2016 23rd International Conference on Pattern Recognition (ICPR)
Cancún Center, Cancún, México, December 4-8, 2016

MSR-CNN: Applying Motion Salient Region Based Descriptors for Action Recognition

Zhigang Tu
School of Computing, Informatics,
Decision System Engineering
Arizona State University
Tempe, USA
Email: Zhigang.Tu@asu.edu

Jun Cao
Intel Corp.
Tempe, USA
Email: jun.cao@intel.com

Yikang Li and Baoxin Li
School of Computing, Informatics,
Decision System Engineering
Arizona State University
Tempe, USA
Email: YikangLi,Baoxin.Li@asu.edu

***Abstract*—In recent years the most popular video-based human action recognition methods rely on extracting feature representations using Convolutional Neural Networks (CNN) and then using these representations to classify actions. In this work, we propose a fast and accurate video representation that is derived from the motion-salient region (MSR), which represents features most useful for action labeling. By improving a well-performed foreground detection technique, the region of interest (ROI) corresponding to actors in the foreground in both the appearance and the motion field can be detected under various realistic challenges. Furthermore, we propose a complementary motion salient measure to select a secondary ROI – the major moving part of the human. Accordingly, a MSR-based CNN descriptor (MSR-CNN) is formulated to recognize human action, where the descriptor incorporates appearance and motion features along with tracks of MSR. The computation can be efficiently implemented due to two characteristics: 1) only part of the RGB image and the motion field need to be processed; 2) less data is used as input for the CNN feature extraction. Comparative evaluation on JHMDB and UCF Sports datasets shows that our method outperforms the state-of-the-art in both efficiency and accuracy.**

***Index Terms*—Action recognition; Motion salient regions; Convolutional Neural Networks**

I. Introduction

The amount of video data available is experiencing explosive growth due to ubiquity of digital recording devices and popularity of video sharing web sites. Human action recognition in video, which is one of the long-standing research topics in computer vision, has been extensively investigated in recent years [1], [4]. In general, action recognition can be considered as a two-step procedure: feature extraction and subsequent classification using these features.

Video-based action recognition is a challenging problem, with many difficulties yet to be resolved. The challenges come from three main aspects [4]: (1) Intra-class (variations within action classes) and inter-class (ambiguities between action classes) variations; (2) Environment and recording settings; and (3) Temporal variations. In this work, we concern how to accurately detect the person and his/her primary moving body part under various complicated conditions, and subsequently localize them. This is a promising way to improve video representation, which eventually determines the result of action recognition, since the performance of action recognition heavily depends on video representation.

Noticeably, after the deep convolutional neural networks (CNN) was applied in [8] to achieve remarkable success in static image classification, extending CNN to extract features for video representations has been widely studied for action recognition [1], [2], [3], [12]. Human actions in video can naturally be viewed as 3D spatio-temporal signals, which are characterized by the temporal evolution of visual appearance governed by motion [10]. In consistence with this characteristic, the approaches used to learn spatiotemporal features to represent spatially and temporally coupled action patterns are exploited. One representative work is [11], which presents two CNNs: one spatial CNN, in which the appearance representations are learned from RGB inputs, and one motion CNN, in which the motion representations are learned from pre-computed optical flow. These two representations are complementary, and better performance was obtained when combining them. We adopt the two-stream network, and improve it by proposing a technique based on motion-salient region (MSR).

In general, features used for human detection belong to global representations, which encode the region of interest (ROI) of a human as a whole. The ROI is normally extracted by leveraging background subtraction or tracking [4]. The global representations depend on the performance of localization, background subtraction or tracking. Furthermore, they are sensitive to variations in viewpoint, background motion, noise and illumination changes. Recently, [2] applied the selective search scheme [7] to produce approximately 2K regions in each frame, and discard the regions that are void of motion according to a motion salient measure (MSM) based on optical flow. However, this method has three drawbacks. First, there is no good method to select the motion salient threshold α, which directly affects the selected regions that are salient in shape and motion, and hence affecting the final accuracy and efficiency of the approach. Second, some subtle actions with small motion could be missed. Third, the selected regions are not necessarily spatially coherent.

On the other hand, Cheron *et al.* [1] obtained the representations derived from human pose. In particular, they used

978-1-5090-4847-2/16/$31.00 ©2016 IEEE

positions of the estimated body joints to define informative regions. The regions corresponding to four body parts – right hand, left hand, upper body and full body, and plus the full image in both the RGB image and the flow field are cropped. This method faces two main problems though. First, human-pose estimation is a difficult task. Pose-estimators should be avoided for action recognition, at least until the performance of the pose estimation being enhanced [5]. Second, using five inputs for CNN feature extraction leads to extensive computation that may not be completely necessary.

Inspired by the above analysis and the current advances in the domain of moving object detection, we formulate an action descriptor based on the identified motion-salient regions (MSRs). The Block-sparse Robust Principal Component Analysis (B-RPCA) technique [13] is employed to detect the human. The B-RPCA method addresses various realistic challenges, e.g., background motions, illumination changes, poor image quality under low light, noise and camouflage, in a unified framework. Not only the foreground individuals can be accurately extracted, but also the implementation is efficient. In addition, since a motion saliency estimation step is applied to compute the support of the foreground regions, spatial coherence is imposed on these target regions. To improve the performance of the B-RPCA technique, we add a velocity angle measure to reduce errors on the consistency of the motion direction. Normally, for the current widely-used action recognition video datasets, the detected MSR in each frame is the full human. According to the obtained motion information of the whole human, we propose another MSM to extract one primary part of the human body, where the movement is most distinctive. The secondary MSR can convey highly discriminative information, which is complementary to the first detected MSR of the whole human. Replacing the four body parts of [1] with our two MSRs, the proposed MSR-based CNN (MSR-CNN) outperforms the closely related state-of-the-art methods: the pose-based CNN (P-CNN), and the regions of interest based spatial- and motion-CNN [2] on both evaluation datasets.

II. MSR-CNN: CNN Features Extraction from MSRs

From the state-of-the-art work [1], [2], [16] and our analysis, it is clear that selecting informative regions for CNN features extraction is a effective way to modify the accuracy of action recognition in video. Fig. 1 and Fig. 2 outline the framework of our MSR-CNN schematically. We detect two complementary MSRs in terms of the improved B-RPCA technique and a MSM separately. As shown in Fig. 1, P1 and P2 are two extracted MSRs. P3 (see Fig. 2) is a 224×224 patch which is obtained by resizing the input full image or the full flow field. This method significant decreases the number of regions need to be processed and allows for faster computation. The two CNNs of [1] are introduced to operate on the MSRs of the RGB image and the optical flow respectively, and correspondingly producing two representations – the appearance-based CNN representation and the motion-based representation. These two representations are captured at each frame and then concatenated over time to form a video representation. At last, the action classification is performed with a linear SVM on the extracted video representation (see Fig. 2).

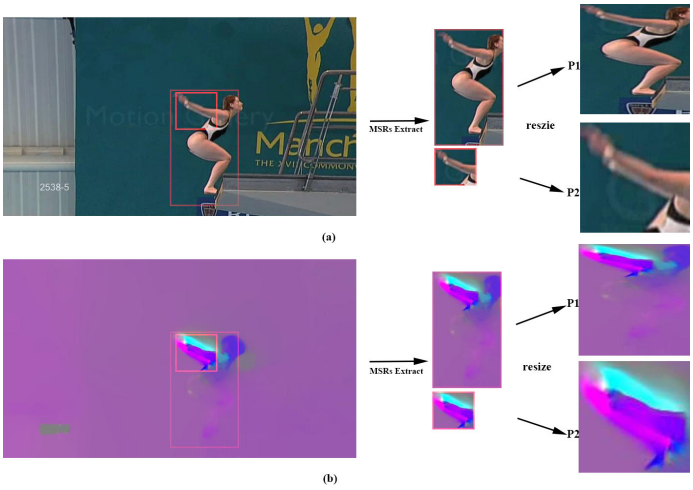

Fig. 1. Input data processing: MSRs extraction and resizing. (a) operation on the RGB input image; (b) operation on the motion field. (P1 (patch1) denotes one MSR – the human body; P2 (patch2) denotes the secondary MSR – motion salient body part)

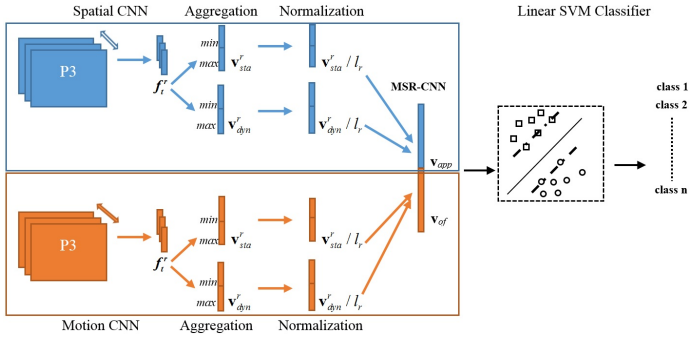

Fig. 2. The framework of our MSR-CNN. From left to right: The resized extracted three patches of each frame. The appearance descriptor of the spatial-CNN and the flow descriptor of the motion-CNN f_t^r is respectively captured per frame t and per region r. Static frame descriptors f_t^r are aggregated across all the frames according to *min* and *max* to get the video descriptor v_{sta}^r. Temporal differences of f_t^r are aggregated to v_{dyn}^r in the same way. Video descriptors are normalized and concatenated over patches r and aggregation schemes into appearance features v_{app} and flow features v_{of}. The final MSR-CNN feature representation is the concatenation of v_{app} and v_{of}. At last, a linear SVM classifier is carried out for action classification.

A. CNN Descriptors

To capture MSR-CNN features, we adopt the same architecture and training procedure as [1]. We apply the MatConvNet toolbox [17] for the convolutional networks. Below, a brief description of our training process is given.

1) Step 1: Processing the input data: To construct a motion-CNN, the optical flow is first calculated for each successive pair of frames according to the method of [18]. Optical flow [19], [20], which describes the pattern of apparent motion of objects in a scene, is of critical importance for

action recognition in video. The *x*-component (i.e. *u*), the *y*-component (i.e. *v*) and the magnitude of the flow are rescaled to the range of [0, 255] like the input RGB images in the following way: $[\widehat{u}, \widehat{v}] = \gamma[u, v] + 128$, where $\gamma = 16$ is the rescale factor. The values smaller than 0 and larger than 255 are discarded. Then, the three components of every flow are stacked to form a 3D image as the input for the motion-CNN. During training, for each selected MSR, we resize it to 224×224 to fit the CNN input layer. To construct a spatial-CNN, for each selected MSR in the RGB image, we also resize it to 224×224.

2) Step 2: Selecting/Training CNN model: Two different CNNs with an identical architecture (similar to [8], with 5 convolutional and 3 fully-connected layers) are employed to obtain the representations of the MSRs on appearance and motion field respectively. The public available model "VGG-f" [21], which is a pre-trained model on the ImageNet challenge database [22], is chosen for spatial-CNN. The state-of-the-art motion network [2], which has been pre-trained on the UCF101 dataset [23], is selected for motion-CNN.

3) Step 3: Aggregation: (1) Formulating a *video descriptor* by aggregating all frame descriptor f_t^r (*r* represents the MSR, *t* denotes the frame at time *t*). In particular, the frame descriptor f_t^r contains $n = 4096$ values which is the output of the second fully-connected layer.

(2) Formulating the *min* and *max aggregation* by calculating the minimum and maximum values for each descriptor dimension *i* over T frames:

$$\begin{aligned} m_i &= \min_{1 \le t \le T} f_t^r(i) \\ M_i &= \max_{1 \le t \le T} f_t^r(i) \end{aligned} \tag{1}$$

(3) Formulating the *static video descriptor* v_{sta}^r by concatenating the time-aggregated frame descriptors:

$$\mathrm{v}_{sta}^r = [m_1, \ldots, m_n, M_1, \ldots, M_n]^T \tag{2}$$

(4) Formulating the *dynamic video descriptor* v_{dyn}^r by concatenating the minimum $\triangle m_i$ and maximum $\triangle m_i$ aggregations of $\triangle f_t^r$

$$\mathrm{v}_{dyn}^r = [\triangle m_1, \ldots, \triangle m_n, \triangle M_1, \ldots, \triangle M_n]^T \tag{3}$$

where $\triangle f_t^r = f_{t+\triangle t}^r - f_t^r$, $\triangle t = 4$ is the time interval.

(5) Formulating a *spatio-temporal MSR-CNN descriptor* by aggregating all the normalized video descriptors for both appearance and motion of all MSRs and different aggregation strategies.

4) Step 4: Classification: The actions are categorized by using a linear SVM classifier trained on the spatio-temporal representations produced by our MSR-CNN.

III. Detection of Motion-salient Regions

Detecting moving objects is an extensively investigated subject [24] and significant progresses have been achieved in the past few years. Most existing techniques may still face some challenges with real data from complicated conditions. In this work, the B-RPCA technique [13] is employed for its overall good performance. Furthermore, an improved B-RPCA is presented to detect the foreground human in the input image. Besides, a MSM is exploited to extract one MSR in the human body detected from the previous step.

A. The B-RPCA Method

To deal with the difficulties in detecting foreground moving objects, Gao *et al.* [13] imposed few constraints to the background. The background can be identified according to a low-rank conditional matrix. Mathematically, the observed video frames can be considered as a matrix $\boldsymbol{M}$, which is a sum of a low-rank matrix $\boldsymbol{L}$ that denotes the background, and a sparse outlier matrix $\boldsymbol{S}$ that consists of the moving objects. Besides, [13] introduced a feedback scheme, and proposed a B-RPCA technique which consists a hierarchical two-pass process to handle the decomposition problem. Three major steps are carried out, which are summarized below to facilitate the discussion of our improvement later (Sect. III-B)

1) Step 1: First-pass RPCA: In this step, a first-pass RPCA in a sub-sampled resolution is applied to fast detect the likely regions of foreground:

$$\min_{\boldsymbol{L},\boldsymbol{S}} \| \boldsymbol{L} \|_* + \lambda \| \boldsymbol{S} \|_1, \qquad s.t. \quad \boldsymbol{M} = \boldsymbol{L} + \boldsymbol{S} \tag{4}$$

where $\| L \|^*$ denotes the nuclear norm of the background matrix $\boldsymbol{L}$, λ is a regularizing parameter which constraints no foreground regions will be missed. The appropriate value $\lambda = 1/\sqrt{\max(m, n)}$. Equation (4) is a convex optimization problem, and it can be solved by applying the inexact augmented Lagrange multiplier (ALM) [27]. Through this first-pass RPCA, all outliers can be identified and stored in the outlier matrix $\boldsymbol{S}$.

2) Step 2: Motion Saliency Estimation: A motion consistency strategy is used to assess the motion saliency of the detected foreground regions and the probability of a block containing the moving objects. Pixels within the blocks captured in the first round of RPCA are tracked by optical flow. After tracking, dense point trajectories are extracted. Firstly, the short trajectories, like $k - j <= 10$ (*j*, *k* represent the frame index, $j, k \in [1, n]$ rely on the trajectory *l*), are removed. Secondly, [26] is applied to estimate the motion saliency of the remaining trajectories according to the consistency of the motion direction. Two benefits are achieved due to the motion saliency estimation: (1) the foreground objects moving in a slow but consistent manner can be better identified; (2) the small local motion comes from inconsistent motions of the background can be further discard. Most of the non-stationary background motions identified and stored in the outlier matrix $\boldsymbol{S}$ in the first step, are filtered off or suppressed.

3) Step 3: Second-pass RPCA: In this step, the λ value is reset according to the motion saliency, which ensures the changes derived from the foreground motion can be completely transferred to the outlier matrix $\boldsymbol{S}$ and avoids to leave any bad presence in the background. The second pass RPCA is implemented as:

$$\min_{\boldsymbol{L},\boldsymbol{S}} \| \boldsymbol{L} \|_* + \sum_i \lambda_i \| \boldsymbol{P}_i(\boldsymbol{S}) \|_F, \quad s.t. \quad \boldsymbol{M} = \boldsymbol{L} + \boldsymbol{S} \tag{5}$$

where $\|\cdot\|_F$ denotes the Frobenius norm of a matrix. $\boldsymbol{P}_i$ is an operator which unstacks every column of $\boldsymbol{S}$ and returns a matrix that represents block *i*. The inexact ALM algorithm is emplyed again to solve the equation (5).

B. The Improved B-RPCA Method

The motion saliency estimation in [13] utilizes the trajectory length and the motion direction of the point trajectories to remove the non-stationary background motion. This strategy is effective to detect the foreground moving objects that keep moving constantly in the scene. If the object stops occasionally, the foreground object cannot be detected via the B-RPCA technique due to the Step 2 operation: motion saliency estimation. Especially for the action recognition datasets, such as JHMDB, the actors may have little motion for some intervals of the actions. To overcome such difficulties for the B-RPCA approach, we propose the following improvements:

1) Relaxing the constraint of the trajectory length to $k - j <= 5$. In this way, falsely-removed foreground due to trajectory length (e.g., when the actor suddenly stops for short moment) will be significantly reduced. Meanwhile, to avoid noise arising from the background, we add the motion derivative constraint similar to MBH [15]. By calculating derivatives of the optical flow components *u* and *v*, the background motion due to locally-constant camera motion will be excluded.

2) Enhancing the consistency measure of the motion direction. Not only the negative direction or positive direction of *u* and *v* along the trajectory, but also the direction variation should be considered. Hence, we add a velocity angle measure as follows:

$$\triangle\theta = \arctan(u_{t+1}/v_{t+1}) - \arctan(u_t/v_t) \in [-\pi/4, \pi/4] \quad (6)$$

where $[u, v] \neq 0$. Same as the motion direction consistency operation, this velocity angle measure is also conducted at positions where the velocity is no-zero along the trajectory.

3) If u_t and v_t satisfy either of the following conditions (Equation 7 or Equation 8), we consider the actor is static between frame *t* and frame $t+1$. Then, we only perform the first ***Step 1*** to detect the actor in the RGB images.

$$range(u_t) < 0.5 \wedge range(mGflow) < 0.5 \quad (7)$$

or

$$range(v_t) < 0.5 \wedge range(mGflow) < 0.5 \quad (8)$$

where $range(u_t)$ denotes the difference between the maximum value of u_t and the minimum value of u_t. ***mGflow*** is the magnitude of the gradients of the optical flow (u_t, v_t), $mGflow = \sqrt{(u_t)_x^2 + (u_t)_y^2 + (v_t)_x^2 + (v_t)_x^2}$, 0.5 is an empirically selected threshold and denotes a half pixel distance.

C. Selecting the MSR of the human

As suggested in [1], [2], [16], selecting suitable MSRs of the actor body is essential, as these body parts are complementary and are potentially helpful for improving action recognition when combined in an appropriate manner. Based on the captured human information of the improved B-RPCA, we introduce a MSM to select the MSR of the human body, where the motion is most distinguishable.

Fig. 3. An outline of our method to select one MSR in the human body. From Left to Right: the result of step 1 – extracting MSR candidates, the result of step 2 – discarding the small MSR candidates, and the result of step 3 – selecting the most salient motion region (the red rectangle).

The region of the foreground actor body is detected and localized via the improved B-RPCA. Accordingly, the location information is useful for identifying other informative body parts which are discriminative. We employ the following steps. (See Fig. 3)

1) Extracting MSR candidates from the detected human body according to a conditional measure defined as:

$$LabH \wedge (mGflow > AmGflow) \wedge (mflow > Amflow) \quad (9)$$

and

$$(|u| > Au) \vee (|v| > Av) \quad (10)$$

where ***LabH*** is the already obtained human body from the last step. ***AmGflow*** is the mean of $mGflow$. $mflow$ is the magnitude of the optical flow $flow = (u, v)$, $mflow = \sqrt{u^2 + v^2}$. ***Amflow*** is the mean of $mflow$. Au and Av is the mean of the horizontal flow u and the vertical flow v.

2) Discarding the small MSR candidates. Different body parts have different motion patterns. In addition, some background motion around the human body may be inaccurately identified by the B-RPCA technique. Due to this implementation, the incorrectly captured background motions could be removed once again. The 3th subfigure in Fig. 3 displays that most of the outliers are suppressed.

$$MSR(i) > \tau \quad (11)$$

where *i* is the index of ***MSR*** candidates. τ is a threshold. If the area of one candidate $MSR(i)$ is smaller than τ, it will be removed. In this paper we set $\tau = 10 \times 10$ experimentally.

3) Capturing the first two largest MSR candidates. We adopt the simple MSM as [2] to select the final MSR by comparing the normalized magnitude of the optical flow between these two candidates:

$$flow_m(R_i) = \frac{1}{|R_i|} \sum_{j \in R_i} flow(j) \quad (12)$$

where $flow_m(R_i)$ is the normalized magnitude of the optical flow in the *i*-th MSR candidate, *j* is the index of the optical flow. The MSR candidate with larger $flow_m(R_i)$ will be finally selected.

TABLE I
RESULTS (% MEAN AVERAGE PRECISION (MAP)) OF THE SPATIAL-MOTION MSR BASED CNN ON THE UCF SPORT DATASET.

Patches	Div.	Golf	Kick.	Lift.	Rid.	Run	S.Board.	Swing1	Swing2	Walk	mAP
P1	100	100	100	100	100	63.89	0	87.67	100	100	85.16
P2	100	100	52.50	100	100	63.89	0	100	100	100	81.64
P3	100	100	52.50	100	100	63.89	63.89	63.89	100	100	84.42
P1+P3	100	100	100	100	100	29.17	52.50	63.89	100	100	84.56
P2+P3	100	100	100	100	100	29.17	25.0	100	100	100	85.42
All	100	100	100	100	100	63.89	100	87.67	100	100	**96.39**

IV. EXPERIMENTS

In this section, we evaluate our MSR-CNN method by testing it on two challenging datasets – UCF Sports [14] and JHMDB [6], and compare it with the state-of-the-art algorithms. In particular, in each dataset, we assess our method in two aspects: 1) whether the improved B-RPCA is effective in detecting the foreground human under complicated situations, and the proposed MSM can extract the MSR of the body part; 2) whether the extracted secondary MSR is complementary to the first one, and if it can further enhance the performance.

A. Evaluation on UCF Sports

Fig. 4 shows the two detected MSRs on 6 different action categories. These actions are operated in various challenging conditions, such as multiple actors and the displacements are larger than the object scale (the 1th and 3th subfigures), the moving area is textureless (the 2th subfigure), occlusion (the 4th subfigure), motion blur (the 5th subfigure) and illumination changes (the 6th subfigure), our two detectors can successfully deal with these difficulties.

Table I shows the results of our proposed MSR based spatial-temporal CNN technique on using different patches. Comparing the first row and the second row, we can find that for some sequences, the extracted CNN features from these two MSRs are different and complementary. Consequently, extracting these two MSRs are necessary as they have different contributions for action recognition. Integrating the three patches together, significant gain is achieved, where the *mAP* is increased from 84.02% (*P3*) to 96.39% (*All*).

B. Evaluation on JHMDB

Fig. 5 shows the action detection and localization performance of our improved B-RPCA as well as the MSM. It is clear that our detectors perform well in complex and realistic situations. For example, in the 5th subfigure, even encountering with the extremely motion blur, the human body and one of his body part are accurately captured.

Table II demonstrates again that different patches play different significant roles in action recognition, and incorporating them can further increase the performance to recognize actions. The results of 71.1% from *All* outperforms other approach more than 4% (comparing with the second best result 68.2% of *P2+P3*).

Table III shows the results of different patches based MSR-CNN in the accuracy manner. The best result still comes from *All* (our proposed MSR-CNN), where the accuracy achieves to 66.2%.

TABLE III
RESULTS (% ACCURACY) OF THE SPATIAL-MOTION MSR BASED CNN ON THE JHMDB DATASET.

	P1	P2	P3	P1 + P3	P2 + P3	All
Accuracy(%)	59.79	58.92	60.78	63.76	65.88	**66.02**

TABLE IV
PERFORMANCE OF OUR MSR-CNN ON THE JHMDB DATASET. WE COMPARE THE MSR-CNN WITH THE STATE-OF-THE-ART RELATED METHODS: ACTION TUBES [2] AND POSE-CNN [1].

Methods	P-CNN (Without GT) [1]	Action Tubes [2]	MSR-CNN
Accuracy(%)	61.1	62.5	**66.02**

C. Comparison with the state-of-the-art

In Table IV, we compare our MSR-CNN approach with the two state-of-the-art algorithms. The accuracy of our method is about 5% better than Pose-CNN (66.02 *vs* 61.1), and about 3.5% more accurate than Action Tubes (66.02 *vs* 62.5). As we have analyzed in Introduction, compared with Pose-CNN, which relies on pose estimation, our detectors handle the difficulties in realistic data without requiring accurate pose estimation. Our method can extract the moving actor body as well as one of its primary moving body part precisely. Compared with Action Tubes, our method outperforms it due to its three drawbacks that degrade its performance. In particular, the empirically-selected motion salient threshold α in Action Tubes is a fixed constant, which is not optimal for all videos. Not only some fine-scale moving objects would be removed, but also some large-scale moving objects in challenging conditions would be incorrectly captured.

Since we focus on extracting complementary motion regions of the human body and its body parts, we do not experiment on integrating the hand-crafted IDT features [15] with our deep-learned MSR-CNN features. The combination can be easily conducted via fusion, and that can further boost the performance (Refer to Pose-CNN [1] for more details).

V. CONCLUSIONS

We propose a motion-salient region based convolutional neural networks (MSR-CNN) for action recognition and localization. The idea is derived from the intrinsic characteristic that only local motion features in the video contribute to the action label. By employing the B-RPCA method and improving its performance in three aspects, the foreground actor can be accurately detected under complex realistic situations. Additionally, based on the motion information obtained from the improved B-RPCA, a simple MSM can be used to efficiently extract a complementary MSR of the human body, which corresponds to the most discriminative motion part of the body. Therefore, it should contribute more to the action recognition. Evaluation on two challenging datasets and comparison with the related state-of-the-art algorithms demonstrate our method achieves superior performance on both the task of action recognition

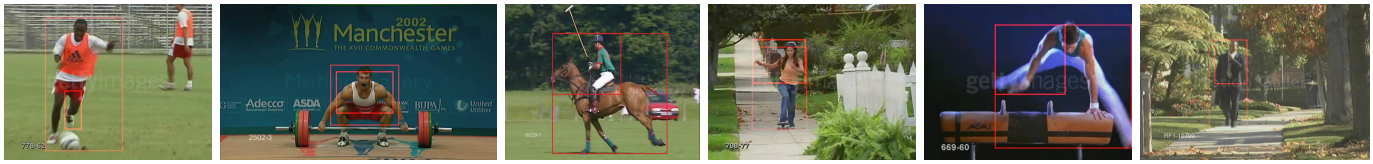

Fig. 4. Results on UCF Sports. Each column represents an action class. The big rectangle corresponds to the extracted foreground human body via the Improved B-RPCA method, the small one corresponds to the extracted secondary MSR via the proposed MSM.

TABLE II
RESULTS (% MEAN AVERAGE PRECISION (MAP)) OF THE SPATIAL-MOTION MSR BASED CNN ON THE JHMDB DATASET. THE RESPECTIVE PERFORMANCE OF P1, P2 AND P3, AND THE COMBINATION PERFORMANCE BY INTEGRATING THEM IN DIFFERENT WAYS ARE SHOWN.

Patches	brushhair	catch	clap	climbstairs	golf	jump	kickball	pick	pour	pullup	push	run	shootball	shootbow	shootgun	sit	stand	wingbaseball	throw	walk	wave	mAP
P1	76.6	54.6	63.3	58.5	88.8	43.4	48.0	57.7	87.4	98.8	82.0	55.3	38.6	80.1	60.1	74.8	72.9	63.4	8.9	85.8	57.3	64.6
P2	93.6	54.0	71.9	45.9	80.8	54.1	54.6	61.8	80.8	97.1	93.0	65.6	48.4	70.1	63.7	65.9	69.5	60.6	22.5	64.7	39.2	64.7
P3	66.9	53.1	51.2	65.9	91.3	47.3	55.2	56.3	97.9	100	85.6	52.8	42.4	91.4	72.2	61.4	66.3	32.7	29.2	86.4	46.9	64.4
P1+P3	76.4	56.5	52.1	53.4	91.3	56.6	59.8	56.3	92.6	100	86.8	52.2	47.5	93.1	67.7	59.7	66.4	49.0	16.8	88.0	65.5	66.1
P2+P3	86.0	51.0	65.5	53.4	91.3	57.0	52.3	61.8	92.6	100	87.3	67.0	50.0	91.8	68.0	64.0	65.6	55.5	20.9	88.0	62.9	68.2
All	89.1	47.3	61.3	54.1	91.3	60.1	59.5	69.4	97.6	100	96.0	71.8	50.8	92.1	71.0	65.5	72.9	60.1	40.3	86.9	55.2	**71.1**

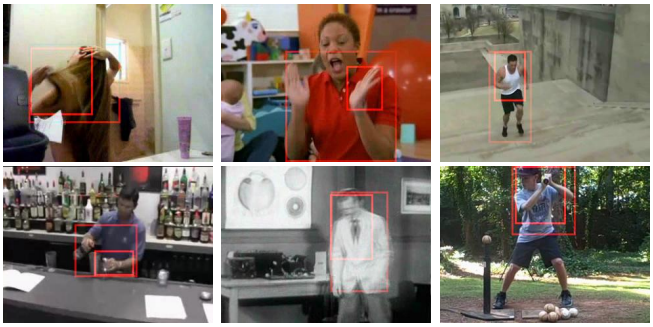

Fig. 5. Results on JHMDB. The big rectangle corresponds to the extracted foreground human body via the Improved B-RPCA method, the small one corresponds to the extracted secondary MSR via the proposed MSM.

and localization. In the future, we plan to design more robust and efficient approaches to extract MSRs, and analyze the type and number of MSRs that can bring the most significant contribution to action recognition.

ACKNOWLEDGMENT

The work was supported in part by ONR grant N00014-15-1-2344 and ARO grant W911NF1410371. Any opinions expressed in this material are those of the authors and do not necessarily reflect the views of ONR or ARO.

REFERENCES

[1] G. Cheron, I. Laptev, and C. Schmid, "P-CNN: Pose-based CNN features for action recognition," in *CVPR*, 2015.
[2] G. Gkioxari and J. Malik, "Finding action tubes," in *CVPR*, 2015.
[3] L. Wang, Y. Qiao, and X. Tang, "Action Recognition with Trajectory-Pooled Deep-Convolutional Descriptors," in *CVPR*, 2015.
[4] R. Poppe, "A survey on vision-based human action recognition," *Image Vis. Comput.*, vol.28, no.6, pp.976–990, 2010.
[5] W. Chen and J. Corso, "Action detection by implicit intentional motion clustering," in *ICCV*, 2015.
[6] H. Jhuang, J. Gall, S. Zuffi, C. Schmid, and M. J. Black, "Towards understanding action recognition," in *ICCV*, 2013.
[7] J. Uijlings, K. van de Sande, T. Gevers, and A. Smeulders, "Selective search for object recognition," *Int. J. Comput. Vis.*, 2013.
[8] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *NIPS*, 2012.
[9] J. Ng, M. Hausknecht, S. Vijayanarasimhan, O. Vinyals, R. Monga, and G. Toderici, "Beyond Short Snippets: Deep Networks for Video Classification," in *CVPR*, 2015.
[10] G. Varol, I. Laptev, and C. Schmid, "Long-term Temporal Convolutions for Action Recognition," in *CVPR*, 2016.
[11] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," in *NIPS*, 2014.
[12] Y. Wang, S. Wang, J. Tang, N. Hare, Y. Chang, and B. Li, "Hierarchical Attention Network for Action Recognition in Videos," in *CoRR abs/1607.06416*, 2016.
[13] Z. Gao, L. F. Cheong, and Y. X. Wang, "Block-Sparse RPCA for Salient Motion Detection," *Trans. Pattern Anal. Mach. Intell.*, vol.36, no.10, pp.1975–1987, 2014.
[14] M. D. Rodriguez, J. Ahmed, and M. Shah, "Action mach: a spatio-temporal maximum average correlation height filter for action recognition," in *CVPR*, 2008.
[15] H. Wang and C. Schmid, "Action recognition with improved trajectories," in *ICCV*, 2013.
[16] P. Weinzaepfel, Z. Harchaoui, and C. Schmid, "Learning to track for spatio-temporal action localization," in *CVPR*, 2015.
[17] A. Vedaldi and K. Lenc, "MatConvNet: Convolutional Neural Networks for MATLAB," in *ACM Int. Conf. Multimedia*, 2015.
[18] T. Brox, A. Bruhn, N. Papenberg, and J. Weickert, "High accuracy optical flow estimation based on a theory for warping," in *ECCV*, 2004.
[19] Z. Tu, N. Aa, C. V. 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.
[20] Z. Tu, R. Poppe, and R. C. Veltkamp, "Weighted local intensity fusion method for variational optical flow estimation," *Pattern Recognit.*, vol.50, pp.223–232, 2016.
[21] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman, "Return of the devil in the details: Delving deep into convolutional nets," in *BMVC*, 2014.
[22] J. Deng, W. Dong, R. Socher, L. Li, K. Li, and F. Li, "ImageNet: A large-scale hierarchical image database," in *CVPR*, 2009.
[23] K. Soomro, A. R. Zamir, and M. Shah, "Ucf101: A dataset of 101 human action classes from videos in the wild," in *CRCVTR-12-01*, 2012.
[24] A. Bugeau and P. Perez, "Detection and segmentation of moving objects in complex scenes," *Comput. Vis. Image Understand.*, 2009.
[25] E. Candes, X. Li, Y. Ma, and J. Wright, "Robust principal component analysis?," *J. ACM*, vol.58, no.3, pp.1–37, 2011.
[26] L. Wixson, "Detecting salient motion by accumulating directionally-consistent flow," *Trans. Pattern Anal. Mach. Intell.*, 2000.
[27] Z. Lin, M. Chen, and Y. Ma, "The augmented Lagrange multiplier method for exact recovery of corrupted low-rank matrix," *Mathematical Programming*, 2010.