016)	33.6	26.2	18.8	-	-
SCNN [4] (2016)	36.3	28.7	19.0	10.3	5.3
CDC [17] (2017)	40.1	29.4	23.3	13.1	7.9
SSAD [23] (2017)	43.0	35.0	24.6	-	-
TURN [18] (2017)	44.1	34.9	25.6	-	-
TPN [24] (2017)	44.1	37.1	28.2	20.6	12.7
TAG [16] (2017)	48.7	39.8	28.2	-	-
R-C3D [25] (2017)	44.9	35.6	28.9	-	-
SS-TAD [13] (2017)	45.7	-	29.2	-	9.6
SSN [15] (2017)	51.9	41.0	29.8	-	-
CBR [19] (2017)	50.1	41.3	31.0	19.1	9.9
ETP [26] (2018)	48.2	42.4	34.2	23.4	13.9
TAL-Net [27] (2018)	53.2	**48.5**	**42.8**	33.8	20.8
BoDS [Ours]	**54.9**	47.2	41.5	**37.5**	**31.6**
ActivityNet (Sports subset)					
[22] (2015)	-	-	33.2	-	-
FG [3] (2016)	-	-	36.7	-	-
TURN [18] (2017)	-	-	37.1	-	-
BoDS [Ours]	51.1	45.0	**38.1**	34.2	29.0

thresholds defined in Fig. 3(b). For all probability thresholds, we see the same behaviour at tIoU threshold between 0.7 and 1. This indicates that all of the highly overlapping proposals, obtained by our method, have probability value equal to 1. Moreover, if we do not have an idea of maximum action duration present in some dataset, we can set τ equal to any large value. This leads to many long duration proposals which will be automatically removed by setting $\rho = 1$.

D. Comparisons

In Table I, we provide the comparison of the proposed method with state-of-the-art approaches. For Thumos14, the method outperforms state-of-the-art approaches including action detection from Frame Glimpses (FG) [3], PSDF [10], SCNN [4], Single-Stream TAD (SS-TAD) [13], SSN [15], the actionness based approach TAG [16], CDC [17], Single Shot Action Detector (SSAD) [23], TURN [18], CBR [19], Temporal Preservation Networks (TPN) [24], Regional C3D (R-C3D) [25], Evolving Temporal Proposals (ETP) [26] and Temporal Action Localization Network (TAL-Net) [27].

Like our method, SS-TAD [13] also produces proposals (right-aligned) of variable duration, however, this requires sliding windows for producing dense training data. Instead, we use the snippet-level data from the training data only for the extraction of the key-snippet which makes the method computationally efficient. Similarly, TURN [18] produces proposals of varying durations by applying sliding windows at multiple temporal scales, which makes them computationally expensive. The top-performing methods i.e. SSN [15], CBR [19], ETP [26] and TAL-Net [27] are built upon multiple networks for proposal generation, refinement, and final detection tasks. However, our method is based upon a single end-to-end framework which does not require extra network(s) for classifying regions into action or background which makes it computationally fast. BoDS achieves significant performance gain under high tIoU threshold of 0.7, where it outperforms TAL-Net [27] by 10.8% mAP.

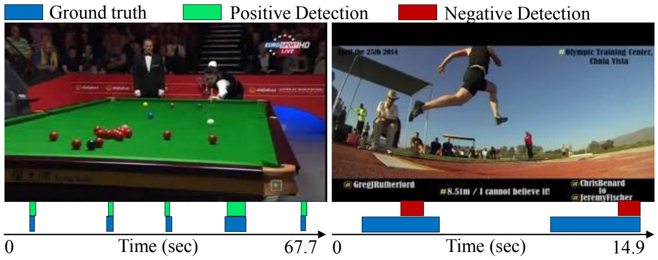


Fig. 4: Qualitative results on two test videos from Thumos14.

For ActivityNet dataset, we compare our results with FG [3], TURN [18] and [22]. Table 1 reports the mAP at different tIoU thresholds whereas other methods [3, 18, 22] only reported the mAP results at tIoU threshold of 0.5. Results show that BoDS resulted in improved detection performance at tIoU threshold of 0.5 for the sports subset of the ActivityNet dataset. For ActivityNet, we adopted the parameter ρ and K from the Thumos14 dataset, this reveals that our proposed method may be generalized to other action datasets as well. BODS-TAD operates at 1279 frames per second (FPS) with C3D features on a single Titan X Pascal GPU. Whereas TURN [18] and R-C3D [25] run at 880 and 1030 FPS respectively.

E. Qualitative Results

Fig. 4 (left) shows the positive detections retrieved by the proposed action detection approach for a video containing multiple Billiards action sequences. The detected region is considered as true positive if the temporal tIoU with ground truth region is greater than or equal to 0.5 and a correct action label is assigned to it. We observe that our method detects all instances of Billiards action, having tIoU greater than 0.75 with the ground truth locations. In Fig. 4 (right) we also provide an example where the BoDS fails to detect correct action labels for a video containing LongJump action sequences. BoDS fails to detect the precise starting and ending time because the video has multiple viewpoints in a single action instance.

IV. Conclusion

In this work, we have proposed a new end-to-end framework, BoDS-TAD, which utilizes the discriminant power of the snippets for detection of true actions. As compared to other methods, which are built upon a "proposal and classification" paradigm, our method does not require a classification stage for proposal extraction which makes it computationally efficient. BoDS-TAD runs at 1279 FPS making it possible for large-scale untrimmed videos. Through experiments, we have shown that this model achieves state-of-the-art performance on the action detection task. It produces proposals which, although are part of actual actions they may be incomplete. In future we will handle this issue by proposing a scheme for rejecting incomplete proposals. For future work, we plan to use this framework as a base module for other video understanding tasks, such as sports video analysis and video summarization.

References

[1] D. Oneata, J. Verbeek, and C. Schmid, "Action and event recognition with fisher vectors on a compact feature set," in *Proceedings of the IEEE international conference on computer vision*, 2013, pp. 1817–1824.

[2] P. Mettes, J. C. van Gemert, S. Cappallo, T. Mensink, and C. G. Snoek, "Bag-of-fragments: Selecting and encoding video fragments for event detection and recounting," in *Proceedings of the 5th ACM on International Conference on Multimedia Retrieval*. ACM, 2015, pp. 427–434.

[3] S. Yeung, O. Russakovsky, G. Mori, and 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.

[4] Z. Shou, D. Wang, and 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.

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

[6] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3d convolutional networks," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 4489–4497.

[7] L. Wang, Y. Qiao, and 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.

[8] L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang, and L. Van Gool, "Temporal segment networks: Towards good practices for deep action recognition," in *European Conference on Computer Vision*. Springer, 2016, pp. 20–36.

[9] S. Nazir, M. H. Yousaf, J.-C. Nebel, and S. A. Velastin, "A bag of expression framework for improved human action recognition," *Pattern Recognition Letters*, vol. 103, pp. 39–45, 2018.

[10] J. Yuan, B. Ni, X. Yang, and 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.

[11] S. Buch, V. Escorcia, C. Shen, B. Ghanem, and J. C. Niebles, "Sst: Single-stream temporal action proposals," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2017, pp. 6373–6382.

[12] V. Escorcia, F. C. Heilbron, J. C. Niebles, and B. Ghanem, "Daps: Deep action proposals for action understanding," in *European Conference on Computer Vision*. Springer, 2016, pp. 768–784.

[13] S. Buch, V. Escorcia, B. Ghanem, L. Fei-Fei, and J. Niebles, "End-to-end, single-stream temporal action detection in untrimmed videos," in *Proceedings of the British Machine Vision Conference (BMVC)*, 2017.

[14] F. Murtaza, M. H. Yousaf, and S. A. Velastin, "Pmhi: Proposals from motion history images for temporal segmentation of long uncut videos," *IEEE Signal Processing Letters*, vol. 25, no. 2, pp. 179–183, 2018.

[15] Y. Zhao, Y. Xiong, L. Wang, Z. Wu, X. Tang, and D. Lin, "Temporal action detection with structured segment networks," *ICCV, Oct*, vol. 2, 2017.

[16] Y. Xiong, Y. Zhao, L. Wang, D. Lin, and X. Tang, "A pursuit of temporal accuracy in general activity detection," *arXiv preprint arXiv:1703.02716*, 2017.

[17] Z. Shou, J. Chan, A. Zareian, K. Miyazawa, and S.-F. Chang, "Cdc: Convolutional-de-convolutional networks for precise temporal action localization in untrimmed videos," in *Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on*. IEEE, 2017, pp. 1417–1426.

[18] J. Gao, Z. Yang, C. Sun, K. Chen, and R. Nevatia, "Turn tap: Temporal unit regression network for temporal action proposals," *arXiv preprint arXiv:1703.06189*, 2017.

[19] J. Gao, Z. Yang, and R. Nevatia, "Cascaded boundary regression for temporal action detection," *arXiv preprint arXiv:1705.01180*, 2017.

[20] J. Sivic and A. Zisserman, "Video google: A text retrieval approach to object matching in videos," in *null*. IEEE, 2003, p. 1470.

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

[22] F. Caba Heilbron, V. Escorcia, B. Ghanem, and J. 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.

[23] T. Lin, X. Zhao, and Z. Shou, "Single shot temporal action detection," in *Proceedings of the 2017 ACM on Multimedia Conference*. ACM, 2017, pp. 988–996.

[24] K. Yang, P. Qiao, D. Li, S. Lv, and Y. Dou, "Exploring temporal preservation networks for precise temporal action localization," *arXiv preprint arXiv:1708.03280*, 2017.

[25] H. Xu, A. Das, and K. Saenko, "R-c3d: region convolutional 3d network for temporal activity detection," in *IEEE Int. Conf. on Computer Vision (ICCV)*, 2017, pp. 5794–5803.

[26] H. Qiu, Y. Zheng, H. Ye, Y. Lu, F. Wang, and L. He, "Precise temporal action localization by evolving temporal proposals," in *Proceedings of the 2018 ACM on International Conference on Multimedia Retrieval*. ACM, 2018, pp. 388–396.

[27] Y.-W. Chao, S. Vijayanarasimhan, B. Seybold, D. A. Ross, J. Deng, and R. 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.