JOINT SPATIO-TEMPORAL ACTION LOCALIZATION IN UNTRIMMED VIDEOS WITH PER-FRAME SEGMENTATION

Xuhuan Duan[1], *Le Wang*[1]*, *Changbo Zhai*[1], *Qilin Zhang*[2], *Zhenxing Niu*[3], *Nanning Zheng*[1], *and Gang Hua*[4]

[1]Xi'an Jiaotong University [2]HERE Technologies [3]Alibaba Group [4]Microsoft Research

ABSTRACT

Inspired by the recent spatio-temporal action localization efforts with tubelets (sequences of bounding boxes), we present a new spatio-temporal action detector Segment-tube, which consists of sequences of per-frame segmentation masks. The proposed Segment-tube detector can temporally pinpoint the starting/ending frame of each action class in the presence of preceding/subsequent interference actions in untrimmed videos. Simultaneously, the Segment-tube detector produces per-frame segmentation masks instead of bounding boxes, offering superior spatial accuracy to tubelets. This is achieved by alternating iterative optimization between temporal action localization and spatial action segmentation. Experimental results on multiple datasets validate the efficacy of the proposed detector.

Index Terms— Action Localization, Action Segmentation, 3D ConvNets, LSTM

1. INTRODUCTION

Joint spatio-temporal action localization has attracted significant attention in recent years [1,2,3,4,5,6,7,8,9,10,11,12,13,14], whose objectives include action classification (determining whether a specific action is present), temporal localization (pinpointing the starting/ending frame of the specific action) and spatio-temporal localization (typically bounding box regression on 2D frames, *e.g.*, [15,16]). Such efforts include local feature based methods [10], convolution neural networks (ConvNets or CNNs) based methods [3,7], and 3D ConvNets based methods [4,13]. Recently, long short-term memory (LSTM) based recurrent neural networks (RNNs) are added on top of CNNs for action classification [5,17] and action localization [6].

Despite the successes of the prior methods, there are multiple limiting factors impeding practical applications. For example, [5,7,8] conduct action recognition only on trimmed videos, where each video contain only one action without interferences from other potentially confusing actions; [3,6,9,10,11,12,13,14] emphasize only on temporal action localization with untrimmed video; while [15,16] implement spatio-temporal action localization in trimmed videos with tubelet-style (sequences of bounding boxes) detectors.

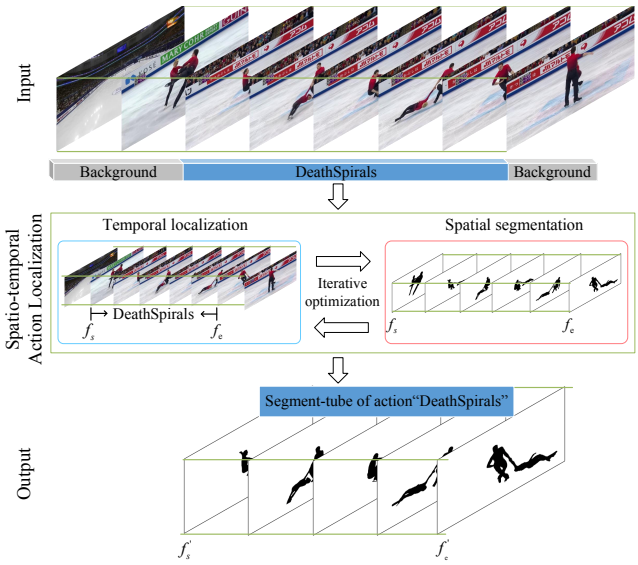

Fig. 1: Flowchart of the proposed Segment-tube detector. An input untrimmed video (*e.g.*, an entire pair figure skating video) typically contains irrelevant preceding and subsequent actions (marked with gray chunks) beyond the relevant frames of a specific action class (*e.g.*, the DeathSpirals, denoted with the blue chunk). The Segment-tube detector alternates the optimization of temporal localization and spatial segmentation iteratively, and outputs a sequence of per-frame segmentation masks with precise starting/ending frames.

With applications in untrimmed videos with improved spatial accuracy in mind, we propose the Segment-tube spatio-temporal action localization detector, as summarized in Fig. 1. Initialized with saliency [18] based image segmentation on individual frames, our method performs temporal action localization with 3D ConvNets and LSTM. In an alternating and iterative manner, the Segment-tube detector refines spatial per-frame segmentation by focusing on frames identified by the temporal localization step. Upon practical convergence, the final spatio-temporal action localization results are obtained in the format of a sequence of segmentation masks (bottom row in Fig. 1).

We conduct extensive experiments on multiple datasets consisting of untrimmed videos, including temporal action lo-

*Corresponding Author, *lewang@xjtu.edu.cn*

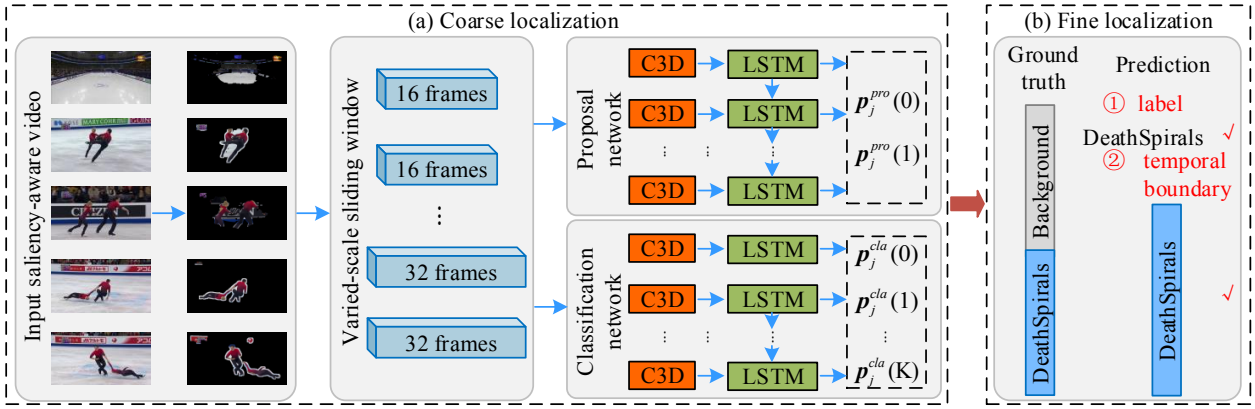

Fig. 2: Overview of the proposed coarse-to-fine temporal action localization.

calization on the THUMOS 2014 dataset [2], and joint spatio-temporal action localization on the ActSeg dataset, which is a newly proposed spatio-temporal action localization dataset with per-frame ground truth segmentation masks. The contributions of this paper are as follows. (1) The Segment-tube spatio-temporal action localization detector is proposed for untrimmed videos, which produces per-frame segmentation masks instead of sequences of bounding boxes. (2) The proposed Segment-tube detector achieves collaborative optimization of temporal localization and spatial segmentation with a new iterative alternation approach. (3) The new ActSeg dataset is proposed, which consists of untrimmed videos with temporal annotations and per-frame ground truth segmentation labels.

2. PROBLEM FORMULATION

Given a video $V = \{f_t\}_{t=1}^{T}$ of T frames, our objective is to determine whether a specific action $k \in \{1, \cdots, K\}$ appears in V, and if so, temporally pinpoint the starting frame $f_s(k)$ and ending frame $f_e(k)$ for action k. Simultaneously, a sequence of segmentation masks $\mathcal{B} = \{B_t\}_{t=f_s(k)}^{f_e(k)}$ within such frame range should be obtained, with B_t being a binary segmentation label for frame t. Practically, B_t consists of a series of superpixels $B_t = \{b_{t,i}\}_{i=1}^{N_t}$, with N_t being the total number of superpixels in frame f_t.

2.1. Temporal Action Localization

A coarse-to-fine action localization strategy is implemented to accurately find the temporal boundaries of the target action k from an untrimmed video, as illustrated in Fig. 2. Inspired by the recent success of ConvNets [19], this is achieved by a cascaded 3D ConvNets with LSTM. The 3D ConvNets consists of eight 3D convolution layers, five 3D pooling layers, and two fully connected layers. The fully-connected 7th layer activation feature is used to represent the video clip. To exploit the temporal correlations, we incorporate a two-layer LSTM [5] using the Peephole implementation (with 256 hidden states in each layer) with 3D ConvNets.

Coarse Action Localization. The coarse action localization determines the approximate temporal boundaries with a fixed step-size (*i.e.*, video clip length). We first generate a set of H saliency-aware video clips $\{h_j\}_{j=1}^{H}$ with variable-length (16 and 32 frames per clip, [20]) sliding window with 75% overlap ratio on the initial segmentation B_o of video V (by using saliency [18]), and proceed to train a cascaded 3D ConvNets with LSTM that couples a proposal network and a classification network. The proposal network is action class-agnostic, it determines whether any actions ($\forall k \in \{1, \cdots, K\}$) are present in clip h_j. The classification network determines whether a specific action k is present in clip h_j. We follow [13] to construct training data from these video clips.

Specifically, we train the proposal network (3D ConvNets with LSTM) to score each video clip h_j with a proposal score $\mathbf{p}_j^{pro} = \left[\mathbf{p}_j^{pro}(1), \mathbf{p}_j^{pro}(0)\right]^T \in \mathcal{R}^2$. Subsequently, a flag label l_j^{fla} is obtained for each clip h_j,

$$l_j^{fla} = \begin{cases} 1, & \text{if} \quad \mathbf{p}_j^{pro}(1) > \mathbf{p}_j^{pro}(0) \\ 0, & \text{otherwise} \end{cases}, \tag{1}$$

where $l_j^{fla} = 1$ denotes the video clip h_j contains an action ($\forall k \in \{1, \cdots, K\}$), and $l_j^{fla} = 0$ otherwise.

A classification network (also a 3D ConvNets with LSTM) is further trained to predict a $(K+1)$-dimensional[1] classification score $\mathbf{p}_j^{cla}$ for each clip that contains an action $\left\{h_j | l_j^{fla} = 1\right\}$, based on which a specific action label $l_j^{spe} \in \{k\}_{k=0}^{K}$ and score $v_j^{spe} \in [0, 1]$ for h_j are assigned,

$$l_j^{spe} = \arg\max_{k=0,\cdots,K} \mathbf{p}_j^{cla}(k), \; v_j^{spe} = \max_{k=0,\cdots,K} \mathbf{p}_j^{cla}(k). \tag{2}$$

Fine Action Localization. With the obtained per-clip specific action labels l_j^{spe}, the fine action localization step predicts the

[1]Class 0 denotes the additional "background" class. Although the proposal network prefilters most "background" clips, a background class is still needed for robustness in the classification network.

video class k' ($k' \in \{1, \cdots, K\}$) and subsequently obtains $f_s(k')$ and $f_e(k')$. We calculate the average of v_j^{spe} over all video clips for each action label l_j^{spe}. We take the label k' with the maximum average predicted score as the predicted action. Subsequently, the action score $\alpha_t(f_t|k')$ and the action label l_t for frame f_t specifically are determined by

$$\alpha_t(f_t|k') = \frac{\sum\limits_{j \in \{j|f_t \in h_j\}} v_j^{spe}}{|\{j|j \in \{f_t \in h_j\}|} , \quad (3)$$

$$l_t = \begin{cases} k', & \text{if} \quad \alpha_t > \gamma \\ 0, & \text{otherwise} \end{cases} , \quad (4)$$

where $|\{\cdot\}|$ denotes the cardinality of set $\{\cdot\}$. We empirically set $\gamma = 0.6$. $f_s(l_t)$ and $f_e(l_t)$ are assigned as the starting and ending frame of a series of consecutive frames sharing the same label l_t, respectively.

2.2. Spatial Action Segmentation

With the obtained temporal localization results, we further conduct spatial segmentation. This problem is cast into a spatio-temporal energy minimization framework,

$$E(\mathcal{B}) = \sum_{s_{t,i} \in V} D_i(b_{t,i}) + \sum_{s_{t,i}, s_{u,v} \in \mathcal{N}_i} S_{iv}(b_{t,i}, b_{u,v}), \quad (5)$$

where $s_{t,i}$ is the ith superpixel in frame f_t, which is computed by SLIC [21]. $D_i(b_{t,i})$ composes the data term, denoting the cost of labeling $s_{t,i}$ with the label $b_{t,i}$ from a color and location based appearance model. $S_{iv}(b_{t,i}, b_{u,v})$ composes the smoothness term, constraining the segmentation labels to be both spatially and temporally consistent from a color based consistency model. $\mathcal{N}_i$ is the spatial neighborhood of $s_{t,i}$ in frame f_t, and temporal neighborhood of $s_{t,i}$ in adjacent frames f_{t-1} and f_{t+1}.

Data Term. With a segmentation $\mathcal{B}$ for V, we estimate two color Gaussian Mixture Models (GMMs) and two location GMMs for the foregrounds and the backgrounds of V, respectively. The corresponding data term $D_i(b_{t,i})$ based on color and location GMMs in Eq. (5) is defined as

$$D_i(b_{t,i}) = -\log\left(\beta U_{b_{t,i}}^{col}(s_{t,i}) + (1-\beta) U_{b_{t,i}}^{loc}(s_{t,i})\right), \quad (6)$$

where β is a parameter controlling the contributions of color $U_{b_{t,i}}^{col}$ and location $U_{b_{t,i}}^{loc}$.

Smoothness Term. We exploit the standard contrast-dependent function [22,23,24] to encourage spatially and temporally adjacent superpixels with similar colors to be assigned with the same label. In Eq. (5), $S_{iv}(b_{t,i}, b_{u,v})$ is then defined as

$$S_{iv}(b_{t,i}, b_{u,v}) = \mathbb{1}_{[b_{t,i} \neq b_{u,v}]} \exp\left(-||\mathbf{c}_{t,i} - \mathbf{c}_{u,v}||_2^2\right), \quad (7)$$

where characteristic function $\mathbb{1}_{[b_{t,i} \neq b_{u,v}]} = 1$ when $b_{t,i} \neq b_{u,v}$, and 0 otherwise. $b_{t,i}$ and $b_{u,v}$ are the segmentation labels of $s_{t,i}$ and $s_{u,v}$, respectively. $\mathbf{c}$ is the color vector.

Optimization. With $D_i(b_{t,i})$ and $S_{iv}(b_{t,i}, b_{u,v})$, we leverage graph cut [25] to minimize the energy function in Eq. (5).

2.3. Iterative & Alternating Optimization

With an initial spatial segmentation B_o of video V using saliency [18, 26], the overall optimization alternates between the temporal action localization in Section 2.1 and spatial action segmentation in Section 2.2. Upon the practical convergence of this iterative process, the final results $\mathcal{B}$ are obtained.

3. EXPERIMENTS AND DISCUSSIONS

We conduct experiments on multiple datasets to evaluate the efficacy of the proposed Segment-tube detector, including 1) temporal action localization task on the THUMOS 2014 dataset [2], and 2) spatio-temporal action localization task on the newly proposed ActSeg dataset. The average precision (AP) and mean average precision (mAP) are employed to evaluate the temporal action localization performance. If an action is assigned the same category label with the ground truth and simultaneously its predicted temporal range overlaps the ground truth at a ratio above a predefined threshold (*e.g.*, 0.5), such temporal localization of an action is deemed correct. The intersection-over-union (IoU) score is utilized to evaluate the spatial action segmentation performance.

Temporal Localization on THUMOS 2014 dataset [2]. We first evaluate the temporal action localization performance on the THUMOS 2014 dataset [2], which is dedicated to localizing actions in long untrimmed videos involving 20 actions. The training set contains 2755 trimmed videos and 1010 untrimmed validation videos. For testing, we use 213 videos that contain relevant action instances. Five existing temporal action localization methods, *i.e.*, AMA [10], FTAP [11], ASLM [12], SCNN [13], and ASMS [3], are included as competing algorithms. AMA [10] combines iDT features and frame-level CNN features to train a SVM classifier. F-TAP [11] leverages high recall temporal action proposals. ASLM [12] uses a length and language model based on traditional motion features. SCNN [13] is an end-to-end segment-based 3D ConvNets framework, including proposal, classification and localization network. ASMS [3] localizes actions by searching for the structured maximal sum. The mAP comparisons are summarized in Table 1, which demonstrates that the proposed Segment-tube evidently outperforms competing algorithms with IoU being 0.3 and 0.5, and is marginally inferior to SCNN [13] with IoU being 0.4.

Table 1: mAP comparisons on the THUMOS 2014 dataset.

IoU threshold	0.3	0.4	0.5
AMA [10]	14.6	12.1	8.5
FTAP [11]	-	-	13.5
ASLM [12]	20.0	23.2	15.2
SCNN [13]	36.3	**28.7**	19.0
ASMS [3]	36.5	27.8	17.8
Segment-tube	**39.8**	27.2	**20.7**

ActSeg dataset. To fully evaluate the spatio-temporal action localization performance, the new ActSeg dataset is introduced. It contains 446 untrimmed videos and 110 trimmed videos of 9 categories in its training split, and 85 untrimmed videos of 9 categories in its testing split. Both temporal annotations and per-frame pixel-wise segmentation labels are included as the ground-truth in all videos.

Mixed Dataset. To maximize the number of videos in each category (see Fig. 3), a mixed dataset is constructed by combining videos of identical action categories from multiple datasets. The training split of the mixed dataset consists of all 446 untrimmed videos and 110 trimmed videos in the proposed ActSeg dataset, 791 trimmed videos from the UCF101 dataset [27], and 90 untrimmed videos from the THUMOS 2014 dataset [2]. The testing split of the mixed dataset consists of all the 85 untrimmed videos from the testing split of the proposed ActSeg dataset.

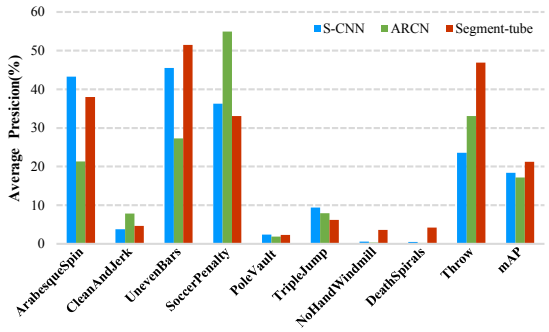

Fig. 3: Histogram of the per-category AP on the testing split of the mixed dataset. IoU threshold = 0.5.

Temporal Localization on Mixed Dataset. SCNN [13] and ARCN [9] are used as competing temporal action localization methods. All three methods are trained on the training split of the mixed dataset. Fig. 3 and Table 2 summarize the per-category AP and mAP, respectively. Our proposed Segment-tube method achieves the best mAP, and it outperforms competing methods in 4 out of 9 action categories.

Spatial Action Segmentation on ActSeg Dataset. The spatial action segmentation task is implemented entirely on the ActSeg dataset, with three competing video object segmentation methods, *i.e.*, VOS [28], FOS [29] and BVS [30]. IoU scores of the proposed Segment-tube method and the three competing methods are summarized in Table 3, with a few typical testing results visualized in Fig. 4. All predicted segmentation masks are visualized as polygons with red edges.

The results in Table 3 demonstrate that the Segment-tube method evidently outperforms VOS [28] and FOS [29], and it is subtly better than the label propagation based BVS method [30]. We speculate that severe occlusions (*e.g.*, in the PoleVault and TripleJump categories) might lead to some performance degradations in BVS [30].

We do not include performance comparisons on joint spatio-temporal localization, because existing methods either implement temporal action localization or spatial action segmentation, but never achieve both simultaneously.

Table 2: mAP comparisons on the mixed dataset.

IoU threshold	ARCN [9]	SCNN [13]	Segment-tube
0.5	17.2	18.4	**21.2**

Table 3: IoU scores on the ActSeg dataset.

Video	VOS [28]	FOS [29]	BVS [30]	Segment-tube
ArabequeSpin	53.9	**82.5**	64.0	80.2
CleanAndJerk	20.1	50.0	**85.9**	84.9
UnevenBars	12.0	40.3	**59.0**	53.2
SoccerPenalty	54.4	38.5	**59.8**	51.4
PoleVault	38.9	41.2	42.6	**46.9**
TripleJump	30.6	36.1	33.5	**55.7**
NoHandWindmill	77.1	73.3	81.8	**84.6**
DeathSpirals	1	66.7	**77.9**	63.1
Throw	33.8	2	**58.7**	53.1
Average	35.8	47.8	62.6	**63.7**

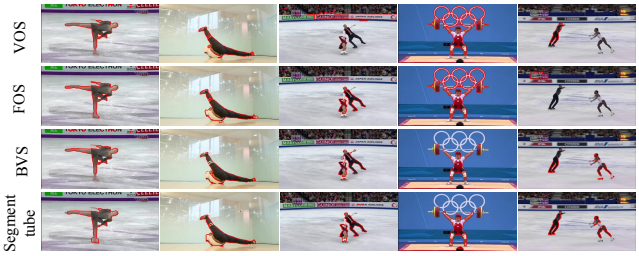

Fig. 4: Examples on the ActSeg dataset. Row 1 $\sim$ 4: VOS [28], FOS [29], BVS [30] and our proposed Segment-tube detector.

4. CONCLUSION

The Segment-tube spatio-temporal action localization detector is proposed, which jointly localize the temporal boundaries and spatial per-frame segmentation masks in untrimmed videos. With the proposed alternating iterative optimization scheme, temporal localization and spatial segmentation could be achieved simultaneously and evident performance gains are observed on multiple datasets.

5. ACKNOWLEDGMENT

This work was supported partly by National Key Research and Development Program of China Grant 2017YFA0700800, NSFC Grants 61629301, 61773312, 61503296, 91748208, and China Postdoctoral Science Foundation Grants 2017T100752 and 2015M572563.

6. REFERENCES

[1] Zhanning Gao, Gang Hua, Dongqing Zhang, Nebojsa Jojic, Le Wang, Jianru Xue, and Nanning Zheng, "Er3:

A unified framework for event retrieval, recognition and recounting," in *CVPR*, 2017.

[2] YG 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," 2014.

[3] Zehuan Yuan, Jonathan C Stroud, Tong Lu, and Jia Deng, "Temporal action localization by structured maximal sums," *arXiv:1704.04671*, 2017.

[4] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri, "Learning spatiotemporal features with 3d convolutional networks," in *ICCV*, 2015.

[5] Jeffrey Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, and Trevor Darrell, "Long-term recurrent convolutional networks for visual recognition and description," in *CVPR*, 2015.

[6] Shugao Ma, Leonid Sigal, and Stan Sclaroff, "Learning activity progression in lstms for activity detection and early detection," in *CVPR*, 2016.

[7] Karen Simonyan and Andrew Zisserman, "Two-stream convolutional networks for action recognition in videos," in *NIPS*, 2014.

[8] Yunbo Wang, Mingsheng Long, Jianmin Wang, and Philip S Yu, "Spatiotemporal pyramid network for video action recognition," in *CVPR*, 2017.

[9] Alberto Montes, Amaia Salvador, and Xavier Giro-i Nieto, "Temporal activity detection in untrimmed videos with recurrent neural networks," *arXiv:1608.08128*, 2016.

[10] Limin Wang, Yu Qiao, and Xiaoou Tang, "Action recognition and detection by combining motion and appearance features," *THUMOS14 Action Recognition Challenge*.

[11] Fabian Caba Heilbron, Juan Carlos Niebles, and Bernard Ghanem, "Fast temporal activity proposals for efficient detection of human actions in untrimmed videos," in *CVPR*, 2016.

[12] Alexander Richard and Juergen Gall, "Temporal action detection using a statistical language model," in *CVPR*, 2016.

[13] Zheng Shou, Dongang Wang, and Shih-Fu Chang, "Temporal action localization in untrimmed videos via multi-stage cnns," in *CVPR*, 2016.

[14] 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 *CVPR*, 2017.

[15] Philippe Weinzaepfel, Zaid Harchaoui, and Cordelia Schmid, "Learning to track for spatio-temporal action localization," in *ICCV*, 2015.

[16] Vicky Kalogeiton, Philippe Weinzaepfel, Vittorio Ferrari, and Cordelia Schmid, "Action tubelet detector for spatio-temporal action localization," *arXiv:1705.01861*, 2017.

[17] Jie Huang, Wengang Zhou, Qilin Zhang, Houqiang Li, and Weiping Li, "Video-based sign language recognition without temporal segmentation," in *AAAI*, 2018.

[18] Ali Borji, Ming-Ming Cheng, Huaizu Jiang, and Jia Li, "Salient object detection: A benchmark," *IEEE T-IP*, 2015.

[19] Lingyan Ran, Yanning Zhang, Wei Wei, and Qilin Zhang, "A hyperspectral image classification framework with spatial pixel pair features," *Sensors*, 2017.

[20] Qilin Zhang and Gang Hua, "Multi-view visual recognition of imperfect testing data," in *ACM MM*, 2015.

[21] Radhakrishna Achanta, Appu Shaji, Kevin Smith, Aurelien Lucchi, Pascal Fua, and Sabine Susstrunk, "S-lic superpixels compared to state-of-the-art superpixel methods," *IEEE T-PAMI*, 2012.

[22] Dong Zhang, Omar Javed, and Mubarak Shah, "Video object co-segmentation by regulated maximum weight cliques," in *ECCV*, 2014.

[23] Le Wang, Gang Hua, Rahul Sukthankar, Jianru Xue, and Nanning Zheng, "Video object discovery and co-segmentation with extremely weak supervision," in *ECCV*, 2014.

[24] Le Wang, Gang Hua, Rahul Sukthankar, Jianru Xue, Zhenxing Niu, and Nanning Zheng, "Video object discovery and co-segmentation with extremely weak supervision," *IEEE T-PAMI*, 2017.

[25] Yuri Boykov, Olga Veksler, and Ramin Zabih, "Fast approximate energy minimization via graph cuts," *IEEE T-PAMI*, 2001.

[26] Le Wang, Jianru Xue, Nanning Zheng, and Gang Hua, "Automatic salient object extraction with contextual cue," in *ICCV*, 2011.

[27] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah, "Ucf101: A dataset of 101 human actions classes from videos in the wild," *arXiv:1212.0402*, 2012.

[28] Yong Jae Lee, Jaechul Kim, and Kristen Grauman, "Key-segments for video object segmentation," in *ICCV*, 2011.

[29] Anestis Papazoglou and Vittorio Ferrari, "Fast object segmentation in unconstrained video," in *ICCV*, 2013.

[30] Nicolas Märki, Federico Perazzi, Oliver Wang, and Alexander Sorkine-Hornung, "Bilateral space video segmentation," in *CVPR*, 2016.