Semantic Cues Enhanced Multimodality Multistream CNN for Action Recognition

Zhigang Tu, *Member, IEEE*, Wei Xie, *Member, IEEE*, Justin Dauwels, *Senior Member, IEEE*, Baoxin Li, *Senior Member, IEEE*, and Junsong Yuan, *Senior Member, IEEE*

***Abstract*— This paper addresses the issue of video-based action recognition by exploiting an advanced multistream convolutional neural network (CNN) to fully use semantics-derived multiple modalities in both spatial (appearance) and temporal (motion) domains, since the performance of the CNN-based action recognition methods heavily relates to two factors: semantic visual cues and the network architecture. Our work consists of two major parts. First, to extract useful human-related semantics accurately, we propose a novel spatiotemporal saliency-based video object segmentation (STS) model. By fusing different distinctive saliency maps, which are computed according to object signatures of complementary object detection approaches, a refined STS maps can be obtained. In this way, various challenges in the realistic video can be handled jointly. Based on the estimated saliency maps, an energy function is constructed to segment two semantic cues: the actor and one distinctive acting part of the actor. Second, we modify the architecture of the two-stream network (TS-Net) to design a multistream network that consists of three TS-Nets with respect to the extracted semantics, which is able to use deeper abstract visual features of multimodalities in multi-scale spatiotemporally. Importantly, the performance of action recognition is significantly boosted when integrating the captured human-related semantics into our framework. Experiments on four public benchmarks—JHMDB, HMDB51, UCF-Sports, and UCF101—demonstrate that the proposed method outperforms the state-of-the-art algorithms.**

***Index Terms*— Action recognition, multi-stream CNN, spatiotemporal saliency estimation, video object detection, semantic cues, multi-modalities.**

Manuscript received January 17, 2018; revised April 1, 2018; accepted April 21, 2018. Date of publication April 25, 2018; date of current version May 3, 2019. This work is supported in part by the Singapore Ministry of Education Academic Research Fund Tier 2 under Grant MOE2015-T2-2-114, in part by the National Natural Science Foundation of China under Grant 61501198, in part by the Natural Science Foundation of Hubei Province under Grant 2014CFB461, and in part by the University at Buffalo. This paper was recommended by Associate Editor G.-J. Qi. *(Corresponding author: Zhigang Tu.)*

Z. Tu is with the State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan 430079, China (e-mail: tuzhigang1986@gmail.com).

W. Xie is with the School of Computer, Central China Normal University, Wuhan 430079, China.

J. Dauwels are with the School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore 637553 (e-mail: jdauwels@ntu.edu.sg).

B. Li is with the School of Computing, Informatics, Decision System Engineering, Arizona State University, Tempe, AZ 85287 USA.

J. Yuan is with the Computer Science and Engineering Department, State University of New York at Buffalo, Buffalo, NY 14260-2500 USA.

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TCSVT.2018.2830102

I. Introduction

With the rapid growth of video data, there is an urgent need for techniques to analyze the video contents automatically [2], [3]. Action recognition, which aims at enabling machines to recognize human actions in real-world videos efficiently and accurately, has drawn a lot of attention in both the academia and industry [4]–[8]. Despite that significant progresses have been achieved recently, action recognition is still a challenging problem. Intra-class and inter-class variations of humans in appearance and motion are the main difficulties. Moreover, varying environments of the actions further increase the complexity of human and motion localization [3]. Actions are characterized by the temporal evolution of visual appearance governed by motion. While spatial cues have been well studied in many current methods, resulting in effective features, there is still a need for developing valid methods for employing temporal cues and their variations in action recognition.

Recently, Convolutional Neural Networks (CNNs), which are able to learn discriminative features from raw data automatically, have shown remarkable success in image analysis, such as image classification [9], object detection [10], [11], human face recognition, and event classification. For action recognition in videos, however, deep learning methods have yet to demonstrate its effectiveness when being compared with traditional hand-crafted approaches [12].

There are two main factors that limit the performance of CNN-based video action recognition. First, compared to still images, videos contain complex conditions such as intra-class variations caused by occlusion and/or change in viewpoint and/or background [26] temporally, and long-range temporal structure is significant to understand the dynamics in action videos [4]. Current CNN architectures are still unable to make full use of the temporal features. The spatiotemporal characteristics of videos call for more efficient network architectures [35]. E.g., a desirable architecture should take advantage of both the temporal (motion) and the spatial (appearance) features with multi-modalities. The second limitation of CNNs concerns the semantic nature of videos. Videos contain rich semantic information, and an action is often related to many semantic visual cues, like the scene, human, and human body parts in interaction [7], [19]. While detecting semantics from videos is difficult, figuring out the correspondences between action types and semantic features is even more difficult. Until

now, only a few works have investigated *how* to effective extract semantic cues and *what* is the role of semantics in video-based action recognition.

These issues motivate us to improve the performance of action recognition in videos with deep learning in two aspects, where the successful two-stream network (TS-Net) [5] is adopted as the baseline in this work: 1) Exploiting useful human-related semantic cues; 2) Improving the architecture of the TS-Net and integrating semantics-derived multi-modalities spatiotemporally into our framework.

Actions are defined as intentional bodily movement of biological agents [13]. This definition reveals that the success of action recognition depends on two visual cues: appearance and motion, and one subject: biological agent (i.e. actor). Inspired by this characteristic, Simonyan and Zisserman [5] proposed a TS-Net, where two CNNs are trained to separately extract appearance and motion information explicitly. In short, one CNN is used to tackle video frames (spatial stream) to capture appearance features for locating a person, and another CNN is applied to process the stacked optical flows (motion stream) to obtain features for detecting bodily movements (action). Late fusion was employed to integrate the softmax scores of two CNNs either by averaging or with a linear classifier. TS-Net faces some challenges: 1) It is unable to represent *what* moves *where*, and *how* the spatial cue and temporal cue evolve over time. To overcome these drawbacks, we introduce a new spatiotemporal fusion architecture upon the TS-Net [6]. 2) The network architecture used in TS-Net is shallow [23]. Recent works demonstrate that deeper CNN architectures boost the performance of action recognition [4], [18]. We test different deeper CNNs for the spatial stream and the temporal stream separately to find more suitable networks. 3) Only two simple modalities (full RBG image and optical flow) are not sufficient to deal with some complicated realistic challenges. We exploit other modalities, which are based on our extracted human-related semantics, to assist action recognition. 4) The TS-Net considers action recognition on a single scale. In realistic videos, action often varies in scales. We extend the TS-Net to capture better features in multi-scale.

Semantic cues, such as scene context [5], hand regions [20], human pose [8], human body [3], [19], and interacting objects [21], are beneficial for video-based action recognition [7]. Especially, the human-related proposals show their attractive advantage in improving the performance of action recognition. However, the state of the art methods [7], [19], [21] simply introduce the image-based object detection techniques, e.g., R-CNN [10], Faster R-CNN [11], selective search [22], to capture object candidates. In fact, they are not suited to be directly applied to detect objects in videos. Besides, these methods produce too many object proposals, but only some of them are useful for recognizing human actions, as one action is usually strongly related to the movement of the actor. In other words, the human body and his/her motion salient part are the most important semantic cues for action recognition.

Towards addressing the above-mentioned two problems, we propose a spatiotemporal saliency based video object segmentation (STS-VOS) model to extract two semantic cues in videos – actor, and its most motion salient body part (we call it acting part). Unlike the previous salient video object detection approaches focusing on exploiting or incorporating different saliency cues, we aim to capture object signatures which can be estimated by any kinds of complementary video object segmentation methods to assist compute more accurate spatiotemporal saliency maps (SSM) [1]. Three distinctive saliency maps, which are obtained from an appearance-dominated approach, a motion-dominated technique, and a deep feature-based method, are incorporated to attain a refined SSM according to an adaptive fusion scheme. At last, based on the computed SSM, an energy minimization framework is designed to segment coherent human body and its most distinctive acting part in a video spatially and temporally.

After detecting the actor and the actor's acting part, together with the full image, we integrate these three types of semantic cues into our advanced multi-stream network (MS-Net). In this way, the segment-level local information, the image-level spatial global scene information, and the video-level spatiotemporal global information can be well fused to support each other. Due to the detected actor and its acting part, our network is able to localize and recognize the action, and is effective to answer *whether* there is an action, if so *where* is the action and performed by *which* actor (see Figure 2). In addition, outliers, like the background clutters outside these captured regions are significantly suppressed. The scene content is also important as it supplies location-independent cues. E.g., differentiating "tennis game" and "badminton game" demands high-level representations to simulate global scene statistics.

This work makes three main contributions:

- We demonstrate that exploiting spatiotemporal consistent human-related semantics, i.e., the actor and its acting part, is effective to improve the performance of video-based action recognition, especially for the action which plays by one actor.
- To capture the actor and its acting part, we present a novel STS-VOS method to detect objects in videos. The complementary object signatures from different object segmentation methods are explicitly used to compute spatiotemporal saliency maps to guide salient video object segmentation via an energy function for the first time.
- We design an advanced semantic cues enhanced spatiotemporal MS-Net to improve the TS-Net in three aspects: exploiting semantics to help formulate valuable input modalities; constructing a multi-scale strategy to extract features from these modalities in different scales; selecting suitable deep network for the spatial stream and the temporal stream separately.
- We conduct experiments on four well-known benchmarks: JHMDB, HMDB51, UCF-Sports and UCF101, on which our method achieves superior performance.

II. Related Work

Numerous studies have been carried out on action recognition in videos. It is beyond the scope of this paper to introduce all of them. Hence, we pay attention to the related works in three categories: 1) semantic cues extraction in videos

in terms of saliency information, 2) TS-Net related algorithms for video processing, and 3) exploiting semantics for TS-Net related methods to recognize human actions in videos.

A. Semantic Cues Extraction in Videos

Semantic cues, especially human and the motion salient regions of human, are very useful for video-based action recognition. Since motion is the most significant cue, a lot of work has considered this feature. 1) Salient motion detection. Wixson [30] proposed to extract salient motion based on the intermediate-stage vision integration of optical flow. 2) Combining appearance and motion saliency cues. Zhai and Shah [31] fused spatial and temporal saliency maps dynamically to construct a spatiotemporal saliency map. Zhong *et al*. [32] presented a video saliency detection algorithm to extract attended regions. One attention model is constructed by fusing spatial and temporal saliency maps. 3) Spatiotemporal consistency optimization. The aforementioned two types of techniques treat video frames one-by-one, without considering the inherent spatiotemporal consistent feature of the video saliency maps. Wang *et al*. [34] designed an energy function to segment object spatiotemporally consistent by using the obtained SSM. Tu *et al*. [1] proposed a spatiotemporal saliency method to extract salient objects in videos. By fusing two distinctive saliency maps, which are calculated from two types of complementary object signatures, a refined spatiotemporal video saliency maps are obtained. This method is robust under multiple complex conditions. How to combine different semantics optimally is a challenging but unsolved problem [14], more contributions need to be conducted in this domain.

B. TS-Net Related Algorithms for Video Processing

Since Simonyan and Zisserman [5] proposed a TS-Net to extract both appearance and motion features for video-based action recognition, the TS-Net has been extended and modified to handle many video tasks. To overcome the drawback of TS-Net which is unable to model long-range temporal structure, Wang *et al.* [4] proposed a temporal segment network (TSN) to extract long-term temporal feature for action recognition. Wu *et al.* [25] applied the Long Short Term Memory (LSTM) [36] to explore long-term temporal dynamics in the visual channel. Varol *et al.* [15] applied neural networks with long-term temporal convolutions to learn video representations. Fernando *et al.* [16] proposed a rank pooling method, which learns a pooling function via ranking machines, to capture the video-wide temporal evolution of a video.

How to optimally fuse different streams is another interesting topic. Wu *et al.* [25] presented a multi-stream framework to capture multimodal features for video classification. A multi-stream multi-class fusion approach was designed to learn the optimal fusion weights of each class with regards the class-specific preferences. Yang *et al.* [26] presented a multilayer and multimodal fusion architecture for classification in videos. They employed four complementary modalities: 2D-CNN on a single RGB frame and flow image, and 3D-CNN on a short clip of RGB frames and flow images. Feichtenhofer *et al.* [6] proposed a spatiotemporal architecture for TS-Net to better extract features for action recognition by proposing a 3D convolutional fusion followed by 3D pooling at the last convolutional layer.

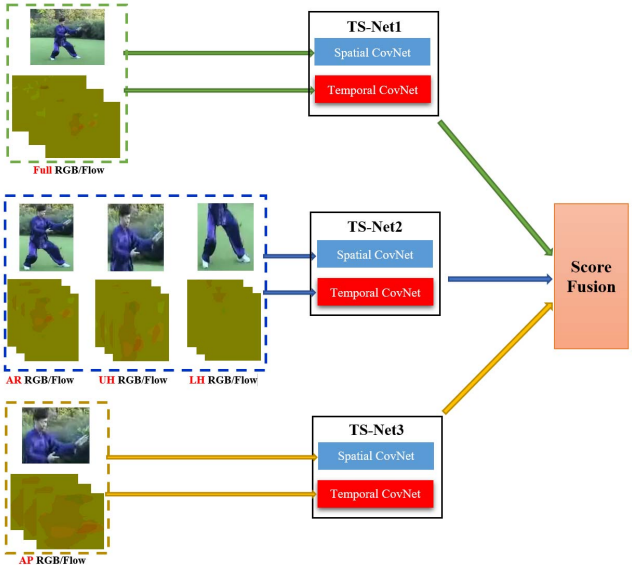

Fig. 1. The proposed MS-Net: consisting of three TS-Nets (i.e. TS-Net1, TS-Net2 and TS-Net3), with respect to the original 2 modalities (i.e. Full RGB and Flow) and our extra extracted semantics derived 8 modalities (i.e. *AR* RGB and Flow, *UH* RGB and Flow, *LH* RGB and Flow, *AP* RGB and Flow).

C. Exploiting Semantics for TS-Net to Recognize Human Action

Cheron *et al*. [8] applied human pose to learn features for video action recognition under the TS-Net. However, pose-estimators should be avoided for action recognition till pose estimation becomes accurate enough [29]. Singh *et al*. [19] exploited a MSB-RNN for fine-grained action detection in videos. Based on the detected bounding box of the person according to a simple state-based tracker, two person-centric streams are obtained in spatial and temporal domains. The tracker performs poorly in capturing humans in videos. Tu *et al*. [3] improved the B-RPCA [37] to capture the actor to assist action recognition, however, the IB-RPCA method of [3] is not good at detecting human if the background is moving or there are multiple moving objects. Gkioxari and Malik [21] employed selective search to produce approximately 2K regions in each frame, and remove the proposals that are void of motion according to a motion salient measure. However, in this method, not only some unnecessary regions are utilized but also the necessary human body cannot be precisely localized. To address the overfitting issue due to insufficient training data, Shu *et al.* [17] proposed a novel architecture of DTNs which is able to transfer cross-domain information from text to image. A semantic-intensive image feature representation is formed to enhance the performance of image classification tasks.

The most related work to us is [7], which integrates various semantic cues (i.e. object proposals) detected by Faster R-CNN into the TS-Net for action recognition. A late fusion (Sum fusion) was adopted to combine the scores computed

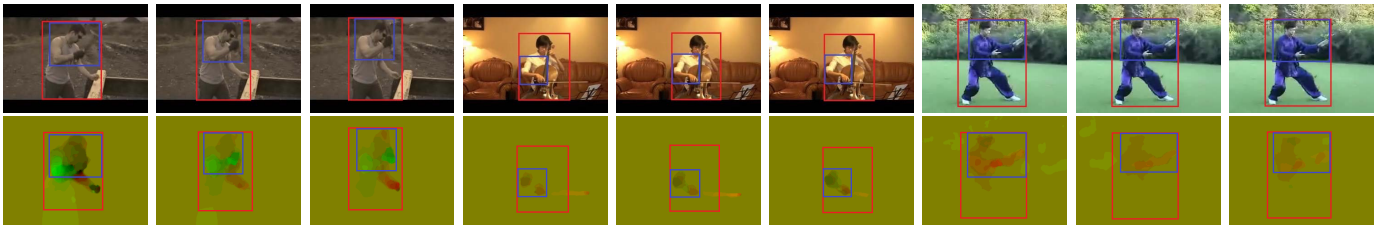

Fig. 2. The detected actor and the actor's acting part on UCF101 by the proposed STS-VOS method. First row and second row are the extracted two types of semantic cues on the RGB image and the flow image respectively.

from these semantic channels for final action prediction. This method has some drawbacks. First, the Faster R-CNN is not good at detecting spatiotemporal consistent object (i.e. object-tube) in videos. Second, too many irrelevant objects or background outliers are captured, and most of them are directly treated as semantic cues to the network. Third, the architecture of the TS-Net is too simple to extract good features when the videos contain challenging conditions. Finally, they ignore the fact that motion salient acting part plays a significant role for action recognition. To handle these defects, we design an effective semantics involved MS-Net in this work.

III. The Proposed Approach

In this section, we describe the proposed method in detail. First, we present a STS-VOS approach to extract semantics: the actor and its acting part (see Figure 2). Second, we design an advanced *MS-Net* which contains *three TS-Nets* to learn features for action recognition. In particular, other than two modalities, i.e. the full RBG image and optical flow [43], [44] (also called scene), those were employed in the original TS-Net, we exploit additional 8 modalities: 1) our detected actor-region (AR) RBG image, 2) upper half (UH) of AR RBG image, 3) lower half (LH) of AR RBG image, 4) acting part (AP) RGB image, 5) AR optical flow, 6) UH of AR optical flow, 7) LH of AR optical flow, and 8) AP optical flow. Besides, we extend these modalities to multi-scale to learn more representative deep descriptors (see Figure 4). Finally, a spatiotemporal convolutional fusion (ST-CF) strategy [6] is introduced to merge distinctive features of spatial stream and temporal stream at the last convolutional layer to full use the spatiotemporal video information (see Figure 5). Figure 1 outlines the general framework schematically.

A. Semantic Cues Detection According to STS-VOS

A video usually shows the movements and interactions of objects over time in some scenarios. To recognize human actions in videos, we decompose a video into three parts – scene, actor and the actor's acting part, to capture visual features. How to effectively extract the human and its acting part is a significantly challenging issue in video-based action recognition. A STS-VOS method is proposed to treat with this problem. The key idea of our STS-VOS method contains two main phases: (1) Introducing the idea of [1] to calculate SSM, which uses some object signatures that can be captured by any effective video-based object segmentation algorithms, allowing mutually complementary channels of saliency maps to be computed, combing these saliency maps to obtain a refined SSM via an adaptive fusion; (2) The SSM is applied to guide the segmentation of both the actor and its acting part in terms of an our designed energy function.

1) Extracting Object Signatures: We introduce three video object segmentation methods to obtain three types of object signatures: the appearance-dominated IB-RPCA method [1], the motion-dominated IFOS approach [1], and the CNN-related object flow (ObjFlow) algorithm [40]. To enhance the performance on detecting object signatures, we improve them to boost their effectiveness for our purpose.

Object Signature 1 (Object Extraction by IB-RPCA) IB-RPCA: The B-RPCA method [37] is able to detect foreground moving objects in complicated scenes with various challenges simultaneously, e.g., background motions, illumination changes, and camouflage. The B-RPCA method contains three major steps: 1) First-pass RPCA; 2) Motion Saliency Estimation (MSE); 3) Second-pass RPCA. To further enhance the performance of the IB-RPCA method of [1], we modify the second step. To suppress the wrongly identified non-stationary motions, we add the velocity angle constraint of [3] to the motion direction consistency measure. In addition, we use the velocity angle constraint to calculate the trajectory. Unlike [1] which sets the trajectory length to 5 frames, we try to track a point as long as possible, however, if the angle difference between frame j and its adjacent frame $j + 1$ is larger than 45° (we set it experimentally), we stop to compute the trajectory. Because, normally, the object points move temporally smoothing and consistent, this measure is useful to reduce the trajectory of the background moving noise which has random motion direction.

Object Signature 2 (Object Extraction by IFOS): The FOS [38] can automatically segment video objects efficiently in unconstrained setting, including fast moving background, non-rigid deformations, objects with arbitrary appearance and motion types. This algorithm contains two primary steps: initial foreground estimation and foreground-background labelling refinement. Tu *et al.* [1] designed an improved LMBD (ILMBD) method [41] to compute motion boundaries based on optical flow to boost the performance of object detection in the first step. We employ the most recently deep-leaning based optical flow method FlowNet 2.0 [42], which is very efficient and effective, to compute optical flow.

Object Signature 3 (Object Extraction by ObjFlow): Tsai *et al.* [40] proposed a method to regard video segmentation and optical flow computation in videos jointly. For video segmentation, they exploited a principled, multi-scale and

spatiotemporal graphical model. Both the CNN deep features [45] and the appearance features are used. This method can tackle various challenges, like cluttered backgrounds, objects with deforming shapes and fast moving.

2) Computing Saliency Maps: After attaining three types of foreground object signatures, we introduce the foreground connectivity [46] to each channel to compute saliency maps with two modifications:

Improvement 1 (Labeling Foreground Superpixels): Srivatsa and Babu [46] computed the objectness maps according to the objectness proposals which are extracted by the method of BING [47]. Then, the foreground can be roughly obtained by thresholding the objectness maps, and superpixels that are part of the foreground are accordingly captured. Since the detected foreground is coarse, the captured foreground superpixels are not precise. In this work, much more accurate foreground superpixels can be estimated based on our segmented object signatures as follow:

$$FP_i \leftarrow P_{O_i} > \gamma \cdot P_{N_i} \tag{1}$$

where $P_i \in P$ represents a superpixel. P_{N_i} denotes the number of pixels belong to a superpixel region R. P_{O_i} is the number of overlapped pixels between a superpixel region of P_i and our detected foreground object signatures. γ is a constant parameter $\in [0, 1]$. $FP_i \in FP$ denotes a superpixel is distinguished as the foreground. Generally, if more than half of a superpixel locates on our detected foregrounds, the superpixel will be labeled as the foreground. We set $\gamma = 0.55$ empirically same as [1]. The SLIC algorithm [48] is used to abstract each video frame into superpixels.

Improvement 2 (Foreground Connectivity): A robust saliency measure called foreground connectivity is utilized to assign saliency values based on superpixel connectivity to the captured foreground. An undirected weighted graph is constructed by using superpixels as nodes. All adjacent superpixels in the image are connected by an edge and the edge weight is set as the Euclidean distance between their mean CIE-Lab values. The geodesic distance between any two superpixels $d_{geo}(P_i, P_j)$ is computed as the accumulated edge weights along their shortest distance on the graph [49]:

$$d_{geo}(P_i, P_j) = \min_{P_1=P_i, P_2, \ldots, P_k=P_j} \sum_{n=1}^{k-1} d(P_n, P_{n+1}) \tag{2}$$

The foreground connectivity of a superpixel P_i is defined:

$$F_{GC}(P_i) = \frac{\sum_{k=1}^{N} d(P_i, P_k) \times \delta(P_k)}{\sum_{k=1}^{N} d(P, P_k) \times (1 - \delta(P_k))} \tag{3}$$

where $\delta(\cdot)$ is 1 if a superpixel is identified as foreground superpixel according to Eq. (1), and N is the total number of superpixels. Same as [46], we also take the reciprocal of F_{GC} and apply it as the foreground weights w^{fg}:

$$w^{fg}(P_i) = 1/F_{GC} \tag{4}$$

Eq. (4) calculates foreground weights for all superpixels. Normally, the foreground weight of a superpixel should be assigned to zero if it is not distinguished as foreground:

$$w^{fg}(P_i) = 0, \quad \forall P_i \notin FP \tag{5}$$

At last, we adopt the saliency optimization method of [46], which incorporates our foreground weights with the background measure of [49], to compute the final saliency maps. Accordingly, a motion-dominated video saliency maps from IFOS (we call it IFOS saliency maps), an appearance-dominated saliency maps from IB-RPCA (we call it IB-RPCA saliency maps), and a deep learning relevant saliency maps (we call it ObjFlow saliency maps) are obtained.

3) Fusing Saliency Maps: The computed three types of saliency maps complement each other. However, simply combining them, e.g., taking the product [66] or average [67], does not necessarily produce a better video saliency maps. We exploit an adaptive fusion method which is implemented in two steps (using the IFOS saliency maps and the IB-RPCA saliency maps for explanation):

Step 1 (Segmenting Object Proposals): We apply adaptive thresholding to roughly segment foreground objects in the IFOS and IB-RPCA video saliency maps in each frame according to the following measure:

$$\begin{aligned} FG_m &\leftarrow S_M > graythresh(S_M) \\ FG_a &\leftarrow S_A > graythresh(S_A) \end{aligned} \tag{6}$$

where S_M represents the IFOS saliency map and S_A denotes the IB-RPCA saliency map, and *graythresh* is the Matlab built-in function [50].

Step 2 (Saliency Maps Fusion): Based on FG_m and FG_a, we can get box regions for each of them, i.e., $mB = \{mB_1, \ldots, mB_M\}$ and $aB = \{aB_1, \ldots, aB_N\}$, and look for overlapped regions between them. Firstly, for these detected box regions in every frame, if any of its intersection-over-union (IOU) score is higher than a threshold τ (we set $\tau = 0.75$), we select the larger region box between them and label them as:

$$LB = B_{1L} \cup \ldots B_{kL} \cup \ldots B_{KL}, \quad (K \leq \min(M, N)) \tag{7}$$

where $B_{kL} = max(aB_n, mB_m)$. Secondly, we find other foreground pixels that are overlapped in other regions:

$$LP \leftarrow (FG_m \cdot FG_a) > 0 \tag{8}$$

The final foreground pixels are labeled as:

$$FG \leftarrow LB \cup LP \tag{9}$$

We fuse S_M and S_A guided by the foreground pixels FG to obtain a spatiotemporal saliency map:

$$STSacy(i) = \begin{cases} S_M(i) & \text{if } i \in FG(i) \& S_M(i) > \eta \\ \max(S_M(i), S_A(i)) & \text{if } i \in FG(i) \& S_M(i) \leq \eta \\ S_M(i) \cdot S_A(i) & \text{if } i \notin FG(i) \end{cases} \tag{10}$$

where i is the pixel index, we set $\eta = 0.8$ to select the good quality S_M same as [33]. High quality motion saliency features are more reliable than appearance saliency features in a video since it is more robust to cluttered backgrounds.

In the same manner, we fuse the ObjFlow saliency maps with the computed SSM $STSacy$ of IFOS and IB-RPCA to get a final refined spatiotemporal video saliency maps. Refer [1] to see the idea of *fusing saliency maps* more.

4) Segmenting Video Objects: Similar to [34] and [38], we regard object segmentation in videos as a pixel labeling problem with two labels (foreground and background). The energy function of [34] is adopted to label $\mathbf{L}$ all the pixels:

$$F(\boldsymbol{L}) = \sum_{t,i} U_i^t(l_i^t) + \lambda_1 \sum_{t,i} A_i^t(l_i^t) + \lambda_2 \sum_{t,i} Ł_i^t(l_i^t) + \lambda_3 \sum_{(i,j)\in N_S} V_{i,j}^t(l_i^t, l_j^t) + \lambda_4 \sum_{(i,j)\in N_T} W_{i,j}^t(l_i^t, l_j^{t+1}) \tag{11}$$

where $(\Omega \times T \rightarrow \Re^3)$ is a video sequence. $x_i \in \mathbf{x}$ represents a point in the image domain Ω, i and j are the discrete index of the pixels. t is the frame index. $l_i^t \in \{0, 1\}$ denotes the binary label of pixel x_i. $\mathbf{L} = \{l_i^t\}_t$ of pixels from all frames denotes a segmentation of a video. U^t, A^t and $Ł^t$ are three unary terms, V^t and W^t are two pairwise terms. In particular, U^t is the saliency term which aims to evaluate the probability of a pixel is foreground or background based on saptio-temporal saliency maps calculated through our prior step. $\lambda_1, \lambda_2, \lambda_3$, and λ_4 are the scalar parameters weight these terms. We set them same as [34] in this work. For more detail, please refer to [34, eq. (7)].

To encourage temporal consistency of foreground objects, we exploit a temporal matching term to the saliency term U^t. The updated saliency term $\widehat{U}^t$ is defined as:

$$\widehat{U}^t = U^t + \lambda_m \Psi_m \tag{12}$$

The matching term Ψ_m is expressed as:

$$\Psi_m(\mathbf{x}) = \begin{cases} \dfrac{1}{N_t} \sum_{\mathbf{x}} \| I_t(\mathbf{x}) - I_{t+1}(\mathbf{x} + \mathbf{w}_{t,t+1}(\mathbf{x})) \|_1 & \text{if } l_i^t = 1 \\ 0 & \text{if } l_i^t = 0 \end{cases} \tag{13}$$

N_t is the number of pixels in the video frame t. I_t is the brightness of frame t. $\mathbf{w}_{t,t+1}$ represents the optical flow between two adjacent frames I_t and I_{t+1} computed by the method [42].

By minimizing our updated energy function with graph-cuts, object segmentation can be accurately achieved. To extract the acting part of the actor, we reset our video saliency maps *STSacy* with a self-adaptive threshold:

$$STSacy(x_i) = 0, \forall STSacy(x_i) < median(STSacy(\mathbf{x})) \tag{14}$$

Finally, the MSM [3] is used to select the acting region where motion is most distinct among the proposals.

B. Semantics Enhanced Multi-Modality MS-Net

Based on our extracted semantic cues, an advanced semantic cues enhanced multi-modality spatiotemporal MS-Net is designed for video action recognition.

1) CNN Architecture: Network architecture plays a crucial role in the designing of high-performance CNN. The CNN_M [23] model, which is trained on the ImageNet dataset and used in TS-Net, is shallow (5 convolutional layers and 3 fully-connected layers). However, the concept of action in videos is more complicated than the concept of object in images, therefore, recognizing human action in videos requires high-level abstraction and inference. The deeper networks VGG16 (13 convolutional layers and 3 fully connected layers) is used to design our network, as it contains high modeling capacity and is able to extract more discriminative features at higher layers.

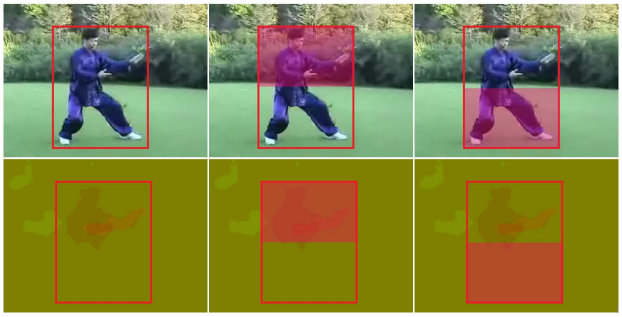

Fig. 3. The captured upper and lower half regions of the actor.

2) Input Modalities: To obtain more powerful abstractions, we exploit extra complementary modalities according to our captured human-related semantics. These modalities boost the discriminate capability of the proposed MS-Net to enable it to extract diverse static and dynamic cues.

a) Full RGB image: where the global spatial scene features are preserved. The TS-Net applies a random cropping to make a 224 × 224 patch and horizontal flipping to augment training samples. It is likely to choose regions near by the image center and training loss descends quickly resulting in overfitting. To address this issue, we introduce the corner cropping approach of [4], where 4 corners and 1 image center regions are cropped to augment the input to reduce overfitting.

b) Full optical flow: where the motion between video frames is described. To construct the flow image, optical flow is first calculated for each successive pair of frames according to [42]. The x-flow component and the y-flow component are rescaled to the range of [0, 255] by a linear transformation. Values smaller than 0 and larger than 255 are truncated. We adopt optical flow stacking with $L = 10$ as [5] and [6].

c) AR RBG image and flow image: where the appearance information of the actor and the motion information due to the movement of the actor can be learned. After detecting the human by our STS-VOS method, we extract the bounding box of the human in both the RGB image and its corresponding flow image. By utilizing the AR, not only background noise is reduced but also the extracted features strongly correspond to the location of the actor. In this way, some actions can be easier to recognize as location variation is removed. Besides, some actions are location dependent. Since some scene shots are captured by static cameras, the actions almost take place at the same image position, e.g., "eat" and "drink" in HMDB51 [57], "GolfSwing" and "Haircut" in UCF101 [56], etc.

d) UH/LH regions of AR RBG image and flow image: where appearance and motion features in half region of the human body are abstracted (see Figure 3). In contrast to [52] which used the left/right/upper/lower half regions for object detection, and Peng and Schmid [53] applied upper/lower half

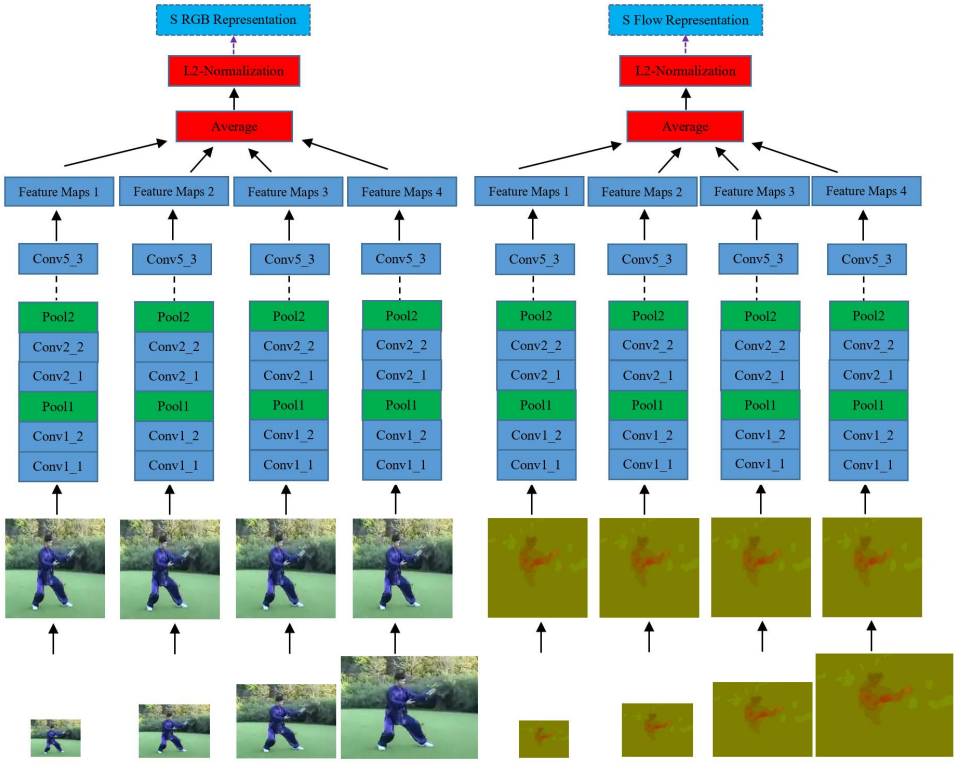

Fig. 4. Using **scene** (S) for explanation (on the TS-Net1) of multi-scale based data augmentation operation. Fusing multi-scale feature representations of one input semantic cue to form the final feature representation in the spatial-stream (e.g., S RGB representation) and the temporal-stream (e.g., S Flow representation) at the last convolutional layer (Conv5_3) of VGG16 model.

of the region proposal network (RPN) proposals for action detection, we utilize UH/LH regions to recognize human actions. Because for human actions, no matter the action is complicated or simple, most of them are horizontal body symmetry. Two benefits can be obtained with this operation: First, the learned features are more robust with respect to occlusions; Second, it is helpful to recognize the actions that body part characteristics are dominated. E.g., "TaiChi", "GolfSwing" are easier to recognize with the upper half region of the actor, while "climb stairs" with respect to the lower half region of the actor.

e) AP RGB image and flow image: where the features related to action in both spatial and temporal domain are extracted. Features, which learned by this high-level semantic cue, have received little attention till now. But the AP directly supplies the action features and they are complementary to the actor and scene cues.

3) Modality in Multi-Scale: The above mentioned modalities are all in one single scale. In natural videos, action instances usually vary in scales. We introduce the multi-scale based data augmentation technique of [4] and [15] to learn more robust deep features by constructing 4 pyramid representations for each of the input semantic cue according to the scale sets: (1) for full image: $\{1/2, 1/\sqrt{2}, 1, \sqrt{2}\}$; (2) for actor-related regions and AP: $\{1/\sqrt{2}, 1, \sqrt{2}, 2\}$. After that, the multi-scale representations of each modality will be rescaled to the same size for training. In the end, the four-scale feature maps of one modality are averaged to produce the final feature representation, and then L2-normalize it [24]. Table VIII shows the results of different methods to combine the multi-scale representations. "Average" performs best among them. Consequently, we apply "Average" fusion in the following experiments. Figure 4 shows the L2-normalized multi-scale merged feature representation (at layer Conv5_3) of scene of the spatial-stream and the temporal-stream.

4) Architecture of Our Network: Feichtenhofer *et al.* [6] proposed a ST-CF method upon the TS-Net with two main improvements: (1) rather than fusing softmax scores, they fuse the two streams features at the last convolutional layer by injecting a convolution fusion layer; (2) Substituting the typical 2D convolution and pooling with 3D convolution followed by 3D pooling to fuse features spatiotemporally.

Unlike the baseline work of [6] which just needs to fuse two modalities, our inputs are much more complicated. With the 3 types of modalities, i.e. scene, the actor-related regions (ARs) – AR, UH/LH regions of the actor, and AP, we construct a MS-Net that consists of three TS-Nets. In particular, one scene related TS-Net (TS-Net1), one ARs based TS-Net (TS-Net2), and one AP based TS-Net (TS-Net3). As shown in Figure 5, for the ARs based TS-Net2, the feature maps of 3 RGB ARs modalities – AR, UH/LH regions of actor in the spatial stream are weighted summed at the last convolutional layer. The weights are set as – AR: UH: LH = 1: 0.5: 0.5. The feature maps of 3 optical flow ARs modalities in the temporal stream are conducted in the same manner. We compare different weights setting in Table VI. We fuse the prediction scores of the three TS-Nets by averaging as [4] and [9] for final action recognition. For each of the three TS-Nets, like the

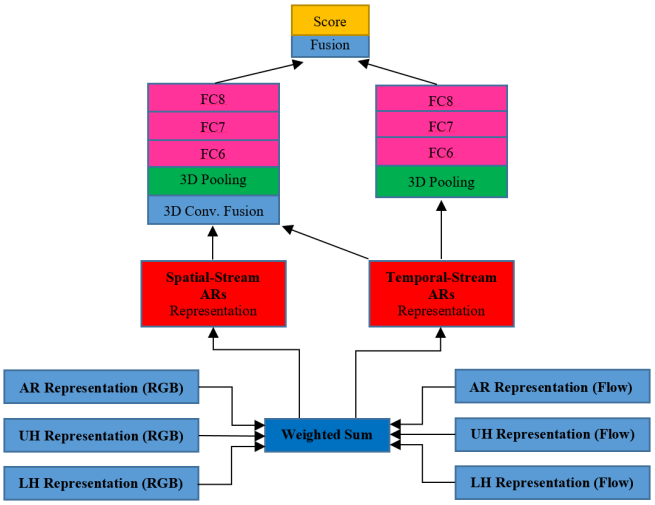

Fig. 5. Using the TS-Net2 for explanation (with the VGG16 architecture). Illustrations of 1) the combination of the feature maps of 3 actor-related regions (ARs: AR, UH, LH) at the last convolutional layer (Conv5_3) and, 2) the fusion of spatial stream and temporal stream via the ST-CF operation.

depiction of TS-Net2 in Figure 5, we adopt the ST-CF to merge the appearance feature and motion feature spatiotemporally. Figure 1 shows the architecture of our framework.

5) Implementation Details: All the input modalities are resized to 224 × 224. We take the strategies of [6] to train three TS-Nets. For TS-Net1, its spatial network and temporal network are trained on the full RGB image and optical flow. For TS-Net2, its spatial network and temporal network are trained on the captured AR RBG image and AR flow image. For TS-Net3, we use TS-Net2. Therefore, the modalities of AR, UH, LH, and AP are all using the architecture of TS-Net2. The pre-trained VGG16 models of [6] are introduced as the initialization, and we fine-tune the spatial and temporal networks with the following modifications: (1) the learning rate is set to 10^{-4} and 5×10^{-3} respectively for the spatial network and the temporal network, and decreased by a factor of 10 as soon as the validation accuracy saturates; (2) the dropout ratio is set as 0.90 for the spatial network and 0.85 for the temporal network. For testing, we use 25 frames randomly sampled from each video.

IV. Experiments

In this section, to evaluate the performance of our proposed MS-Net, four popular benchmark datasets are used for experimenting and analyzing: JHMDB [55], HMDB51 [57], UCF-Sports [54], and UCF101 [56]. Extensive experiments are conducted to test the effectiveness of our model which is based on the VGG16 architecture in five aspects: (1) the effect of the detected semantic visual cues according to our STS-VOS method; (2) the performance of the multiple modalities in multi-scale; (3) the influence of the spatiotemporal fusion strategy; (4) the effect of deeper architecture; (5) comparing our method with the state of the art algorithms.

A. Datasets and Evaluation Protocols

JHMDB is a subset of HMDB51, with 928 video clips of 21 different actions. Each action contains 36 to 55 video clips. One clip includes the number of frames range from 15 to 40 with frame size 320 × 240.

HMDB51 consists of 6766 action videos which are divided into 51 action categories. The videos are collected from a wide range of sources, including movies and online videos.

UCF-Sports includes 150 video clips with 10 different actions. The videos are captured in cluttered, dynamic environments, and each of them corresponds to one action.

UCF101 contains 13320 videos categorized into 101 action classes. For one class, there are more than 100 video clips, and each video clip with an average length of 180 frames. They cover a large range of activities such as sports and human-object interaction. It is a challenging dataset as the captured videos change significantly in scale, illumination, background and camera motion.

For JHMDB and UCF-Sports, we follow the suggested evaluation protocol of [55] and [54], and report the average accuracy over the three splits. Besides, to evaluate the performance of our extracted semantic cues – AR, UH, LH, and AP, we compute the mAP at the video level as in [3]. For HMDB51 and UCF101, we utilize the split 1 for experimental analysis and report final accuracy averaged over the three splits.

B. Evaluation the Effect of Our STS-VOS Method

Figure 6 shows the precision-recall (PR) curves and the mean absolute error (MAE) [68] of the estimated SSM of 4 different approaches – the baseline salient video object detection method [1], the saliency maps combination techniques, i.e. product, average, and our adaptive fusion method, on three benchmark datasets. Our method outperforms the other three approaches on all these datasets. Comparing to the baseline method [1], the MAE gain of us reaches to 20.18%, 5.64%, and 12.58% on the datasets SegTrack [69], SegTrack V2 [70], and Ten-Video-Clips [71] respectively. It demonstrates that improving the two object segmentation methods in [1], adding other valid complementary object signatures are effective to further boost the performance of detecting salient objects in videos. In addition, the results also reveal that our adaptive fusion strategy is useful, which performs much better than the classical product and average fusion schemes.

Figure 7 shows the quantitative comparisons of actor segmentation on the UCF-Sports dataset. The measure Accuracy-Overlap Thresholds of [29] is used for evaluation. The IOU scores for the segmentations are computed and the overlap threshold varies from 0.1 to 0.6. Our method performs better than all of them at every threshold, which demonstrating that the proposed STS-VOS approach is able to integrate the advantages of the three segmentation techniques and extract the actors accross the whole video robustly and accurately.

C. Evaluation the Effect of Our Extracted Semantic Cues

In this subsection, to evaluate the effect of the captured human-related semantic cues by our STS-VOS approach, we perform the following studies: whether our exploited STS-VOS method is effective to detect the actor, and the captured actor is better than the objective proposals extracted by other

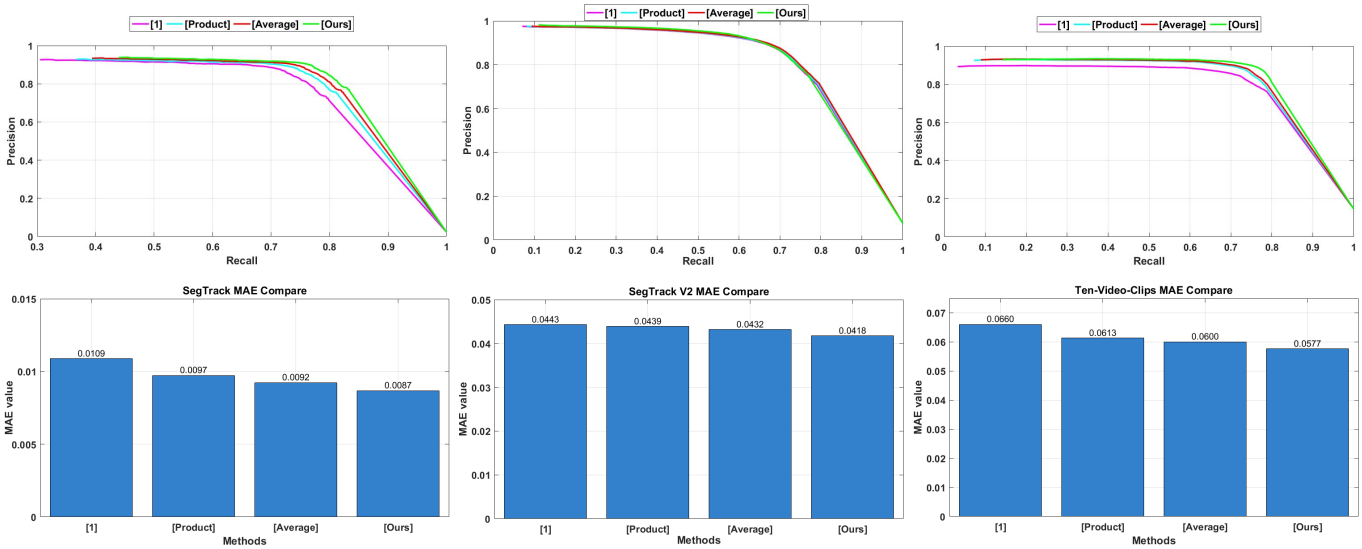

Fig. 6. Comparison of PR curves (top row) and MAE (bottom row) on datasets SegTrack (left column), SegTrack v2 (middle column) and Ten-Video-Clips (right column).

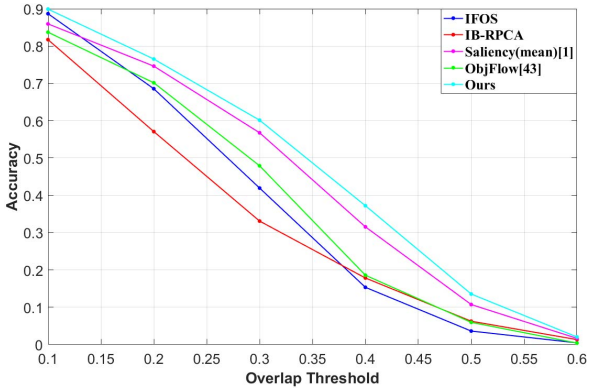

Fig. 7. Comparison of actor segmentation on the UCF-Sports.

object detection methods for action recognition; whether the UH and LH regions of the actor as well as the AP are beneficial for action recognition; whether these semantic cues are complementary and would boost the performance of recognizing human actions when combing them all. To conduct these investigations, we utilize the TS-Net with VGG16 architecture as the baseline, and implement three kinds of experiments on JHMDB, UCF-Sports and UCF101 datasets for analysis. In particular, as the actor bounding box of JHMDB and UCF-Sports are annotated, we also use them for the purpose of action detection evaluation.

First, we study the effect of the human detection quality on action recognition by comparing different object detection methods. Six excellent video-based methods, i.e. one state-based video tracker (suggested in [19]) – extended Lucas-Kanade (ELK) [39], five video segmentation methods – FOS and B-RPCA as well as their variants IFOS and IB-RPCA, and the CNN feature integrated ObjFlow, and two currently popular used image-based deep learning approaches – R-CNN and Faster R-CNN, are selected for comparison. The human region of interests (ROIs) are computed by directly running the source code supplied by the authors. We only use the detected objects (bounding boxes) as the input. To deal with multiple object ROIs captured by R-CNN and Faster R-CNN, we adopt the multiple instance learning technique of [7] to integrate the scores by a max operation. Table I shows that the quality of the detected ROIs significantly influence the accuracy of action recognition. Our STS-VOS method perform best to extract AR in videos to supply features for action recognition. The accuracy of action recognition reaches to 62.11% by using the AR detected by our STS-VOS, which is 2.93% higher than employing IFOS (IFOS outperforms FOS, B-RPCA and IB-RPCA), is 3.99% better than using ObjFlow, and is 6.22% more precise than utilizing ELK. The results reveal that exploiting advanced video-based human detector is a promising way to enhance the performance of video-based action recognition. In addition, the results of the two deep learning based image object detection methods, i.e., R-CNN and Faster R-CNN are comparable to the five hand-crafted video segmentation methods, but are much worse than our STS-VOS approach. This is because the proposed STS-VOS method is able to combine valid complementary object signatures of different sources, leading to it is able to detect the actor-tube in videos accurately under various complex cases, and also some irrelevant objects are greatly compressed where these irrelevant objects are not useful for action recognition.

There is no dataset supplies the ground truth of AP, thus we cannot directly evaluate the performance of AP detection. Instead, we are able to indirectly evaluate the influence of AP detection on recognizing human actions. Table II shows the accuracy of action recognition by setting different thresholds to extract AP. When setting the threshold to the adaptive *median* value (see Eq. (14)), best result is obtained. It displays that the quality of the detected AP affects the performance of action recognition explicitly.

Second, we investigate the influence of UH region (A-Upper) and LH region (A-Lower) of the actor. We show the per-class mAP of the TS-Net with VGG16 on JHMDB

TABLE I

ACTION RECOGNITION PERFORMANCES OF THE TS-NET USING VGG16 ACROSS THE THREE SPLITS ON JHMDB DATASET WITH DIFFERENT SOURCES OF ACTOR ROIS CAPTURED BY DIFFERENT METHODS. GT REPRESENTS THE GROUND TRUTH ACTOR ANNOTATION

	ELK [39]	B-RPCA [37]	IB-RPCA	FOS [38]	IFOS	ObjFlow [40]	R-CNN [10]	Faster R-CNN [11]	**Ours**	GT
Acc.(%)	55.89	56.81	58.93	58.02	59.18	58.12	58.45	58.90	**62.11**	63.87

TABLE II

ACTION RECOGNITION PERFORMANCE OF THE TS-NET3 ON SPLIT 1 OF UCF101 WITH DIFFERENT SETTINGS OF THRESHOLDS TO EXTRACT AP

	0.25	0.5	0.75	*mean*	*median*
Acc.(%)	90.53	91.01	90.67	90.79	**91.28**

in Table III and on UCF-Sports in Table IV. Among the three semantics, the *A* obtains the top performance in mAP, which indicates that the complete AR is critical importance for an action. The UH region and LH region have different performances in different datasets. For example, on JHMDB, the *A-Upper* achieves approximate result to *A* (63.3% vs 64.9%), and outperforms *A-Lower* by 4.9%. However, on the UCF-Sports dataset, the *A-Upper* performs worst, which is 5.22% lower than *A-Lower*. This reflects that these two datasets consist of different types of actions, i.e., for JHMDB the upper region of the action is most discriminative (daily actions), while for UCF-Sports the actions are dominated in the lower region (sport actions). The most accurate result can be obtained by combing the three semantics. *All* (the three actor-related regions) outperforms *A* by 2.5% on JHMDB and by 4.85% on UCF-Sports. Consequently, it is essential to extract the UH/LH regions of the actor and integrate them together with the actor for action recognition in videos as these semantic cues are complementary. To further validate this characteristic, we conduct an experiment on split one of the UCF101 dataset. As shown in Table V, the accuracy reaches to 91.97% ("*All*") when incorporating these human-related semantics.

For some actions, the half region of actor outperforms the whole actor. For instance, on both JHMDB and UCF-Sports datasets, *A-Upper* performs better than *A* for "Golf". Besides, *A-Upper* gains 4.4% for "Shootgun" and 3.1% for "Sit" over *A* on JHMDB. On the other side, *A-Lower* outperforms *A* by 2.9% for "Climbstairs" on JHMDB and by 23.78% for "Swing1" on UCF-Sports. Taking "Golf" for explanation, the most significant features to recognize this action are distributed in the UH region of the actor due to the intentional bodily movement of the handle and the upper torso.

Third, we evaluate the performance of the three types of cues, i.e., scene (S), actor-related regions (ARs), and the acting part (AP), which are corresponding to TS-Net1, TS-Net2, and TS-Net3 respectively. Table VII demonstrates the effectiveness of our idea that exploiting useful semantic cues are helpful for recognizing human action in a video. Especially, when comparing the second (S), third (ARs) and fourth (AP) columns to the fifth (S + ARs) and sixth (S + AP) columns, we can find that the results from the integrated semantic cues are better than using one single semantic cue. This can also be observed in the seventh (S + ARs + AP) column, when incorporating the captured two types of semantics with the original full image, the recognition accuracy is improved by 1.72% (91.35% to 93.07%). Because of the exploited ARs and AP supply strong complementary feature information to the scene. Besides, the features distributed at the ARs and the AP are closely relevant to human actions. Moreover, the learned features from these two types of semantic visual cues are robust with respect to occlusions, and the background noise is significantly suppressed in these regions. Therefore, all the TS-Net1, TS-Net2, and TS-Net3 are important and necessary, no one is more important than other two, and combing the three TS-Nets is easier to recognize human actions and is able to boost the recognition accuracy. The result of S (TS-Net1), which is sightly more accurate than AP (TS-Net3), reveals that scene context supplies very significant image-level spatial cue for action recognition. ARs (TS-Net2) performs better than both S (TS-Net1) and AP (TS-Net3), because an action is conducted by an actor who implements intentional bodily movement. One of the most crucial reasons why AP (TS-Net3) obtains the worst result among the three semantics is that action is scene context and actor related. The AP, which constrained to the bodily movement region, only supplies the body acting information locally. It can be easily confused if the actions have similar bodily movement, such as "EyeMakeup" and "Lipstick." However, the AP (TS-Net3) is quite useful if combined with other cues, e.g., S (TS-Net1) and ARs (TS-Net2). Last but not least, another one contribution due to the extracted two types of semantics is that, as shown in Figure 2, the *AR* is able to answer *who* is playing the action, and the *AP* is able to answer *where* is the action happening.

D. Evaluation of Modalities in Multi-Scale

In reality, an action can occur at arbitrary scales. We conduct multi-scale operation to increase robustness to scale changes. Table IX shows the effect of our multi-scale method. By formulating each modality into 4 scales, we are able to extract features in different scales. This is helpful to extract better local deep features and improving the recognition accuracy. Comparing to single-scale, the accuracy is boosted by 0.68% (93.07% vs 93.75%) on the UCF101 dataset with the application of our multi-scale strategy.

E. Evaluation of Spatiotemporal Fusion

How to fuse the spatial stream and the temporal stream spatially and temporally is still a difficult problem. We introduce the ST-CF method [6] to fuse the two steam features spatiotemporally, and evaluate whether this strategy is effective. Table X shows the results of ST-CF and the traditional linearly averaging based late fusion as [7] on split one of UCF101. Clearly, the ST-CF scheme obtains better result than the

TABLE III

PER-CLASS MAP ON SPLIT 1 OF JHMDB DATASET WITH RESPECT TO MULTIPLE ACTOR-RELATED REGIONS. *A* REPRESENTS THE DETECTED ACTOR, *A-Upper* AND *A-Lower* REPRESENT THE UPPER HALF REGION AND LOWER HALF REGION OF THE ACTOR

Regions	B.hair	Catch	Clap	C.stairs	Golf	Jump	Kickball	Pick	Pour	Pullup	Push	Run	S.ball	S.bow	S.gun	Sit	Stand	S.baseball	Throw	Walk	Wave	mAP
A	77.3	52.8	61.7	60.6	87.9	46.2	51.4	57.5	85.7	97.2	85.1	54.6	41.8	80.5	61.1	72.2	73.7	64.1	10.2	85.0	56.1	64.9
A-Upper	**83.1**	51.2	58.5	56.2	82.7	53.5	48.8	55.2	81.6	96.8	81.3	51.4	39.5	74.8	**65.5**	**75.3**	64.9	66.7	9.4	80.8	51.6	63.3
A-Lower	65.6	49.8	31.2	**63.5**	84.2	49.1	50.3	52.7	78.2	96.1	83.0	52.9	33.6	70.1	51.2	66.7	66.1	51.2	7.8	75.3	47.4	58.4
All(ARs)	82.5	51.3	**64.6**	60.9	**90.4**	**55.7**	51.0	57.1	**86.3**	97.0	**87.5**	**72.4**	**43.0**	79.8	70.7	74.4	73.8	**66.4**	**10.5**	84.7	55.9	**67.4**

TABLE IV

PER-CLASS MAP ON SPLIT 1 OF UCF-SPORTS DATASET WITH RESPECT TO MULTIPLE ACTOR-RELATED REGIONS

Regions	Diving	Golf	Kicking	Lifting	RidingHorse	Run	SkateBoarding	Swing1	Swing2	Walk	mAP
A	100.0	63.89	100.0	100.0	100.0	63.89	87.67	63.89	87.67	100.0	86.70
A-Upper	100.0	**87.67**	52.50	100.0	100.0	25.00	63.89	63.89	87.67	100.0	78.06
A-Lower	100.0	52.50	52.50	100.0	100.0	52.50	87.67	**87.67**	**100.0**	100.0	83.28
All(ARs)	100.0	87.67	100.0	100.0	100.0	52.50	87.67	87.67	100.0	100.0	**91.55**

TABLE V

ACCURACY OF THE TS-NET2 ON SPLIT 1 OF UCF101 WITH RESPECT TO MULTIPLE ACTOR-RELATED REGIONS

	A	A-Upper	A-Lower	All(ARs)
Acc.(%)	91.46	90.26	91.13	91.97

TABLE VI

ACCURACY OF THE TS-NET2 ON SPLIT 1 OF UCF101 WITH RESPECT TO MULTIPLE ACTOR-RELATED REGIONS WHICH ARE COMBINED WITH DIFFERENT WEIGHTS – AR: UH: LH

All(ARs)	1:1/4:1/4	1:1/3:1/3	1:1/2:1/2	1:1:1	1/2:1:1
Acc.(%)	91.86	91.90	**91.97**	91.63	91.50

TABLE VII

ACCURACY OF THE MS-NET ON SPLIT 1 OF UCF101 WITH THREE SEMANTIC CUES BASED INPUT MULTI-MODALITIES. S REPRESENTS THE SCENE – FULL RGB IMAGE AND FULL FLOW IMAGE (TS-NET1), ARS DENOTES THE ACTOR-RELATED REGIONS (TS-NET2), AP REPRESENTS THE ACTING PART OF THE ACTOR (TS-NET3)

	S	ARs	AP	S+ARs	S+AP	All (S+ARs+AP)
Acc.(%)	91.35	91.97	91.28	92.75	92.66	93.07

TABLE VIII

ACCURACY OF THE MS-NET ON SPLIT 1 OF UCF101 WITH RESPECT TO COMBINE THE FEATURE REPRESENTATIONS OF EACH MODALITY IN MULTI-SCALE WITH DIFFERENT METHODS

Multi-scale Combination (S)	Max	Concatenation	Average
Acc.(%)	90.07	91.19	91.35
Multi-scale Combination (ARs)	Max	Concatenation	Average
Acc.(%)	90.21	91.72	91.97
Multi-scale Combination (AP)	Max	Concatenation	Average
Acc.(%)	90.02	91.03	91.28

traditional fusion (93.91% vs 93.66%), because, the ST-CF is able to learn correspondences between highly abstract spatial stream features and temporal stream features. Furthermore, it can incorporate the features of spatial and temporal steams over time.

TABLE IX

ACCURACY OF THE MS-NET WITH MULTI-MODALITIES ON SPLIT 1 OF UCF101 IN MULTI-SCALE

	Multi-Modalities	Multi-Modalities+Multi-scale
Acc.(%)	93.07	93.66

TABLE X

ACCURACY OF MS-NET WITH MULTI-MODALITIES ON SPLIT 1 OF UCF101 BY USING TWO FUSION STRATEGIES

	Late Fusion	ST-CF [6]
Acc.(%)	93.66	93.91

TABLE XI

ACCURACY OF OUR SPATIOTEMPORAL MS-NET ON SPLIT 1 OF UCF101 WITH DIFFERENT NETWORK ARCHITECTURES

Architectures	Acc.(%)
S:CNN_M, T:CNN_M	86.31
S:VGG16, T:CNN_M	90.77
S:VGG19, T:CNN_M	90.82
S:VGG16, T:VGG16	**93.91**
S:VGG19, T:VGG19	93.68
S:VGG19, T:VGG16	**93.92**

F. Evaluation of Deeper Architectures

Previous works [4], [6], [18] reveal that deeper network architectures are helpful for enhancing action recognition in videos. We evaluate the performance of usually employed shallow network architecture: CNN_M [23] and recent very deeper network architectures: VGG16 and VGG19 [51] in our MS-Net. As shown in Table XI, when using the deeper VGG16 to replace the CNN_M, much better results are obtained. The accuracy improvement reaches to 7.6% ("S:VGG16, T:VGG16" 93.91% vs "S:CNN_M, T:CNN_M" 86.31%). However, when comparing the results of "S:VGG19, T:VGG19" with "S:VGG19, T:VGG16", we can find that the deeper VGG19 performs worse than VGG16 in the temporal stream. Only slightly gain is attained when utilizing VGG19 to substitute VGG16 in the spatial stream ("S:VGG19, T:VGG16" 93.92% vs "S:VGG16, T:VGG16" 93.91%). This

TABLE XII
COMPARISON (ACC.(%)) WITH STATE OF THE ART METHODS ON THE UCF101 AND HMDB51 DATASETS (AVERAGE ON 3 SPLITS)

Methods	UCF101	HMDB51
iDT+FV [28]	85.9	57.2
iDT+HSV [61]	87.9	61.1
iDT+MIFS [62]	89.1	65.1
C3D [60]	85.2	-
Two-Stream Net [5] (VGG_M)	88.0	59.4
Two-Stream Net [59] (VGG16)	91.4	58.5
Two-Stream SR-CNNs [7] (VGG16)	92.6	-
Two-Stream Conv. Fusion [6] (VGG16)	92.5	65.4
Multilayer Multimodal Fusion [26] (VGG16, C3D)	91.6	61.8
ActionVLAD [63] (VGG16)	92.7	66.9
Ours (VGG16)	**93.9**	**67.2**
C3D+iDT [60]	90.4	-
TDD+FV [27]	90.3	63.2
Two-Stream Conv. Fusion+iDT [6]	93.5	69.2
ActionVLAD+iDT [63]	93.6	69.8
Ours (VGG16)+iDT	**94.8**	**70.4**
TSN (3 modalities, BN-Inception) [4]	94.2	69.4
ST-ResNet+iDT [64]	94.6	70.3

TABLE XIII
COMPARISON OF SPEED (FPS) FOR TESTING ON THE UCF101

Methods	TSN (3 modalities, BN-Inception) [4]	Ours
Speed (FPS)	5	4.3

reflects that the spatial stream and the temporal stream have different characteristic. Taking a further investigation on the relationship between the deeper network and the tow-stream related architectures is a future important research topic for the task of action recognition in videos.

G. Comparison With State of the Arts

We compare our proposed method with the state of the art video-based action recognition approaches over the three splits on UCF101 and HMDB51 in Table XII. The results of these methods are quoted from the original papers. The comparison is classified into 4 groups:

1) In the top group, the results of three famous hand-crafted algorithms [28], [61], [62] are shown. Especially, the improved Dense Trajectories (iDT) is one of the most successful hand-crafted features for action recognition currently. Our method outperforms these three approaches by at least 4.8% and 2.1% (comparing with the best result of [62]) on UCF101 and HMDB51 respectively.

2) In the middle top group, seven deep learning methods, which utilize a comparable baseline architecture to ours (VGG16), are selected for evaluation. In contrast to the original TS-Net [5], the results of our method are much better (93.9% vs 88.0% on UCF101, 67.2% vs 59.4% on HMDB51). Besides, compared to the most recent two-stream based method ActionVLAD [63] which uses the same VGG16 architecture as us, our results are 1.2% and 0.3% more accurate on UCF101 and HMDB51. Yang *et al.* [26] exploited a new framework to combine multiple layers and multiple modalities of CNNs. To model the long-term temporal information over an entire video, it proposed FC-RNN instead of utilizing the standard RNN structure. However, [26] performs not only much worse than us (91.6% vs 93.9% on UCF101, 61.8% vs 67.2% on HMDB51), but also worse than [4], [6], and [7], where these three methods do not employ LSTM, a variant of RNN, same as us.

3) In the middle bottom group, the deep learning methods that combine iDT are chosen for analysis. It is clear that integrating the iDT into our framework, the action recognition performance is further enhanced by 0.9% on UCF101 and 3.2% on HMDB51 respectively. Our method performs best among them.

4) In the bottom group, the TSN [4] and ST-ResNet [64], which based on ultra-deep architectures like BN-Inception [58] and ResNet [65], are chosen for comparison. Both TSN and ST-ResNet achieve high performance. However, it is interesting to note that together with the iDT, our VGG16 based method still performs better than these ultra-deep two-stream baselines. On UCF101, our method outperforms TSN by 0.6% and outperforms ST-ResNet by 0.2%. On HMDB51, our result is 1.0% and 0.1% more accurate than TSN and ST-ResNet. The superior performance of our method demonstrates that exploiting useful semantic cues, explicitly integrating multi-modalities in multi-scale, designing good fusion techniques, and using proper deep architectures are able to boost the performance of TS-Net on action recognition significantly.

The proposed method obtains a reasonable computational efficiency. To be specific, as shown in Table XIII, for testing, it can averagely process about 4.3 frames per second (FPS) on the UCF101 dataset with one Titan-X GPU. The proposed method performs slightly slower than TSN [4] but gets more accurate results. Which demonstrates that our method achieves a good balance between efficiency and accuracy.

V. CONCLUSION

In this paper, we proposed an advanced spatiotemporal MS-Net method, which stems from the TS-Net, for video-based action recognition. We designed a STS-VOS method to segment human-related semantic cues in videos. By fusing distinctive video saliency maps which are estimated from disparate object signatures, we can obtain a robust video SSM. An energy function is then constructed based on SSM to segment both the actor and the actor's acting part. These two semantic cues are leveraged as input modalities, together with other modalities (i.e. the scene, and the UH/LH regions of the actor), we formulate five types of RGB/Flow modalities (10 modalities) to help improving the performance of the TS-Net to recognize human actions. Besides, we extend these modalities into multi-scale to learn more robust deep features. Finally, a ST-CF method was adopted to fuse the feature maps of each type of RGB/Flow modalities spatiotemporally. In the future, we plan to exploit other semantic cues and propose more effective fusion approaches to combine the spatial and temporal streams for action recognition in videos.

A. Limitations

1) The proposed STS-VOS method is still not good at capturing the primary actor if the scene contains multiple

humans, the background large-scale objects have apparent movement, etc.

2) Extracting the human-related semantics, i.e. AR and AP, to construct new streams is more useful for the case where an action is conducted by one actor, while it is not very beneficial for recognizing the actions that there are multiple actors in the scene, or the scene includes multiple activities.

3) The whole architecture is not end-to-end as it needs to fuse the scores of the three TS-Nets, and it requires to preprocessing the RGB and flow images to obtain semantic cues (i.e., AR and AP) with our STS-VOS method separately.

B. Future Work

There are several ways in which we can address these limitations, and further enhance the practical application of our approach. First, to boost the performance of our STS-VOS method, we can integrate more effective object signatures which are computed by the newly deep CNN methods, and improving the fusion approach to combine different signatures. Second, to handle the scene contains multiple actors and multiple activities, we should design an adaptive approach to combine the three TS-Nets based on the number of detected object candidates according to ObjFlow. If an action video contains multiple actors or multiple activities, we should reduce the weights of TS-Net2 and TS-Net3. Third, incorporating the feature representations at the last convolutional layer and exploit a unified spatio-temporal loss function to optimize the architecture end-to-end is a promising way to improve the performance of action recognition. Third, establishing an effective dataset which contains the ground truth of AP, is a good work to evaluate the performance of AP detection, and find out the influence of AP detection on action recognition.

REFERENCES

[1] Z. Tu *et al.*, "Fusing disparate object signatures for salient object detection in video," *Pattern Recognit.*, vol. 72, pp. 285–299, Dec. 2017.

[2] H. Ye, Z. Wu, R.-W. Zhao, X. Wang, Y.-G. Jiang, and X. Xue, "Evaluating two-stream CNN for video classification," in *Proc. 5th ACM Int. Conf. Multimedia Retr.*, 2015, pp. 435–442.

[3] Z. Tu, J. Cao, Y. Li, and B. Li, "MSR-CNN: Applying motion salient region based descriptors for action recognition," in *Proc. 23rd Int. Conf. Pattern Recognit.*, Dec. 2016, pp. 3524–3529.

[4] L. Wang *et al.*, "Temporal segment networks: Towards good practices for deep action recognition," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 20–36.

[5] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 568–576.

[6] C. Feichtenhofer, A. Pinz, and A. Zisserman, "Convolutional two-stream network fusion for video action recognition," in *Proc. Comput. Vis. Pattern Recognit.*, 2016, pp. 1933–1941.

[7] Y. Wang, J. Song, L. Wang, O. Hilliges, and L. Gool, "Two-stream SR-CNNs for action recognition in videos," in *Proc. Brit. Mach. Vis. Conf.*, 2016, pp. 1–12.

[8] G. Chéron, I. Laptev, and C. Schmid, "P-CNN: Pose-based CNN features for action recognition," in *Proc. Int. Conf. Comput. Vis.*, 2015, pp. 3218–3226.

[9] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.

[10] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. Comput. Vis. Pattern Recognit.*, 2014, pp. 1–8.

[11] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 91–99.

[12] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei, "Large-scale video classification with convolutional neural networks," in *Proc. Comput. Vis. Pattern Recognit.*, 2014, pp. 1–8.

[13] L. Wang, Y. Qiao, X. Tang, and L. Van Gool, "Actionness estimation using hybrid fully convolutional networks," in *Proc. Comput. Vis. Pattern Recognit.*, 2016, pp. 1–10.

[14] S. Chang, G.-J. Qi, C. C. Aggarwal, J. Zhou, M. Wang, and T. S. Huang, "Factorized similarity learning in networks," in *Proc. Int. Conf. Data Mining*, Dec. 2014, pp. 60–69.

[15] G. Varol, I. Laptev, and C. Schmidy, "Long-term temporal convolutions for action recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, 2017. [Online]. Available: https://ieeexplore.ieee.org/document/7940083/, doi: 10.1109/TPAMI.2017.2712608.

[16] B. Fernando, E. Gavves, J. Oramas M., A. Ghodrati, and T. Tuytelaars, "Rank pooling for action recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 4, pp. 773–787, Apr. 2017.

[17] X. Shu, G.-J. Qi, J. Tang, and J. Wang, "Weakly-shared deep transfer networks for heterogeneous-domain knowledge propagation," in *Proc. 23rd ACM Int. Conf. Multimedia*, 2015, pp. 35–44.

[18] C. Szegedy *et al.*, "Going deeper with convolutions," in *Proc. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 1–9.

[19] B. Singh, T. K. Marks, M. Jones, O. Tuzel, and M. Shao, "A multi-stream bi-directional recurrent neural network for fine-grained action detection," in *Proc. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 1961–1970.

[20] M. Ma, H. Fan, and K. M. Kitani, "Going deeper into first-person activity recognition," in *Proc. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 1894–1903.

[21] G. Gkioxari and J. Malik, "Finding action tubes," in *Proc. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 759–768.

[22] J. R. R. Uijlings, K. E. A. van de Sande, T. Gevers, and A. W. M. Smeulders, "Selective search for object recognition," *Int. J. Comput. Vis.*, vol. 104, no. 2, pp. 154–171, Apr. 2013.

[23] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman, "Return of the devil in the details: Delving deep into convolutional nets," in *Proc. Brit. Mach. Vis. Conf.*, 2014, pp. 1–28.

[24] H. Jégou, M. Douze, C. Schmid, and P. Pérez, "Aggregating local descriptors into a compact image representation," in *Proc. Comput. Vis. Pattern Recognit.*, Jun. 2010, pp. 3304–3311.

[25] Z. Wu, Y.-G. Jiang, X. Wang, H. Ye, and X. Xue, "Multi-stream multi-class fusion of deep networks for video classification," in *Proc. ACM Multimedia Conf.*, 2016, pp. 791–800.

[26] X. Yang, P. Molchanov, and J. Kautz, "Multilayer and multimodal fusion of deep neural networks for video classification," in *Proc. ACM Multimedia Conf.*, 2016, pp. 978–987.

[27] L. Wang, Y. Qiao, and X. Tang, "Action recognition with trajectory-pooled deep-convolutional descriptors," in *Proc. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 4305–4314.

[28] H. Wang and C. Schmid, "Action recognition with improved trajectories," in *Proc. Int. Conf. Comput. Vis.*, Dec. 2013, pp. 3551–3558.

[29] W. Chen and J. J. Corso, "Action detection by implicit intentional motion clustering," in *Proc. Int. Conf. Comput. Vis.*, Dec. 2015, pp. 3298–3306.

[30] L. E. Wixson, "Detecting salient motion by accumulating directionally-consistent flow," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 8, pp. 774–780, Aug. 2000.

[31] Y. Zhai and M. Shah, "Visual attention detection in video sequences using spatiotemporal cues," in *Proc. 14th ACM Int. Conf. Multimedia*, 2006, pp. 815–824.

[32] S.-H. Zhong, Y. Liu, F. Ren, J. Zhang, and T. Ren, "Video saliency detection via dynamic consistent spatio-temporal attention modelling," in *Proc. 27th AAAI Conf. Artif. Intell.*, 2013, pp. 1063–1069.

[33] J. Yang *et al.*, "Discovering primary objects in videos by saliency fusion and iterative appearance estimation," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 26, no. 6, pp. 1070–1083, Jun. 2016.

[34] W. Wang, J. Shen, and F. Porikli, "Saliency-aware geodesic video object segmentation," in *Proc. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 3395–3402.

[35] J. Carreira and A. Zisserman, "Quo vadis, action recognition? A new model and the kinetics dataset," in *Proc. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 4724–4733.

[36] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.

[37] Z. Gao, L.-F. Cheong, and Y.-X. Wang, "Block-sparse RPCA for salient motion detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 10, pp. 1975–1987, Oct. 2014.
[38] A. Papazoglou and V. Ferrari, "Fast object segmentation in unconstrained video," in *Proc. Int. Conf. Comput. Vis.*, Dec. 2013, pp. 1777–1784.
[39] S. Oron, A. Bar-Hille, and S. Avidan, "Extended Lucas-Kanade tracking," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 142–156.
[40] Y.-H. Tsai, M.-H. Yang, and M. J. Black, "Video segmentation via object flow," in *Proc. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 3899–3908.
[41] P. Weinzaepfel, J. Revaud, Z. Harchaoui, and C. Schmid, "Learning to detect motion boundaries," in *Proc. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 2578–2586.
[42] E. Ilg, N. Mayer, T. Saikia, M. Keuper, A. Dosovitskiy, and T. Brox, "FlowNet 2.0: Evolution of optical flow estimation with deep networks," in *Proc. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 1647–1655.
[43] Z. Tu, N. van der Aa, C. Van Gemeren, and R. C. Veltkamp, "A combined post-filtering method to improve accuracy of variational optical flow estimation," *Pattern Recognit.*, vol. 47, no. 5, pp. 1926–1940, 2014.
[44] Z. Tu, W. Xie, J. Cao, C. van Gemeren, R. Poppe, and R. C. Veltkamp, "Variational method for joint optical flow estimation and edge-aware image restoration," *Pattern Recognit.*, vol. 65, pp. 11–25, May 2017.
[45] C. Ma, J.-B. Huang, X. Yang, and M.-H. Yang, "Hierarchical convolutional features for visual tracking," in *Proc. Int. Conf. Comput. Vis.*, Dec. 2015, pp. 3074–3082.
[46] R. S. Srivatsa and R. V. Babu, "Salient object detection via objectness measure," in *Proc. Int. Conf. Image Process.*, Sep. 2015, pp. 4481–4485.
[47] M.-M. Cheng, Z. Zhang, W.-Y. Lin, and P. Torr, "BING: Binarized normed gradients for objectness estimation at 300fps," in *Proc. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 3286–3293.
[48] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Süsstrunk, "SLIC superpixels compared to state-of-the-art superpixel methods," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 11, pp. 2274–2282, Nov. 2012.
[49] W. Zhu, S. Liang, Y. Wei, and J. Sun, "Saliency optimization from robust background detection," in *Proc. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 2814–2821.
[50] N. Otsu, "A threshold selection method from gray-level histograms," *IEEE Trans. Syst., Man, Cybern.*, vol. SMC-9, no. 1, pp. 62–66, Jan. 1979.
[51] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. Int. Conf. Learn. Represent.*, 2015, pp. 1–14.
[52] S. Gidaris and N. Komodakis, "Object detection via a multi-region and semantic segmentation-aware CNN model," in *Proc. Int. Conf. Comput. Vis.*, Dec. 2015, pp. 1134–1142.
[53] X. Peng and C. Schmid, "Multi-region two-stream R-CNN for action detection," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 744–759.
[54] M. D. Rodriguez, J. Ahmed, and M. Shah, "Action MACH a spatio-temporal maximum average correlation height filter for action recognition," in *Proc. Comput. Vis. Pattern Recognit.*, Jun. 2008, pp. 1–8.
[55] H. Jhuang, J. Gall, S. Zuffi, C. Schmid, and M. J. Black, "Towards understanding action recognition," in *Proc. Int. Conf. Comput. Vis.*, Dec. 2013, pp. 3192–3199.
[56] K. Soomro, A. Zamir, and M. Shah, "UCF101: A dataset of 101 human actions classes from videos in the wild," *CoRR*, pp. 1–6, Dec. 2012.
[57] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre, "HMDB: A large video database for human motion recognition," in *Proc. Int. Conf. Comput. Vis.*, Nov. 2011, pp. 2556–2563.
[58] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 1–11.
[59] X. Wang, A. Farhadi, and A. Gupta, "Actions $\sim$ transformations," in *Proc. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 2658–2667.
[60] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3D convolutional networks," in *Proc. Int. Conf. Comput. Vis.*, Dec. 2015, pp. 4489–4497.
[61] X. Peng, L. Wang, X. Wang, and Y. Qiao, "Bag of visual words and fusion methods for action recognition: Comprehensive study and good practice," *Comput. Vis. Image Understand.*, vol. 150, pp. 109–125, Sep. 2016.
[62] Z. Lan, M. Lin, X. Li, A. Hauptmann, and B. Raj, "Beyond Gaussian pyramid: Multi-skip feature stacking for action recognition," in *Proc. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 204–212.
[63] R. Girdhar, D. Ramanan, A. Gupta, J. Sivic, and B. Russell, "ActionVLAD: Learning spatio-temporal aggregation for action classification," in *Proc. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 3165–3174.
[64] C. Feichtenhofer, A. Pinz, and R. Wildes, "Spatiotemporal residual networks for video action recognition," in *Proc. Adv. Neural Inf. Process. Syst.*, 2016, pp. 3468–3476.
[65] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 770–778.
[66] R. Margolin, A. Tal, and L. Zelnik-Manor, "What makes a patch distinct?" in *Proc. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 1139–1146.
[67] W. Kim, C. Jung, and C. Kim, "Spatiotemporal saliency detection and its applications in static and dynamic scenes," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 21, no. 4, pp. 446–456, Apr. 2011.
[68] F. Perazzi, P. Krähenbuhl, Y. Pritch, and A. Hornung, "Saliency filters: Contrast based filtering for salient region detection," in *Proc. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 733–740.
[69] D. Tsai, M. Flagg, A. Nakazawa, and J. M. Rehg, "Motion coherent tracking using multi-label MRF optimization," *Int. J. Comput. Vis.*, vol. 100, no. 2, pp. 190–202, 2012.
[70] F. Li, T. Kim, A. Humayun, D. Tsai, and J. M. Rehg, "Video segmentation by tracking many figure-ground segments," in *Proc. Int. Conf. Comput. Vis.*, Dec. 2013, pp. 2192–2199.
[71] K. Fukuchi, K. Miyazato, A. Kimura, S. Takagi, and J. Yamato, "Saliency-based video segmentation with graph cuts and sequentially updated priors," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Jun./Jul. 2009, pp. 638–641.

Zhigang Tu received the master's degree in image processing from the School of Electronic Information, Wuhan University, China, in 2010, and the Ph.D. degree in computer science from Utrecht University, The Netherlands, in 2015. From 2015 to 2016, he was a Post-Doctoral Researcher with Arizona State University, USA. He is currently a Post-Doctoral Research Fellow with the School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore. His research interests include motion estimation, object segmentation, object tracking, action recognition and localization, and anomaly detection.

Wei Xie received the B.E. degree in electronic information engineering and the Ph.D. degree in communication and information system from Wuhan University, China, in 2004 and 2010, respectively. From 2010 to 2013, he was an Assistant Professor with the Computer School, Wuhan University. He is currently an Associate Professor with the Computer School, Central China Normal University, China. His research interests include motion estimation, super resolution reconstruction, image fusion, and image enhancement.

Justin Dauwels received the Ph.D. degree in electrical engineering from the Swiss Polytechnical Institute of Technology, Zurich, in 2005. He was a Post-Doctoral Fellow with the RIKEN Brain Science Institute from 2006 to 2007 and a Research Scientist with Massachusetts Institute of Technology from 2008 to 2010. He is currently an Associate Professor with the School of Electrical and Electronic Engineering, Nanyang Technological University (NTU), Singapore. He is also the Deputy Director of the ST Engineering NTU Corporate Lab, which comprises over 100 Ph.D. students, research staff members, and engineers, developing novel autonomous systems for airport operations and transportation. His research interests are in data analytics with applications to intelligent transportation systems, autonomous systems, and analysis of human behaviour and physiology. He was a JSPS Post-Doctoral Fellow in 2007, a BAEF Fellow in 2008, a Henri-Benedictus Fellow of the King Baudouin Foundation in 2008, and a JSPS Invited Fellow in 2010 and 2011. His research on intelligent transportation systems has been featured by the BBC, Straits Times, Lianhe Zaobao, Channel 5, and numerous technology websites.

Baoxin Li (SM'04) received the Ph.D. degree in electrical engineering from University of Maryland, College Park, in 2000. From 2000 to 2004, he was a Senior Researcher with SHARP Laboratories of America, Camas, WA, USA, where he was the Technical Lead in developing SHARPs HiIMPACT Sports technologies. From 2003 to 2004, he was also an Adjunct Professor with Portland State University, Portland, OR, USA. He is currently a Full Professor and the Chair of computer science and engineering with Arizona State University, Phoenix, AZ, USA. He holds nine issued U.S. patents. His current research interests include computer vision and pattern recognition, image/video processing, multimedia, medical image processing, and statistical methods in visual computing. He was a recipient of the SHARP Laboratories President Award in 2001 and 2004, the SHARP Laboratories Inventor of the Year Award in 2002, and the National Science Foundations CAREER Award from 2008 to 2009.

Junsong Yuan (M'08–SM'14) received the bachelor's degree from Huazhong University of Science and Technology, Wuhan, China, in 2002, through the Special Class for the Gifted Young, the M.Eng. degree from National University of Singapore, in 2005, and the Ph.D. degree from Northwestern University in 2009. He was an Associate Professor with Nanyang Technological University (NTU), Singapore. He is currently an Associate Professor with the Computer Science and Engineering Department, State University of New York at Buffalo. His research interests include computer vision, video analytics, gesture and action analysis, large-scale visual search, and mining. He received the Best Paper Award from International Conference on Advanced Robotics (ICAR17), the 2016 Best Paper Award from IEEE TRANSACTIONS ON MULTIMEDIA, the Doctoral Spotlight Award from the IEEE Conference on Computer Vision and Pattern Recognition (CVPR09), the Nanyang Assistant Professorship from NTU, and the Outstanding EECS Ph.D. Thesis Award from Northwestern University.

He served as a Guest Editor for *International Journal of Computer Vision*. He is currently a Senior Area Editor of *Journal of Visual Communications and Image Representations* and an Associate Editor of IEEE TRANSACTIONS ON IMAGE PROCESSING and IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY. He is the Program Co-Chair of ICME18 and VCIP15 and the Area Chair of ACM MM18, ICPR18, CVPR17, ICIP1817, and ACCV1814.