Hierarchical Self-Attention Network for Action Localization in Videos

Rizard Renanda Adhi Pramono, Yie-Tarng Chen, and Wen-Hsien Fang
National Taiwan University of Science and Technology, Taipei, Taiwan, R.O.C.
Email: {d10702801, ytchen, whf}@mail.ntust.edu.tw

Abstract

This paper presents a novel Hierarchical Self-Attention Network (HISAN) to generate spatial-temporal tubes for action localization in videos. The essence of HISAN is to combine the two-stream convolutional neural network (CNN) with hierarchical bidirectional self-attention mechanism, which comprises of two levels of bidirectional self-attention to efficaciously capture both of the long-term temporal dependency information and spatial context information to render more precise action localization. Also, a sequence rescoring (SR) algorithm is employed to resolve the dilemma of inconsistent detection scores incurred by occlusion or background clutter. Moreover, a new fusion scheme is invoked, which integrates not only the appearance and motion information from the two-stream network, but also the motion saliency to mitigate the effect of camera motion. Simulations reveal that the new approach achieves competitive performance as the state-of-the-art works in terms of action localization and recognition accuracy on the widespread UCF101-24 and J-HMDB datasets.

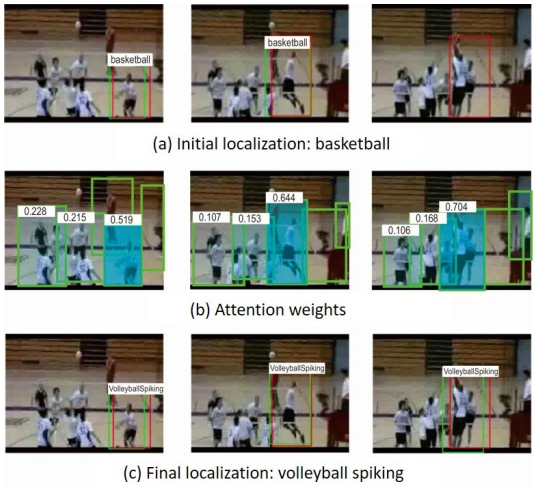

Figure 1: Examples of the impact of self-attention on the group action scenario: (a) initial localization before employing spatio-temporal attention; (b) the attention weight scores are at the top of the bounding boxes; (c) the final localization supervised by the attention, where the localization and the ground truth are in green and red, respectively.

1. Introduction

Owing to its vast potential applications in video content analysis such as video surveillance [1] and video captioning [2], action localization, which performs action classification and generates sequences of bounding boxes related to the locations of the actors, has received much research attention over the past few years. Action localization, however, encounters not only common issues in action recognition such as background clutter, occlusion, intraclass variation, and adverse camera motion, but also the challenging issues that the videos may be untrimmed and have multiple action instances.

A variety of algorithms has been proposed for action recognition and localization [4–7]. For instance, Zolfaghari *et al.* [5] utilized a markov chain model to aggregate multi-stream features. Alwando *et al.* [6] considered an efficient dynamic programming (DP) approach to search for multiple action paths and used an iterative refinement algorithm to obtain more precise bounding boxes. Singh *et al.* [8] incorporated a single shot multi-box detector (SSD) with an incremental DP scheme to generate action tubes with low complexity. The aforementioned methods [4–8], however, consider each frame separately without using the temporal relationship information across the frames, and thus is usually unable to detect actions that contain a sequence of sub-actions such as cricket bowling and basketball. To address this issue, Yang *et al.* [9] proposed a cascaded proposal generation scheme with a location anticipation network to leverage the sequential information across the adjacent frames. Hou *et al.* [10] trained a 3D

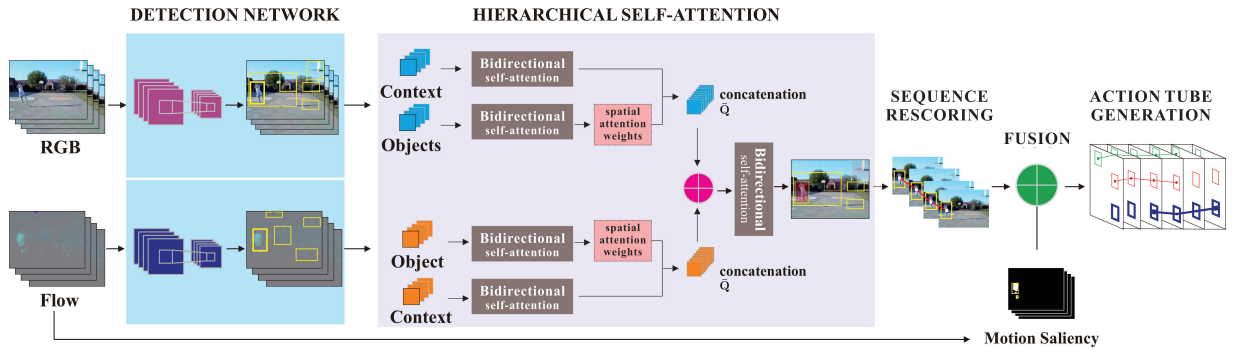

Figure 2: The pipeline of the proposed architecture that comprises of five main steps: First, action detection is conducted using faster R-CNN [3]. Next, HISAN provides spatio-temporal attention to improve the localization accuracy, followed by an SR scheme to rectify low detection scores. The fourth step is a new fusion scheme including motion saliency to reinforce motion information. The last step is action tube generation.

convolutional network to exploit the temporal information from the adjacent frames. Kaloigeton *et al.* [11] proposed a tubelet detector, which can simultaneously produce a sequence of bounding boxes and their detection scores from multiple frames. However, [10, 11] call for high training complexity compared to the 2D convolutional networks. He *et al.* [12] employed a long short-term memory (LSTM) to model the temporal information within action tubes. Li *et al.* [13] considered a recurrent detection network that made use of multi-context from multiple frames to localize actions. However, LSTM processes the information sequentially so in general it has difficulty in learning temporal dependency at distant positions [14]. Gu *et al.* [15] utilized a two-stream inflated 3D ConvNet (I3D) [16] to preserve the temporal information of the two-stream faster R-CNN [6, 7]. Recently, a 3D generalization of capsule network, which can learn different characteristics of actions without using region proposal network (RPN), was proposed in [17]. However, both [15] and [17] have high computational complexity and require a large volume of training data to fully converge.

This paper presents a novel Hierarchical Self-Attention Network (HISAN) to produce spatial-temporal tubes for action localization in videos. The essence of HISAN is to combine the two-stream convolutional neural network (CNN) with the newly devised hierarchical bidirectional self-attention mechanism, which comprises of two levels of bidirectional self-attention to efficaciously capture not only the long-term temporal dependency information but also the spatial context information, to render more precise localization. As shown in Fig. 1, HISAN can learn the structure relationship of key actors to improve the localization accuracy when dealing with the group action scenario, which is difficult to recognize with only a single frame. Also, a sequence rescoring (SR) algorithm is employed to resolve the dilemma of inconsistent detection scores incurred by occlusion or background clutter. Moreover, a new fusion scheme is invoked, which integrates both of the appearance and motion information from the two-stream network, and the motion saliency to mitigate the effect of camera motion that obscures the motion information. Simulations reveal that the new approach achieves competitive performance compared with the state-of-the art works in terms of action localization and recognition accuracy on the widespread UCF101-24 and J-HMDB datasets.

The main contributions of this work are as follows: (i) a two-stream CNN with innovative hierarchical bidirectional self-attention is presented, where both of the spatio-temporal attention and spatial context information are employed to boost the localization accuracy. To the authors' knowledge, it is the first time that self-attention is utilized for action localization; (ii) an SR algorithm, which can rectify the inconsistent detection scores, is employed to reduce the adverse effect of occlusion and background clutter; (iii) a new fusion scheme, which incorporates the motion saliency, is addressed to alleviate the influence of camera motion.

2. Related Works

A vast amount of CNN object detectors has been addressed for action localization [6–9, 11, 13]. The current object detectors can be categorized as either proposal-based [3, 18] or proposal-free [19–21]. Ren *et al.* [3] considered a region proposal network (RPN) to lower the training cost in generating the region proposals. Dai *et al.* [18] developed a position-sensitive region-of-interest (RoI) pooling to resolve the problem of translation invariance in detection. Even though this approach is faster than [3], the detection accuracy is inferior. Red-

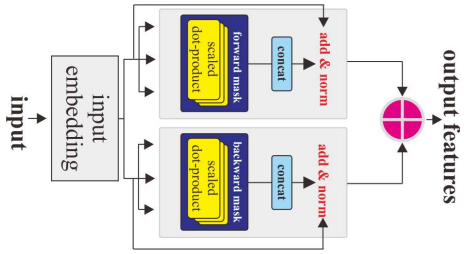

Figure 3: Bidirectional self-attention network.

mon *et al.* [21] designed a fully convolutional network to run multi-scale training with low complexity. SSD [19] used a fixed number of anchors as [3], and multi-scale feature maps to handle objects with different sizes and ratios. Both approaches [19, 21] trade accuracy with complexity and can not locate small-scale objects well [22].

Miscellaneous CNN architectures have been focusing on how to integrate information from multiple modalities to boost the action recognition and localization accuracy. For instance, Simonyan *et al.* [23] developed a two-stream CNN with a late fusion strategy to aggregate the spatial and motion information. Ji *et al.* [24] replaced the conventional 2D-CNN with a 3D ConvNet to capture temporal information from multiple adjacent frames. A markov chain model was employed in [5] to fuse multi-stream features. Choutas *et al.* [25] proposed a stream of human joint information to complement the two-stream architecture.

Attention mechanism has shown to be effective to enhance the performance of CNN when learning fine-grained action in videos [26–29]. Girdhar *et al.* [26] proposed top-down and bottom-up attention to replace the conventional CNN pooling approach. Fang *et al.* [27] built an attention model that focused on correlation of crucial body parts to recognize human-object interactions. Actor-attention regularization was developed in [28] to supervise the spatio-temporal attention on important action regions surrounding the actors. Li *et al.* [29] devised a spatio-temporal attention with diversity regularization to learn various human body parts to identify a person from several different view points.

Temporal dependency has been extensively investigated to obtain more discriminative CNN descriptors. A common solution is to combine recurrent neural network (RNN) or its variant, LSTM, with CNN architectures. For instance, Li *et al.* [30] considered a convolutional soft-attention LSTM to guide motion-based attention around the location of actions. Li *et al.* [13] integrated a two-stage detection network with LSTM to produce more accurate detection. Shi *et al.* [31] replaced the traditional RNN kernels with radial basis function to predict future actions. Recently, a non-local neural network was proposed in [32], which fused temporal dependency information into CNN architectures for video classification. In contrast to the above approaches, our work combines the strength of self-attention [14, 33] in learning temporal dependency with CNN-based object detectors to obtain more precise action localization.

3. Methodology

In this section, we begin with a brief introduction of the action detection network in Sec. 3.1. Our focus is then on three main parts. First, a hierarchical self-attention network is described in Sec. 3.2. An effective SR algorithm to resolve the problem of inconsistent detection scores is considered in Sec. 3.3. Finally, a new fusion scheme that incorporates the motion saliency is addressed in Sec. 3.4. For easy reference, the proposed architecture is illustrated in Fig. 2.

3.1. Detection Network

We use faster R-CNN [3] that consists of spatial and motion CNNs, each of which includes two stages, region proposal generation, and bounding box regression and classification, as our action detection network. First, RPN generates a prescribed number of region proposals, which are likely to contain actions. Thereafter, softmax action scores are given to each region proposal based on the CNN features of the corresponding regions. As [4], the action detection network is built on top of VGG-16. The spatial-CNN takes RGB images as input while the motion-CNN works on optical flow images generated by [34]. The optical flow images are created by stacking the flows in the horizontal and vertical directions.

3.2. Hierarchical Self-Attention Network

This subsection describes the proposed HISAN that provides spatio-temporal attention to rectify inaccurate bounding boxes from the detection network. HISAN consists of several bidirectional self-attention units to model the long-term temporal dependency information.

3.2.1 Bidirectional Self-Attention Unit

We consider the bidirectional self-attention network, as depicted in Fig. 3, that integrates both of the past and future context information to resolve the ambiguity when different videos contain similar movement patterns in the first few frames [33, 35]. Bidirectional self-attention calculates the response of a position in a sequence by relating it to all of the other positions without causality restriction [33]. Since the number of frames in each video may be different, we divide every video into a number of video units with a fixed length of

T_L. Given a sequence of features from a video unit $\mathbf{V} = [\mathbf{v}_1, \cdots, \mathbf{v}_{T_L}] \in \mathbb{R}^{C \times T_L}$, where C is the feature length, the self-attention output at position i, $\mathbf{q}_i$, is defined as [14]

$$\mathbf{q}_i = e(\mathbf{v}_i) + \frac{e(\mathbf{v}_i)}{d(\mathbf{v}_i)} \sum_{\forall k} f(\mathbf{v}_i, \mathbf{v}_k) \tag{1}$$

where $e(\cdot)$ is a linear embedding layer, $d(\cdot)$ is a normalization term, and the pairwise function $f(\mathbf{v}_i, \mathbf{v}_k)$ computes the dot product between the features at positions i and k.

To imbue the pairwise function with temporal alignment information, here, inspired by [14, 33], we modify (1) by adding positional masks. The positional masks are imposed on the self-attention function using either a positional forward mask $\mathcal{M}^f$ or a backward mask $\mathcal{M}^b$, which are defined respectively as

$$\begin{cases} \mathcal{M}^f_{i,k} = 0, & i < k \\ -\infty, & \text{otherwise} \end{cases} \tag{2}$$

$$\begin{cases} \mathcal{M}^b_{i,k} = 0, & i > k \\ -\infty, & \text{otherwise.} \end{cases} \tag{3}$$

Thereby, (1), including the positional masks, is now given by

$$\mathbf{q}^f_i = e(\mathbf{v}_i) + \frac{e(\mathbf{v}_i)}{d(\mathbf{v}_i)} \sum_{\forall k} \sigma(\mathcal{M}^f_{i,k} + f(\mathbf{v}_i, \mathbf{v}_k)) \tag{4}$$

$$\mathbf{q}^b_i = e(\mathbf{v}_i) + \frac{e(\mathbf{v}_i)}{d(\mathbf{v}_i)} \sum_{\forall k} \sigma(\mathcal{M}^b_{i,k} + f(\mathbf{v}_i, \mathbf{v}_k)) \tag{5}$$

where $\sigma(\cdot)$ is a sigmoid function [36]. For each direction, we simultaneously apply the self-attention function on P parallel heads across the feature subspace of dimension C [14], say the feature representation with the forward mask is $\mathbf{Q}^f = [\mathbf{Q}^f_1, \cdots, \mathbf{Q}^f_P]$, where $\mathbf{Q}^f_p = [\mathbf{q}^{f,p}_i, \cdots, \mathbf{q}^{f,p}_{T_L}]^T \in \mathbb{R}^{T_L \times (C/P)}$, $p = 1, \cdots, P$. Afterward, to effectively aggregate the features from these two directions, we combine them together by

$$\bar{\mathbf{Q}} = (\mathbf{w}_f \mathbf{1}^T) \circ \mathbf{Q}^f + (\mathbf{w}_b \mathbf{1}^T) \circ \mathbf{Q}^b \tag{6}$$

in which $\circ$ is an element-wise multiplication [37] and $\mathbf{1} \in \mathbb{R}^{\mathbf{C}}$ is an all-ones vector. All of the weights $\mathbf{w}_f$ and $\mathbf{w}_b \in \mathbb{R}^{T_L}$ are updated in the same manner, *e.g.* $\mathbf{w}_f$ is computed as

$$\mathbf{w}_f = \sigma\left(\mathbf{Q}^f \mathbf{b}_e + \frac{(\mathbf{Q}^f + \mathbf{Q}^b)}{2} \mathbf{w}_e\right) \tag{7}$$

where $\mathbf{b}_e, \mathbf{w}_e \in \mathbb{R}^C$ are weight vectors optimized during the training. Note that such a new combination requires about the same complexity as the original transformer encoder.

3.2.2 Hierarchical Architecture

As shown in Fig. 2, HISAN, which can produce two levels of information, is designed to learn the locations of key actors. The first level aggregates multiple person-object interactions and the context information while the second level integrates the first-level features over time to locate the action. The first level consists of two bidirectional self-attention units, where the first unit processes the spatio-temporal features from multiple bounding boxes while the other one takes the contextual features from video frames. The spatial location is represented by a bounding box $\mathbf{x}_{i,j}$ and a feature vector $\mathbf{u}_{i,j}$, obtained from a fully connected layer of faster R-CNN, where i and j are indices for frames and bounding boxes, respectively. The spatio-temporal features can be expressed as a weighted sum of the feature vectors as

$$\boldsymbol{\phi}_i(\{\mathbf{v}_{i,j}\}, \{\eta_{i,j}\}) = \sum_{j=1}^{N} \eta_{i,j} \times \mathbf{u}_{i,j} \tag{8}$$

where N is the maximum number of bounding boxes in each frame and $\eta_{i,j}$ is the soft attention weight for the bounding box $\mathbf{x}_{i,j}$. The attention weights are normalized such that $\sum_{j=1}^{N} \eta_{i,j} = 1$. The attention weights are updated based on the output of the first bidirectional self-attention unit as follows:

$$\boldsymbol{\eta} = \bar{\mathbf{Q}} \mathbf{W}_{b_1} + \boldsymbol{\phi} \mathbf{W}_{b_2} \tag{9}$$

where $\boldsymbol{\eta} = [\eta_{1,1}, \cdots, \eta_{T_L,N}]$, $\boldsymbol{\phi} = [\boldsymbol{\phi}_1, \cdots, \boldsymbol{\phi}_{T_L}]^T \in \mathbb{R}^{T_L \times C}$, and $\mathbf{W}_{b_1}$ and $\mathbf{W}_{b_2} \in \mathbb{R}^{C \times N}$ are linear weights.

In the second level, the bidirectional representation and the attention weights from the spatial and motion networks are averaged and then propagated to another bidirectional self-attention network. A softmax classifier is then employed to obtain the class probabilities. Note that the hierarchical self-attention network computes not only the attention weights but also the classification score $\mathbf{y}_{i,j} = \{y^0_{i,j}, \cdots, y^{cls}_{i,j}\}$, where cls is the number of classes and $y^0_{i,j}$ is the background score indicating the probability of no action existing in frame i. We use the classification score $y^c_{i,j}$ and the attention weight $\eta_{i,j}$ to adjust the detection score for a specific class c by

$$s_c(\mathbf{x}_{i,j}) = s^{init}_c(\mathbf{x}_{i,j}) \times \eta_{i,j} \times y^c_{i,j} \tag{10}$$

where $s^{init}_c(\mathbf{x}_{i,j})$ is the initial detection score from the detection network.

3.3. Sequence Rescoring

In our framework, the frame-level detection is linked together with a DP algorithm which gives penalty to

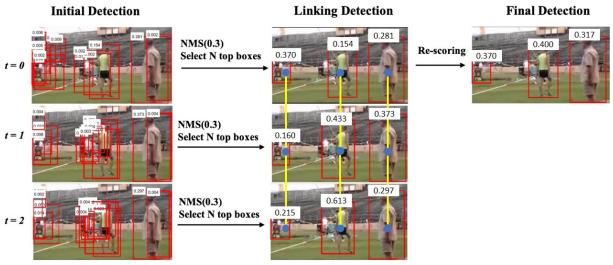

Figure 4: Overview of the SR algorithm, which consists of three stages: the selection, linking and rescoring of the bounding boxes.

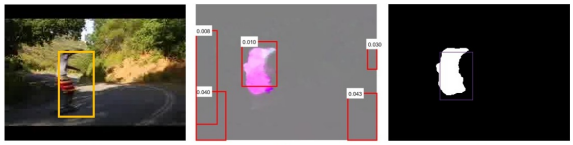

Figure 5: Overview of the motion saliency: (a) the key actor inside the yellow box; (b) the bounding boxes from the motion-CNN; (c) the salient region in white.

bounding boxes that do not overlap in time. However, in some cases, the detection score is low due to occlusion or background clutter. In this scenario, even though the overlap is high, the bounding box may not be linked to the correct path because of the low detection score. To overcome this setback, we devise an SR algorithm succeeding the output of HISAN.

This algorithm is divided into three stages, as depicted in Fig. 4. First, a non-maximum suppression (NMS) [38] is invoked to reduce the number of bounding boxes to $N_{nms} < N$ in every frame of a sequence of frames, which span an interval from T_s to T_e. Afterward, every bounding box is linked together with the bounding box with the maximum overlap in the adjacent frames. If the overlap is below a prescribed threshold, the link is terminated. In the last stage, the bounding box in the current frame is then rescored by

$$s_c(\mathbf{x}_{i=T_s,j}) = \max\left(\frac{\sum_{i=T_s}^{T_e} s_c(\mathbf{x}_{i+1,l_i})}{T_e - T_s + 1}, s_c(\mathbf{x}_{T_s,j})\right) \quad (11)$$

where $l_i = \underset{l}{\text{argmax}} \ \text{IoU}(\mathbf{x}_{i+1,l}, \mathbf{x}_{i,l_{i-1}})$, $i > T_s$, and $\mathbf{x}_{T_s,l_{T_s}} = \mathbf{x}_{T_s,j}$, in which IoU denotes the intersection-over-union [4]. Based on (11), the low detection score due to occlusion and background clutter can be enhanced so the bounding box can be linked to the correct path.

3.4. Fusion Strategy

We consider a new fusion scheme, which incorporates motion saliency, to highlight the motion information. The motion saliency is included based on the consideration that false detection may occur from the motion-CNN due to small camera movement. As an example, given an RGB image in Fig. 5 (a), the moving actor cannot be distinguished based on the motion-CNN scores, as shown in Fig. 5 (b). On the contrary, the motion saliency, as shown in Fig. 5 (c), captures the correct region associated with the moving actor.

Suppose f_m^i is the magnitude of the optical flow in frame i and I_i is the entire region in this frame. If G is a region in frame i, then $g_m^i(G) = \frac{1}{|G|}\sum_{u\in G} f_m^i(u)$ indicates how motion salient G is. G is considered motionless if $g_m^i(G) < \delta$, where δ is a prescribed threshold. Let H denote the entire set of salient regions defined by $H = \{G | g_m^i(G) \geq \delta, G \in I_i\}$. A collection of unconnected regions $\{H_1, \cdots, H_{sp}\} \subseteq H$ are labelled from the region H based on 8-connected pixel connectivity. The motion saliency score of the bounding box $\mathbf{x}_{i,j}$ is thus defined as

$$w(\mathbf{x}_{i,j}) = \mu \times \max_{o\in\{1,\cdots,sp\}} \frac{|\mathbf{x}_{i,j} \cap H_o|}{|\mathbf{x}_{i,j} \cup H_o|} \quad (12)$$

where $|\cdot|$ denote the cardinality of a set of pixels. Note that the motion saliency in (12) is different from the ones addressed in [6, 39]. Compared with [6], (12) is more amenable to multiple action instances because every bounding box is associated with at most one salient region. Also, in [39], H is employed directly to filter motionless proposals, so it is likely to yield more false negatives. Based on (12), if the bounding box $\mathbf{x}_{i,j}$ has a larger saliency score, it is more probable that it encompasses a moving actor.

Given the bounding boxes from the spatial-CNN and the motion-CNN, $\{\mathbf{x}_{i,j}\}$, $j = 1, \cdots, 2N_{nms}$, the final detection score of each bounding box is then given by

$$s_c^*(\mathbf{x}_{i,j}) = s_c(\mathbf{x}_{i,j}) + w(\mathbf{x}_{i,j}) \quad (13)$$

Table 1: Parameter settings for training the faster R-CNN and HISAN.

	Faster R-CNN	HISAN
Pre-trained model	ImageNet [40]	-
Loss Function	Log + smooth L1 [3]	Cross entropy [41]
Optimizer	SGD	Adam ($\beta_1 = 0.9$, $\beta_2 = 0.0009$)
Learning rate	0.001	0.0001
Momentum	0.9	-
Decay	0.0005	0.01
Iterations	320k (UCF101-24), 180k (J-HMDB)	120k

3.5. Action Tube Generation

After the fusion, the frame-level detection boxes are linked together to generate action tubes. Note that action localization and multi-object tracking are two different problems since the former requires action classification

to link actions across frames. Also, as opposed to multi-object tracking, in general only key actors are localized in the action localization problems [39]. Therefore, we opt to use a lightweight DP algorithm instead of a more sophisticated multi-object tracking algorithm [42, 43], which uses a data association algorithm to link trackers with detections. Denote a set of video paths in a video with M frames as T and a set of bounding boxes $\{\mathbf{x}_{i,j}\}$ along with the final detection scores $\{s_c^*(\mathbf{x}_{i,j})\}$ obtained from the fusion scheme discussed in Sec. 3.4. From T, we attempt to find sets of bounding boxes from $i = 1$ to $i = M$, which are likely to contain a single action instance. These sets are termed action tubes. Following [6], the action tubes are found by maximizing the accumulative score given by

$$\sum_{T_1,\cdots,T_{2N_{nms}}} \sum_{i=1}^{M-1} A_c(\mathbf{x}_{i,j}, \mathbf{x}_{i+1,g}), \tag{14}$$

where $A_c(\mathbf{x}_{i,j}, \mathbf{x}_{i+1,g}) = s_c^*(\mathbf{x}_{i+1,g}) + \alpha \times \mathsf{IoU}(\mathbf{x}_{i,j}, \mathbf{x}_{i+1,g})$, in which j and g are ordered by the paths $\{T_n\}_{n=1}^{2N_{nms}}$ and α is a weighting parameter. The optimization problem can be solved using the multiple path search algorithm [6] that simultaneously finds all possible paths within one iteration.

In an untrimmed video, an action usually occupies only a fraction of the entire video duration. Consequently, it is required to find the temporal duration of the action within the action tube. To do so, we use the same algorithm as [4], which uses DP to solve the tube energy maximization with the restriction of smoothness of scores across consecutive frames.

4. Experimental Results

4.1. Datasets

The experiments are conducted using two widespread action localization datasets: UCF101-24 [44] and J-HMDB [45], both of which have a variety of characteristics, to validate our approach.

UCF101-24 dataset [8,44]: This dataset contains 3194 annotated videos and 24 action classes. It encompasses several viewpoints of actors, illumination conditions, and camera movements. Most of the videos in this dataset are untrimmed.

J-HMDB dataset [45]: This dataset is composed of 928 trimmed videos and 21 action classes. Several challenges encountered in this dataset include occlusion, background clutter, and high inter-class similarity.

4.2. Experimental Settings

The learning process consists of training faster R-CNN and HISAN, which are conducted separately. The faster R-CNN is trained without feature sharing [4]. For easy reference, Table 1 summarizes the hyperparameters of these training procedures. All of the experiments are based on the same protocols provided by UCF101-24 [4,44] and J-HMDB [4,45]. We use a video unit with length T_L = 30 and 15 for UCF101-24 and J-HMDB, respectively, which is decided by the minimum length of the videos in the datasets. We choose the feature dimension, $C = 4096$, as the dimension of the *fc7* of the detection network. We set the number of heads as $P = 8$ and a dropout rate $= 0.1$ as suggested in [14, 33].

We apply box voting scheme [46] to the bounding boxes from faster R-CNN with a prescribed IoU threshold $= 0.5$. Following [6], the parameters α and δ are determined via the grid search with cross-validation, and a 5×5 spatial window is used to construct R. The length of the SR interval is set as 15 and the parameter μ as 0.7. We use NMS with threshold $= 0.3$ and set N_{nms} as 5 to have the same number of bounding boxes per frame as [8]. The widespread video and frame mean-average precision (mAP) are employed as metrics of accuracy for the spatio-temporal action tube detection. Following [7,8,10], we assess the classification accuracy based on the action tube with the largest accumulated score.

Table 2: Action localization results on UCF101-24 with various combinations of strategies.

Strategy			Video mAP				Frame mAP
HISAN	SR	Motion Saliency	0.2	0.5	0.75	0.5:0.95	0.5
			72.96	43.69	15.83	19.02	63.80
✓			78.35	47.28	21.76	23.30	70.47
✓	✓		79.05	48.44	21.87	23.71	71.03
✓	✓	✓	**80.42**	**49.50**	**22.35**	**24.05**	**73.71**

Table 3: Action localization results on J-HMDB with various combinations of strategies.

Strategy			Video mAP				Frame mAP
HISAN	SR	Motion Saliency	0.2	0.5	0.75	0.5:0.95	0.5
			72.50	71.24	43.52	42.09	60.21
✓			84.09	82.98	48.78	48.56	76.04
✓	✓		84.30	83.27	48.92	49.02	76.53
✓	✓	✓	**85.97**	**84.02**	**52.76**	**50.50**	**76.72**

4.3. Ablation Studies

• **Impact of Hierarchical Bidirectional Self-Attention**: First, we inspect the performance improvement with the spatio-temporal attention generated by HISAN on UCF101-24 and J-HMDB, as shown respectively in Tables 2 and 3, from which we can note that with the incorporation of this mechanism, the video mAP of the two-stream CNN can be improved by about 2.5% to 5% and 5% to 12% on UCF101-24 and J-HMDB, respectively. Also, the frame mAP can be enhanced by about 6% and 16% on UCF101-24 and J-HMDB, respectively. This is because this mechanism

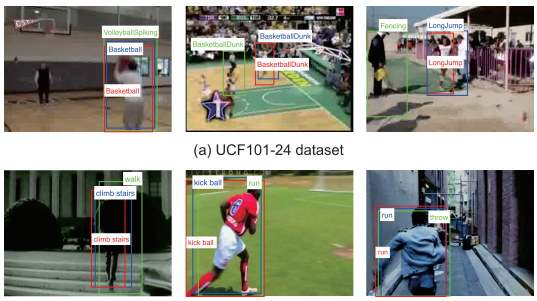

Figure 6: Some action localization results with and without HISAN are in blue and green, respectively, while the ground truth is in red.

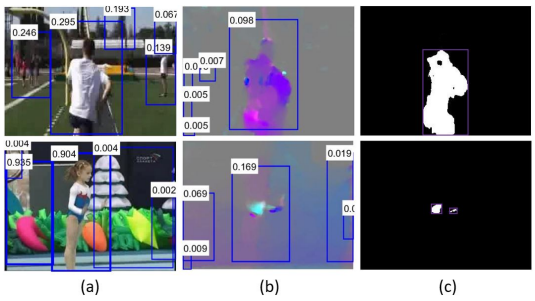

Figure 7: Example cases where the new fusion helps the localization: (a) the detection by the spatial-CNN; (b) the detection by the motion-CNN; (c) the motion-saliency boxes.

exploits temporal dependency to guide the attention on the location of the action. The improvement on J-HMDB is more significant than that on UCF101-24, as the former has many action classes with similar sequences of sub-actions, as depicted in Fig. 6, which requires more temporal dependency information to classify the actions. The effect of the spatio-temporal attention is also illustrated in Fig. 1, from which we can see that self-attention can help locate the actions, especially in the group action scenario that is difficult to recognize with only the information of a single frame.

• **Impact of SR:** Next, we scrutinize the effect of the SR algorithm, which is devised to handle the inconsistent detection scores due to occlusion. As shown in Tables 2 and 3, together with SR the video mAP can be further boosted by about 0.3% to 1.5% and 0.2% to 0.4% on UCF101-24 and J-HMDB, respectively. Moreover, the frame mAP can be improved by about 0.5% on both datasets. The improvement on J-HMDB is less because there is only a single action instance in all videos, so there is less occlusion in this dataset.

• **Impact of the New Fusion:** Finally, we examine the new fusion scheme, which incorporates motion saliency to diminish the effect of small camera motion. We can notice from Tables 2 and 3 that the new scheme improves the video mAP by about 1.1% to 2.3% and 0.7% to 1.7% on UCF101-24 and J-HMDB, respectively. Furthermore, the frame mAP can be boosted by about 2% and 0.2% on UCF101-24 and J-HMDB, respectively. The improvement on UCF101-24 is more substantial, as the videos in this dataset contain more camera motion. As an illustration, some cases where the motion saliency helps action localization are depicted in Fig. 7, from which we can see that the saliency maps contain the true region with actions so the low detection scores from both of the spatial-CNN and motion-CNN can be bolstered with the motion-saliency score via (13).

Based on the observations from the above simulations, to attain superior performance, the proposed HISAN is equipped with the SR algorithm and the new fusion in the following simulations.

4.4. Comparison with State-of-the-Art Works

This section compares our method, which uses either VGG-16 backbone or more sophisticated ResNet101 + FPN [47], with some recently proposed approaches for action localization and for action recognition.

First, we consider action localization problem. The comparison with ten baselines, including Zolfaghari *et al.* [5], Alwando *et al.* [6], Singh *et al.* [8], CPLA [9], T-CNN [10], ACT [11], TPN [12], RTP + RTN [13], Gu *et al.* [15], and Duarte *et al.* [17], in terms of video mAP for different IoU's on UCF101-24 is shown in Table 4, from which we can note that CPLA [9] provides better performance than [5, 10, 12] with the incorporation of an anticipation network that reuses the detection in the previous frames to rectify inaccurate detection in the current frame. By using an iterative refinement scheme, [6] consistently outperforms [9] for all IoU's. ROAD [8] incorporates SSD to obtain competitive localization performance with low complexity. ACT [11] utilizes a multi-frame object detector to simultaneously regress the bounding boxes from a sequence of frames. A combination of recurrent and multi-context information is explored in [13] to enhance the detection accuracy. Gu *et al.* [15] integrates the two-stream I3D and faster R-CNN to attain more accurate localization. Duarte *et al.* [17], which considers a capsule network, attains the best results. However, capsule networks call for high complexity due to a routing-by-agreement mechanism. Apart from [15, 17], our approach outperforms all of the aforementioned methods by incorporating the hierarchical bidirectional self-attention to boost the detection accuracy and employing a new fusion scheme with motion saliency to leverage the motion information.

Table 4: Comparison of the action localization performance on UCF101-24. The best results are bold-faced.

Method	Video mAP			
	0.2	0.5	0.75	0.5:0.95
T-CNN [10]	39.20	-	-	-
Zolfaghari *et al.* [5]	47.61	26.79	-	-
TPN [12]	71.69	-	-	-
CPLA [9]	73.54	37.80	-	-
Alwando *et al.* [6]	72.90	41.10	-	-
ROAD [8]	73.50	46.30	15.00	20.40
ACT [11]	76.50	49.20	19.70	23.40
RTP + RTN [13]	77.90	-	-	-
Gu *et al.* [15]	-	59.90	-	-
Duarte *et al.* [17]	**97.40**	**82.0**	**26.12**	**36.20**
Ours w/ VGG-16	80.42	49.50	22.35	24.05
Ours w/ ResNet-101+FPN	82.30	51.47	23.48	24.93

Table 5: Comparison of the action localization performance on J-HMDB. The best results are bold-faced.

Method	Video mAP			
	0.2	0.5	0.75	0.5:0.95
ROAD [8]	73.80	72.00	44.50	41.60
ACT [11]	74.20	73.70	52.10	44.80
Zolfaghari *et al.* [5]	78.20	73.47	-	-
T-CNN [10]	78.40	76.90	-	-
TPN [12]	79.70	76.96	-	-
Alwando *et al.* [6]	79.78	78.26	-	-
Gu *et al.* [15]	-	78.60	-	-
RTP + RTN [13]	82.7	81.30	-	-
Duarte *et al.* [17]	**95.40**	61.95	3.01	19.06
Ours w/ VGG-16	85.97	84.02	52.76	50.50
Ours w/ ResNet-101+FPN	87.59	**86.49**	**53.83**	**51.26**

For J-HMDB, we make a comparison with nine baselines, including Zolfaghari *et al.* [5], Alwando *et al.* [6], ROAD [8], T-CNN [10], ACT [11], TPN [12], RTP + RTN [13], Gu *et al.* [15], and Duarte *et al.* [17]. From Table 5, we can note that [17], which incorporates a capsule network to learn more semantic information, can attain the best performance on IoU = 0.2. However, it does not work well on this smaller but challenging dataset as its performance drops substantially for higher IoU's. T-CNN [10] excels [5, 8, 11] by using a 3D CNN to generate more precise 3D proposals. TPN [12] achieves slightly better results than [10] by using LSTM to learn video-level information. Using faster R-CNN with an iterative refinement scheme to obtain more accurate bounding boxes, [6] obtains superior performance over [12]. Gu *et al.* [15] attains slightly better performance by using a two-stream I3D to localize actions more precisely. RTP + RTN [13] attains even superior performance by integrating recurrent mechanism into both of the proposal and classification networks. Our method surpasses [13] for all IoU's. This is because our HISAN can learn long-term temporal dependency that is crucial in detecting actions with similar sub-actions such as 'climb stairs' and 'walk'.

Next, we compare the action recognition performance on UCF101-24 with some of the above baselines, which reported their performance on this problem, as shown in Table 6, from which we can see that T-CNN [10] achieves better performance compared with [8, 48] by exploiting the discriminative features provided by 3D ConvNet. Using an RBF kernelized RNN coupled with adversarial training strategy, [31] substantially outperforms [10]. Our approach, which utilizes the hierarchical bidirectional self-attention to leverage the temporal information, demonstrates the best performance.

Table 6: Comparison of action recognition results on UCF101-24 and J-HMDB. The best results are bold-faced.

Method	Accuracy	
	UCF101-24	J-HMDB
Temporal Fusion [48]	89.27	-
ROAD [8]	92.00	63.00
T-CNN [10]	94.40	67.20
RBF Kernelized RNN [31]	98.00	73.00
R-STAN [28]	-	79.20
PoTion [25]	-	85.50
Ours w/ VGG-16	**99.45**	**86.80**

The comparison of the action recognition performance is also made on J-HMDB with some of the above baselines, as shown in Table 6, from which we can see that [31] outperforms [10] because the temporal dependency is not well trained in the 3D ConvNet as opposed to the RBF kernelized RNN. R-STAN [28], a unified two-stream LSTM network that provides attention to regions surrounding the actions, achieves higher accuracy. PoTion [25] surpasses [28] by combining pose motion network with the two-stream I3D. Our approach achieves the best performance by learning long-term temporal dependency and spatial context information.

5. Conclusions

This paper has developed an effective architecture, HISAN, a combination of two-stream CNN with the newly devised hierarchical bidirectional self-attention for action localization in videos, to learn the long-term temporal dependency and the spatial context information. In addition, an SR algorithm is employed to rectify the inconsistent detection scores and a new motion saliency assisted fusion scheme is addressed to highlight the motion information. Simulations show that the new approach attains competitive performance compared with state-of-the-art methods on the UCF101-24 and J-HMDB datasets.

Acknowledgment

This work was supported by the Ministry of Science and Technology, R.O.C. under contracts MOST 107-2221-E-011-124 and MOST 107-2221-E-011-078-MY2.

References

[1] In Su Kim, Hong Seok Choi, Kwang Moo Yi, Jin Young Choi, and Seong G Kong. Intelligent visual surveillance—a survey. *International Journal of Control, Automation and Systems*, 8(5):926–939, 2010.

[2] Jeffrey Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, and Trevor Darrell. Sequence to sequence-video to text. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4534–4542, 2015.

[3] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In *Proceedings of the Neural Information Processing Systems*, pages 91–99, 2015.

[4] Suman Saha, Gurkirt Singh, Michael Sapienza, Philip HS Torr, and Fabio Cuzzolin. Deep learning for detecting multiple space-time action tubes in videos. In *Proceedings of the British Machine Vision Conference*, pages 58.1–58.13, 2016.

[5] Mohammadreza Zolfaghari, Gabriel L Oliveira, Nima Sedaghat, and Thomas Brox. Chained multi-stream networks exploiting pose, motion, and appearance for action classification and detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2904–2913, 2017.

[6] Erick Hendra Putra Alwando, Yie-Tarng Chen, and Wen-Hsien Fang. CNN-based multiple path search for action tube detection in videos. *To appear in IEEE Transactions on Circuits and Systems for Video Technology*, 2018.

[7] Xiaojiang Peng and Cordelia Schmid. Multi-region two-stream R-CNN for action detection. In *Proceedings of the European Conference on Computer Vision*, pages 744–759, 2016.

[8] Gurkirt Singh, Suman Saha, Michael Sapienza, Philip HS Torr, and Fabio Cuzzolin. Online real-time multiple spatiotemporal action localisation and prediction. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3637–3646, 2017.

[9] Zhenheng Yang, Jiyang Gao, and Ram Nevatia. Spatio-temporal action detection with cascade proposal and location anticipation. In *Proceedings of the British Machine Vision Conference*, 2017.

[10] Rui Hou, Chen Chen, and Mubarak Shah. Tube convolutional neural network (T-CNN) for action detection in videos. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5822–5831, 2017.

[11] Vicky Kalogeiton, Philippe Weinzaepfel, Vittorio Ferrari, and Cordelia Schmid. Action tubelet detector for spatio-temporal action localization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4405–4413, 2017.

[12] Jiawei He, Zhiwei Deng, Mostafa S Ibrahim, and Greg Mori. Generic tubelet proposals for action localization. In *Proceedings of the IEEE Winter Conference on Applications of Computer Vision*, pages 343–351, 2018.

[13] Dong Li, Zhaofan Qiu, Qi Dai, Ting Yao, and Tao Mei. Recurrent tubelet proposal and recognition networks for action detection. In *Proceedings of the European Conference on Computer Vision*, pages 303–318, 2018.

[14] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Proceedings of the Advances in Neural Information Processing Systems*, pages 5998–6008, 2017.

[15] Chunhui Gu, Chen Sun, David A Ross, Carl Vondrick, Caroline Pantofaru, Yeqing Li, Sudheendra Vijayanarasimhan, George Toderici, Susanna Ricco, Rahul Sukthankar, et al. AVA: A video dataset of spatio-temporally localized atomic visual actions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6047–6056, 2018.

[16] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017.

[17] Kevin Duarte, Yogesh Rawat, and Mubarak Shah. Videocapsulenet: A simplified network for action detection. In *Advances in Neural Information Processing Systems*, pages 7610–7619, 2018.

[18] Jifeng Dai, Yi Li, Kaiming He, and Jian Sun. R-FCN: Object detection via region-based fully convolutional networks. In *Proceedings of the Advances in Neural Information Processing Systems*, pages 379–387, 2016.

[19] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. SSD: Single shot multibox detector. In *Proceedings of the European Conference on Computer Vision*, pages 21–37, 2016.

[20] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 779–788, 2016.

[21] Joseph Redmon and Ali Farhadi. YOLO9000: Better, faster, stronger. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6517–6525, 2017.

[22] Jonathan Huang, Vivek Rathod, Chen Sun, Menglong Zhu, Anoop Korattikara, Alireza Fathi, Ian Fischer, Zbigniew Wojna, Yang Song, Sergio Guadarrama, et al. Speed/accuracy trade-offs for modern convolutional object detectors. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7310–7311, 2017.

[23] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. *Proceedings of the Neural Information Processing Systems*, pages 568–576, 2014.

[24] Shuiwang Ji, Wei Xu, Ming Yang, and Kai Yu. 3D convolutional neural networks for human action recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(1):221–231, 2013.

[25] Vasileios Choutas, Philippe Weinzaepfel, Jérôme Revaud, and Cordelia Schmid. PoTion: Pose motion representation for action recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7024–7033, 2018.

[26] Rohit Girdhar and Deva Ramanan. Attentional pooling for action recognition. In *Proceeding of the Advances in Neural Information Processing Systems*, pages 34–45, 2017.

[27] Hao-Shu Fang, Jinkun Cao, Yu-Wing Tai, and Cewu Lu. Pairwise body-part attention for recognizing human-object interactions. In *Proceedings of the European Conference on Computer Vision*, pages 51–67, 2018.

[28] Wenbin Du, Yali Wang, and Yu Qiao. Recurrent spatial-temporal attention network for action recognition in videos. *IEEE Transactions on Image Processing*, 27(3):1347–1360, 2018.

[29] Shuang Li, Slawomir Bak, Peter Carr, and Xiaogang Wang. Diversity regularized spatiotemporal attention for video-based person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 369–378, 2018.

[30] Zhenyang Li, Kirill Gavrilyuk, Efstratios Gavves, Mihir Jain, and Cees GM Snoek. VideoLSTM convolves, attends and flows for action recognition. *Computer Vision and Image Understanding*, 166:41 – 50, 2018.

[31] Yuge Shi, Basura Fernando, and Richard Hartley. Action anticipation with RBF kernelized feature mapping RNN. In *Proceedings of the European Conference on Computer Vision*, pages 305–322, 2018.

[32] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7794–7803, 2018.

[33] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[34] Thomas Brox, Andrés Bruhn, Nils Papenberg, and Joachim Weickert. High accuracy optical flow estimation based on a theory for warping. In *Proceedings of the European Conference on Computer Vision*, pages 25–36. Springer, 2004.

[35] Bharat Singh, Tim K Marks, Michael Jones, Oncel Tuzel, and Ming Shao. A multi-stream bi-directional recurrent neural network for fine-grained action detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1961–1970, 2016.

[36] 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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2625–2634, 2015.

[37] William McGuire, Richard H Gallagher, and H Saunders. *Matrix Structural Analysis*. 2000.

[38] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 580–587, 2014.

[39] Georgia Gkioxari and Jitendra Malik. Finding action tubes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 759–768, 2015.

[40] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009.

[41] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. The MIT Press, 2016.

[42] Yu Xiang, Alexandre Alahi, and Silvio Savarese. Learning to track: Online multi-object tracking by decision making. In *Proceedings of the IEEE International Conference on Computer Vision*, 2015.

[43] Amir Sadeghian, Alexandre Alahi, and Silvio Savarese. Tracking the untrackable: Learning to track multiple cues with long-term dependencies. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 300–311, 2017.

[44] Khurram Soomro, Amir Roshan Zamir, and M Shah. UCF101: A dataset of 101 human actions classes from videos in the wild. *Center for Research in Computer Vision*, 2012.

[45] Hueihan Jhuang, Juergen Gall, Silvia Zuffi, Cordelia Schmid, and Michael J Black. Towards understanding action recognition. In *Proceedings of the IEEE Conference on Computer Vision*, pages 3192–3199, 2013.

[46] S. Gidaris and N. Komodakis. Object detection via a multi-region and semantic segmentation-aware cnn model. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1134–1142, 2015.

[47] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2117–2125, 2017.

[48] Zhaoxuan Fan, Tianwei Lin, Xu Zhao, Wanli Jiang, Tao Xu, and Ming Yang. An online approach for gesture recognition toward real-world applications. In *Proceeding of the International Conference on Image and Graphics*, pages 262–272, 2017.