Online Action Recognition based on Skeleton Motion Distribution

Bangli Liu[1]
bangli.liu@port.ac.uk
Zhaojie Ju[1]
zhaojie.ju@port.ac.uk
Naoyuki Kubota[2]
kubota@tmu.ac.jp
Honghai Liu[1]
honghai.liu@port.ac.uk

[1] School of Computing, University of Portsmouth, Portsmouth, UK
[2] Graduate School of System Design, Tokyo Metropolitan University, Tokyo, Japan

Abstract

Online action recognition which aims to jointly detect and recognize actions from video streams, is an essential step towards a comprehensive understanding of human behavior. However, it is challenging to accurately locate and recognize the occurrence of actions from noisy data streams. This paper proposes a skeleton motion distribution based method for effective online action recognition. Specifically, an adaptive density estimation function is built to calculate the density distribution of skeleton movements. Observing that each action has a unique motion distribution, we detect the occurrence of actions by identifying the transition of the motion distribution in a video stream. Once the starting point of an action is detected, a snippet-based classifier is proposed for online action recognition, which continuously identifies the most likely action class. Experimental results demonstrate that our method outperforms the state-of-the-art methods in terms of both detection accuracy and recognition precision.

1 INTRODUCTION

Human action recognition has received growing interest due to its wide range of applications in public surveillance, elderly care, and human-computer interaction. Although significant work has been done for offline action recognition where the actions are pre-segmented by providing the starting and ending frames [5, 9, 15, 16, 26], their performance remains unclear when applied to realistic scenarios where no prior information regarding the action's trigger time is available. Most of the common situations require the algorithms to automatically process the data stream without any prior information [3]. For example, the assisted robot should be able to provide immediate help for elderly people with minimal latency if they are going to fall down. Some work has been done for the similar task named early action recognition [12, 17, 25], which predicts actions before they are fully finished. However, this type of methods assumed that the starting time of actions is known beforehand. This solution can only be regarded as a partial answer to online action recognition since more attention is focused on recognition instead of detection. Compared to isolated action

recognition and early action recognition, online action recognition is significantly more challenging for two reasons: firstly, the execution boundaries of various action categories need to be detected accurately; secondly, only partial actions might be available for recognition due to the performance of action detection algorithm, thus the action recognition algorithm should be capable of recognizing actions from different action fragments.

In this paper, we address these problems by developing a novel skeleton motion distribution based method to simultaneously perform action detection and classification in continuous videos. The unique movement characteristics of each action class make the corresponding motion sequence has very different distribution from each other. Thus the distribution property at the beginning of a new action deviates so much from the previous action in a video sequence. The occurrence frame of actions is detected depending on the deviator of its density distribution with respect to the previous action. Once an action is detected, a snippet-based classifier is designed to process the observed video immediately for action classification. This classifier is performed in fragment level which could reduce the influence of false detections caused by noises and thus improve the accuracy.

The main contributions of our proposed approach are listed as follows: (1) jointly automatic action detection and recognition from continuous videos; (2) a skeleton motion distribution based framework to achieve effective action detection; (3) a snippet-based classifier to accurately recognize detected actions. The remainder of this paper is organized as follows: Section 2 reviews related work for human action detection and recognition. Section 3 introduces the proposed action detection and recognition. Section 4 reports experimental results as well as the comparison with the state-of-the-art methods. Section 5 summarizes the work of this paper.

2 Related Work

The development of cost-effective RGB-D sensors has encouraged a good deal of approaches using 3D skeleton joints proposed for human action recognition [8, 10, 20, 24, 32]. Qiao *et al*. [24] applied a trajectorylet based on local feature representation to capture ample static and dynamic information of actions. Exemplar-SVM was then used to select a discriminative trajectorylet to describe each action. Vemulapalli *et al*. [32] made use of the rotations and translations between body parts to model relative 3D geometry relation, with which human motion was encoded as curves in the Lie group. Although these methods have achieved satisfactory classification performance on delicately trimmed actions, the detection performance of these methods remains unclear when applied to realistic applications where action boundaries are not known. Some researchers proposed methods to recognize actions with appropriate latency or even before they are completely executed. For example, the sequential max-margin event detectors (SMMED) [14] and the easy naive Bayes framework [7] were proposed to identify actions when enough observations are obtained. Hoai *et al*. [12] assumed that a known action is going to happen in each video, and trained a structured output SVM to detect the action early. The assumption of the start point or the action content indeed limits their applicability due to practical video streams normally containing a lot of irrelevant negative actions.

Online action recognition which aims to jointly detect and recognize actions in untrimmed videos has received increasing attention in recent years. Some existing methods handle the problems using heuristics based sliding window approaches. For example, Song *et al*. [30] implemented online gesture interpretation and segmentation continuously by combining a

latent dynamic conditional random field and a temporal sliding window. In [18], each video frame was assigned a label based on the comparison between its representation and template representations. Instead of using sliding window, continuous action sequences were firstly segmented into isolated actions which were then input to classifiers for recognition in [2, 27, 29, 33]. Shao *et al*. [27] used the local maxima/minima value of varying motion gradients for action segmentation. Wu *et al*. [33] divided the action sequence into overlapping clips, which were clustered as action-words. They learned an action-topics model based on these action-words to represent their co-occurrence and temporal relations. The action segmentation was completed by assigning an action topic to each clip. The segments that may contain actions were firstly localized via a multi-stage convolutional neural network, and the action category in the segments was then identified in [29]. The main concern of these methods is that the timely responses which are essential in some scenarios (e.g., assistant life, public surveillance) might not be provided since action classification is performed after the complete actions are segmented. To speed up the performance of online action recognition, Bloom *et al*. [1] proposed to detect the action points [23, 28] that indicate the action peak frames. However, identifying a sole time instance might easily cause false detections when peaks of different actions are similar.

In this paper, the occurrence of actions is effectively detected via a skeleton motion distribution based framework. Once the starting time is determined, a snippet-based classifier is utilized to classify the actions continuously via a sliding window strategy. The influence of false detections could be eased by this fragment level classification.

3 Proposed Method

This paper proposes a skeleton motion distribution based framework for joint action detection and recognition in continuous videos. Firstly, the density distribution of skeleton motion is estimated, with which the local density relationship is calculated to determine the action boundary. Secondly, a snippet-based classifier is proposed to classify action clips in different stages continuously, followed by a smoothing strategy to increase the confidence of the detected action category.

3.1 Adaptive Density Estimation for Action Detection

3.1.1 Skeleton Motion

This paper utilizes the displacement offset of each joint to describe skeleton motion. Given a human skeleton motion sequence $d_1, d_2, ..., d_n \in N$, where N is the frame number. The coordinate of joints is translated to the hip-center coordinate system to make the following extracted feature invariant to different location. $d_t = [\Delta^1_{x_t,y_t,z_t}, \Delta^2_{x_t,y_t,z_t}, ..., \Delta^{20}_{x_t,y_t,z_t}]$ is the offset displacement feature of skeleton joints in 3D space at time t. Herein, skeleton displacement offset is computed as follows:

$$\begin{cases} \triangle x^i_t = x^i_t - x^i_1, \\ \triangle y^i_t = y^i_t - y^i_1, \\ \triangle z^i_t = z^i_t - z^i_1, \end{cases} \tag{1}$$

x^i, y^i, z^i are coordinates of the $i-th$ joint. $\triangle x^i_t, \triangle y^i_t, \triangle z^i_t$ are displacement offset in three directions with respect to the initial position. To obtain a compact representation, locality preserving projection (LPP) [11] is applied to reduce the dimensionality of skeleton motion

d. A transformation matrix W is calculated to map motion data d to a set of points p in a low dimensional space, such that $p = W^T d$. LPP can optimally preserve the local neighborhood structure of the data by restoring the neighbor relationship between original data in the weighted adjacency graph.

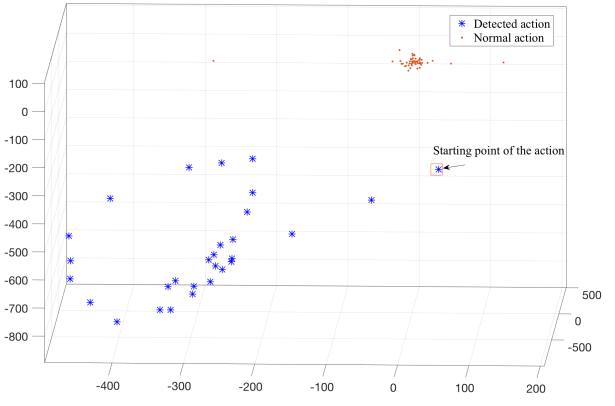

Figure 1: Example of action detection. The red dots and the blue stars are samples from the normal action and *running* action, respectively. The blue star in the rectangle is the starting point of *running*. The estimated density of *running* points even at the beginning of the action has huge difference compared to that of the normal action, which means our method could detect the occurrence of the action quickly.

In a video stream, when a different action begins to happen, the distribution of its motion data deviates so much from the previous action, as shown in Fig. 1. Thus, this difference is used to detect the occurrence of actions. Assuming the mapped motion data $\{p_1, p_2, ..., p_n\}$ is the random variable in a feature space from a distribution with an unknown density $q(p)$. The density $q(p)$ is obtained via a kernel density estimation function, which is a nonparametric density estimation without any assumptions about the underlying distributions, and has witnessed a great success in detecting outliers [19] and background subtraction [6, 22]. This paper adopts a multivariate Gaussian function with zero mean and unit standard deviation as the density function. Thus, the estimated $\widehat{q}(p)$ at point p can be formulated in Eq. 2:

$$\widehat{q}(p) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{(h(p_i))^k} \cdot \frac{1}{(\sqrt{2\pi})^k} exp(-\frac{||p-p_i||^2}{2(h(p_i))^2}) \tag{2}$$

where k is the dimensionality of data samples, and $h(p_i)$ is the bandwidth at point p_i.

Motion data is mostly multimodal, even in the same action category, resulting in that the data from different modality has different density. The estimated density might be not precise if the whole set of motion data is used to estimate the density. On the other hand, the distribution of a point should have the similar distribution with its neighbor points. To this end, the density estimation of a point is adapted by using its neighbor $Km(p)$. Thus, Eq. 2 is modified as follows:

$$\widehat{q}(p) \propto \frac{1}{m}\sum_{p_i \in Km(p)} \frac{1}{(\sqrt{2\pi}h(p_i))^k} exp(-\frac{||p-p_i||^2}{2(h(p_i))^2}) \tag{3}$$

where $Km(p)$ includes m samples ($m << n$) belonging to the neighbor of the point p. Compared to the holistic comparison, the local measure depending on $Km(p)$ yields more effective density estimation by reducing computation costs from the whole data set (n samples) to local neighbors (m samples).

To enhance the robustness of the density estimation function against the change of data distribution for different subjects or action instances, an adaptive bandwidth, $h(p_i) = h \cdot d_m(p_i)$, is achieved by considering the distance $d_m(p_i)$ between the point p_i and its $m-th$ neighbor, as shown in Eq. 4. This improvement makes the bandwidth small in density action samples while big in sparse ones, thus enables the density estimation function adaptive to various data density.

$$\widehat{q}(p) \propto \frac{1}{m} \sum_{p_i \in Km(p)} \frac{1}{(\sqrt{2\pi}h \cdot d_m(p_i))^k} exp(-\frac{||p-p_i||^2}{2(h \cdot d_m(p_i))^2}) \tag{4}$$

Since the density distribution of data points from different actions is distinctive due to their distinct movement characteristics, Observing that the density distribution of a different action starts to deviate very much from the previous action at its beginning in Fig. 1, the relative density relationship is used to effectively describe this difference, as denoted in Eq. 5:

$$LDR(p) \propto \frac{\sum_{p_i \in Km(p)} \frac{\widehat{q}(p_i)}{m}}{\widehat{q}(p) + c \cdot \sum_{p_i \in Km(p)} \frac{\widehat{q}(p_i)}{m}} \tag{5}$$

where $LDR(p)$ denotes the ratio between the density at p and the average density of its neighbors. c is a scaling constant to avoid infinity values of LDR caused by very small estimated density at the point p.

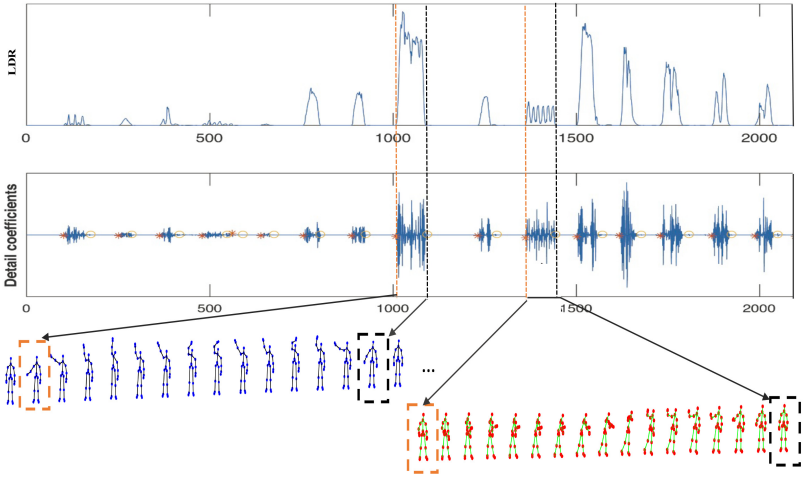

Figure 2: The local density ratio of a continuous video and action starting /ending points detection through wavelet transform.

Fig. 2 shows that the local estimated density of targeted actions has apparent difference compared to normal actions, and this value will convert dramatically at its starting and end-

ing due to the change of the density distribution from action to action. This sudden convert referred to as action boundaries in action sequences can be regarded as a type of impulses, which could be detected via wavelet transform. The wavelet transform is a powerful technique for analyzing irregular data, owing to its great capacity in providing the frequency and corresponding time location information of signals. This makes the wavelet transform suitable for detecting impulses occurring at any time. Approximation and detail coefficients are outputs from the low-pass and high-pass filter, respectively. The significant difference in density distribution between the occurrence of the action and its neighbors allows us to detect impulses during a time series with a high accuracy.

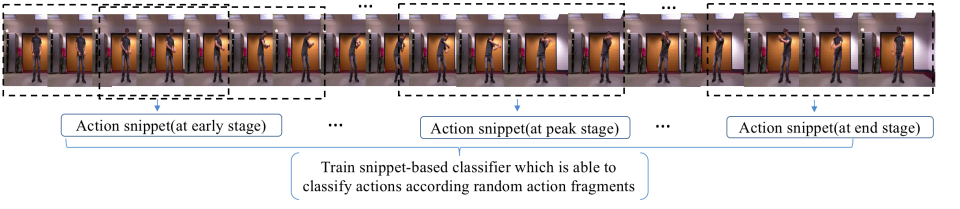

Figure 3: Snippet-based classifier. Action snippets in different performing stage are generated from the continuous video via sliding window.

3.2 Online Action Recognition

Since action durations exhibit considerable variability and only partial action observations might be available due to the performance of action detection algorithm, we explore features from snippets incorporating partial segments of actions in different performing stages. Fig. 3 shows a schematic illustration of our classification. Action snippets are randomly generated from untrimmed videos using a sliding window strategy. The proposed snippet-based classifier takes advantage of local temporal information, thus makes it robust to variations in execution time.

Compared to recognizing actions after they are totally completed, recognizing actions from partial action observations is more challenging due to the limited information. To make the feature descriptor discriminative for different categories, the 3D moving trend feature of joints [21] is adopted for action representation. The moving directions of skeleton joints, calculated using Eq. 6, are divided into various semantic words, and then a histogram which can quantitatively reflect the moving trend property is built by accumulating directions over the whole sequence.

$$\mathbf{v_t^i} = \{x_t^i - x_{t-1}^i, y_t^i - y_{t-1}^i, z_t^i - z_{t-1}^i\} \tag{6}$$

Cosine similarity and displacement are combined for soft voting during histogram quantification, as formulated in Eq. 7:

$$cos\theta_j^i(t) = \frac{\mathbf{v_j} \cdot \mathbf{v_t^i}}{\|\mathbf{v_t^i}\|\|\mathbf{v_j}\|}, j \in [1,m] \tag{7}$$

where $\mathbf{v_j} \in \mathbf{V}$ is the defined semantic words.

Table 1 lists the detailed procedure of action detection and recognition. The action boundaries are detected depending on the local density relationship in continuous action sequences. And then the snippet-based classification is processed to continuously classify

partial actions. Our method achieves lower computation cost compared to continuous recognition over the whole video, because the action recognition is processed intermittently if and only if the occurrence of actions is detected. In addition, The classifier performing in fragment level could reduce the influence of false detections and thus improve the performance.

Table 1: Framework of the proposed online action recognition.

Input: skeleton motion set $D = d_1, d_2, ..., d_t$ at $t = 1, 2, ..., T$.
Snippet-based classifier C, window size $length$, stride $step$ and threshold δ.

Output: Start points $StartP$ and end points $EndP$ of actions, action class $Label$

Initialization: $StartP = 0$, $EndP = 0$.
While $t < T$
· Map data d_t to a low dimension space using LPP: $p_i = W^T d_i$;
· Compute local density relationship $LDR(p_i)$;
· Detect $StartP$ and $EndP$ according to detail coefficients $cD1$ and δ using dwt:
$cD1[t] = \sum_{-\infty}^{+\infty} y[k]h[2t-k]$, where h is the high-pass filter;
If $cD1(t) > \delta$
$StartP = t$;
Start snippet-based action recognition from time t using sliding window;
Assign each snippet a specific class label $label(i)$ by the classifier C.
util $cD1(t) < \delta$
$EndP = t$;
Smooth the labels of snippets from the detected sequence over time $StartP$ to $EndP$;
Select the final $Label$ with highest probability to the detected sequence;
End
End

4 Experiment Results

The MAD database [14] is an RGB-D database providing continuous videos. It has 40 sequences and each of them contains 35 activities (e.g., *running, crouching, jumping, throw, basketball dribble, kicking*). Activities in each sequence are performed continuously with null class inserted between two activities. Three modalities were recorded in this database: RGB video, depth, and 3D coordinates of 20 human body joints. At test time, local density relationship is computed using Eq. 5 and then passed through the wavelet transform for detecting action boundaries. For a fair comparison, we performed five-fold-cross-validation as set in [14]. At train time, a snippet-based 36-class classifier is trained, where action snippets representing different action stage of one class were picked up from videos.

Table 2: Average Detection and Recognition results on *MAD* database(%)

Methods	*Prec*	*Rec*
SVM+DP [13]	28.6	51.4
SMMED [14]	59.2	57.4
ENB [7]	76.1	73.6
Method in [4]	72.1	79.7
Proposed	**84.8**	**80.8**

Table 2 gives the comparison of the average performance in terms of *precision percentage (Prec)* and *Recall percentage (Rec)* that defined in [14]. *TN*: number of correctly detected events who has 50% overlap with the ground truth event; *GTN*: Number of all ground truth

events; *DN*: number of detected events. $Prec = \frac{TN}{DN}$ and $Rec = \frac{TN}{GTN}$. From the table, we can see that our method achieves over 80% accuracies in *Prec* and *Rec* which outperform the other compared methods. Specifically, the *Prec* of our method is approximately 60% higher than SVM+DP [13], 25% higher than SMMED [14], and 8.7% higher than ENB [7]. On the other hand, our method improves the *Rec* rate by 7.2% compared to the result of ENB. Higher precision percentage and recall percentage of the proposed method indicate that our method has the capability to precisely recognize actions within detected actions and accurately detect actions from continuous videos.

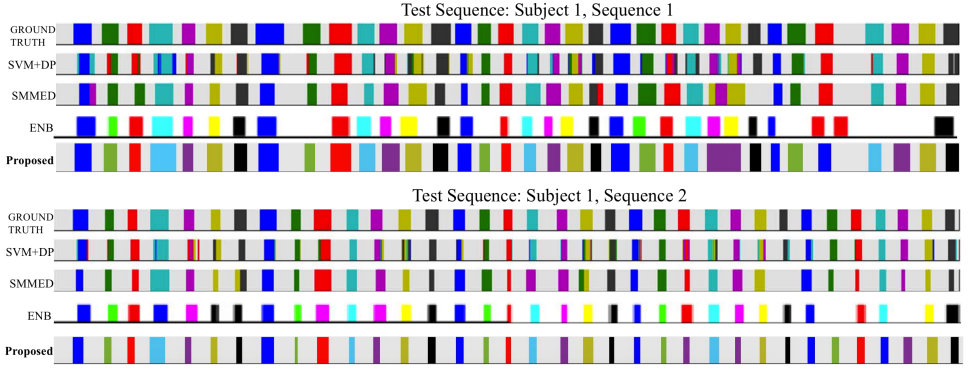

Figure 4: The comparison of the proposed method with SVM+DP [13], SMMED [14] and ENB [7] on two test sequences in the *MAD* database.

The action detection performance of different methods on two sample action sequences is also compared in Fig. 4. We reproduce the result of SVM+DP and SMMED from [14] and the result of ENB from [7]. It can be seen from these color bars that the performance of our method is the closest to the ground truth compared to other listed methods. Although SVM+DP and SMMED could detect almost all action occurrence, the accuracy of classification within a detected action is relatively low. In spite of better performance in classification within detected actions of ENB, its missing detection rate decreases. Our method is able to detect actions correctly with a lower missing detection rate than ENB and performs a higher classification accuracy than all listed methods. Furthermore, Fig. 5 reports the average detection accuracy of each action class. The majority of actions could be detected with the accuracy around 80%. The relatively low accuracy in action 4 ('walking'), action 22 ('Right Arm Dribble'), and action 23 ('Right Arm Pointing to the Ceiling') might be the result of the action similarity.

In addition, to demonstrate the classification performance of our method for fully completed actions, Fig. 6 reports the confusion matrix among 36 action categories. It can be seen that there is over 90% accuracy in almost all of the actions, and some actions such as 'Crouching', 'Jumping', and 'Left Arm Punch', could be succesfully recognized. These outstanding performance indicates that the adopted feature descriptor as well as the classifier have an outstanding capacity of distinguishing various actions. The quite similar movements of 'Right Arm Back Receive' and 'Swing from Left' result in the big confusion between each other.

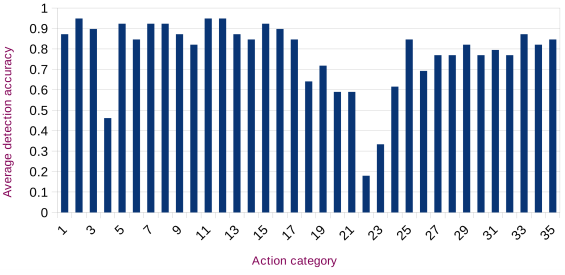

Figure 5: Average detection accuracy of each action category.

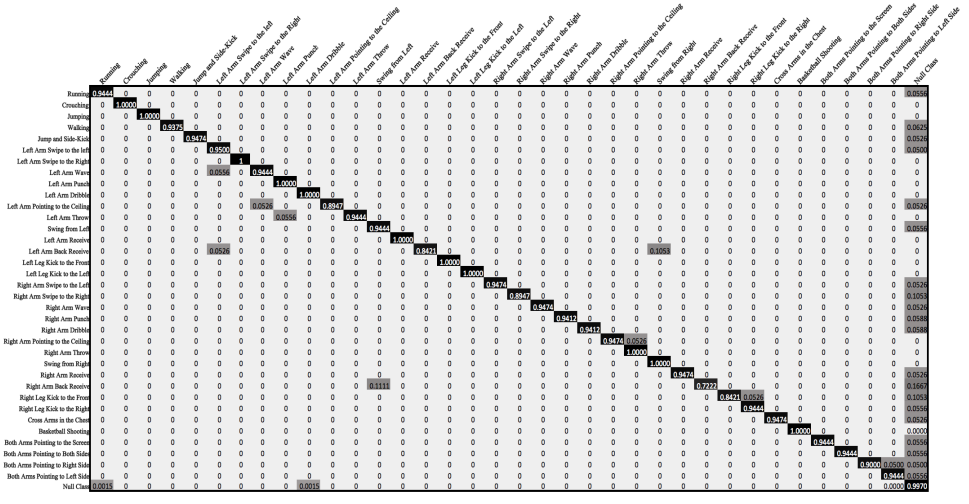

Figure 6: Confusion matrix on the *MAD* database.

5 Conclusion

This paper proposed a skeleton motion distribution based method to deal with the action detection and recognition problem in online action recognition. An adaptive density estimation function was developed for calculating the density distribution of skeleton motion in different actions. The transition of the density distribution from action to action was investigated for effective action detection. Furthermore, a snippet-based classifier which can handle action fragments was trained for the sequential action recognition once actions are detected. The comparison with the state-of-the-art methods in the publicly available database has shown that our method obtained outstanding results including detection and classification performance. However, the limitation of the proposed method is that the adopted skeleton data may not be that reliable due to occlusions or noise. In the future, we plan to improve the detection accuracy by fusing RGB and depth information.

6 Acknowledgements

The authors would like to acknowledge support from DREAM project of EU FP7-ICT (grant no. 611391)

References

[1] Victoria Bloom, Vasileios Argyriou, and Dimitrios Makris. Linear latent low dimensional space for online early action recognition and prediction. *Pattern Recog.*, 72: 532–547, 2017.

[2] Xiujuan Chai, Zhipeng Liu, Fang Yin, Zhuang Liu, and Xilin Chen. Two streams recurrent neural networks for large-scale continuous gesture recognition. In *Int. Conf. Pattern Recog. Workshops*, pages 31–36, 2016.

[3] Roeland De Geest, Efstratios Gavves, Amir Ghodrati, Zhenyang Li, Cees Snoek, and Tinne Tuytelaars. Online action detection. In *European Conference on Computer Vision*, pages 269–284. Springer, 2016.

[4] Maxime Devanne, Stefano Berretti, Pietro Pala, Hazem Wannous, Mohamed Daoudi, and Alberto Del Bimbo. Motion segment decomposition of rgb-d sequences for human behavior understanding. *Pattern Recognition*, 61:222–233, 2017.

[5] Yong Du, Wei Wang, and Liang Wang. Hierarchical recurrent neural network for skeleton based action recognition. pages 1110–1118, 2015.

[6] Ahmed Elgammal, Ramani Duraiswami, David Harwood, and Larry S Davis. Background and foreground modeling using nonparametric kernel density estimation for visual surveillance. *Proceedings of the IEEE*, 90(7):1151–1163, 2002.

[7] Hugo Jair Escalante, Eduardo F Morales, and L Enrique Sucar. A naive bayes baseline for early gesture recognition. *Pattern Recognition Letters*, 73:91–99, 2016.

[8] Salvatore Gaglio, Giuseppe Lo Re, and Marco Morana. Human activity recognition process using 3-d posture data. *IEEE Transactions on Human-Machine Systems*, 45 (5):586–597, 2015.

[9] Yao Guo, Youfu Li, and Zhanpeng Shao. Dsrf: A flexible trajectory descriptor for articulated human action recognition. *Pattern Recog.*, 2017.

[10] Yao Guo, Youfu Li, and Zhanpeng Shao. Dsrf: A flexible trajectory descriptor for articulated human action recognition. *Pattern Recognition*, 76:137–148, 2018.

[11] Xiaofei He and Partha Niyogi. Locality preserving projections. 16(2003), 2003.

[12] Minh Hoai and Fernando De la Torre. Max-margin early event detectors. *International Journal of Computer Vision*, 107(2):191–202, 2014.

[13] Minh Hoai, Zhen-Zhong Lan, and Fernando De la Torre. Joint segmentation and classification of human actions in video. *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 3265–3272, 2011.

[14] Dong Huang, Shitong Yao, Yi Wang, and Fernando De La Torre. Sequential max-margin event detectors. *European conference on computer vision*, pages 410–424, 2014.

[15] Ahmad Jalal, Yeon-Ho Kim, Yong-Joong Kim, Shaharyar Kamal, and Daijin Kim. Robust human activity recognition from depth video using spatiotemporal multi-fused features. *Pattern Recog.*, 61:295–308, 2017.

[16] Yu Kong and Yun Fu. Bilinear heterogeneous information machine for rgb-d action recognition. *Computer Vision and Pattern Recognition (CVPR), 2015 IEEE Conference on*, pages 1054–1062, 2015.

[17] Yu Kong, Dmitry Kit, and Yun Fu. A discriminative model with multiple temporal scales for action prediction. pages 596–611, 2014.

[18] Kaustubh Kulkarni, Georgios Evangelidis, Jan Cech, and Radu Horaud. Continuous action recognition based on sequence alignment. *International Journal of Computer Vision*, 112(1):90–114, 2015.

[19] Longin Jan Latecki, Aleksandar Lazarevic, and Dragoljub Pokrajac. Outlier detection with kernel density functions. In *International Workshop on Machine Learning and Data Mining in Pattern Recognition*, pages 61–75. Springer, 2007.

[20] Ivan Lillo, Juan Carlos Niebles, and Alvaro Soto. Sparse composition of body poses and atomic actions for human activity recognition in rgb-d videos. *Image and Vision Computing*, 59:63–75, 2017.

[21] Bangli Liu, Hui Yu, Xiaolong Zhou, Dan Tang, and Honghai Liu. Combining 3d joints moving trend and geometry property for human action recognition. In *Systems, Man, and Cybernetics (SMC), 2016 IEEE International Conference on*, pages 000332–000337. IEEE, 2016.

[22] Anurag Mittal and Nikos Paragios. Motion-based background subtraction using adaptive kernel density estimation. In *Computer Vision and Pattern Recognition, 2004. CVPR 2004. Proceedings of the 2004 IEEE Computer Society Conference on*, volume 2, pages II–II. Ieee, 2004.

[23] Sebastian Nowozin and Jamie Shotton. Action points: A representation for low-latency online human action recognition. *Microsoft Research Cambridge, Tech. Rep. MSR-TR-2012-68*, 2012.

[24] Ruizhi Qiao, Lingqiao Liu, Chunhua Shen, and Anton van den Hengel. Learning discriminative trajectorylet detector sets for accurate skeleton-based action recognition. *Pattern Recognition*, 66:202–212, 2017.

[25] MS Ryoo. Human activity prediction: Early recognition of ongoing activities from streaming videos. *Computer Vision (ICCV), 2011 IEEE International Conference on*, pages 1036–1043, 2011.

[26] Amir Shahroudy, Tian-Tsong Ng, Qingxiong Yang, and Gang Wang. Multimodal multipart learning for action recognition in depth videos. 2016.

[27] Ling Shao, Ling Ji, Yan Liu, and Jianguo Zhang. Human action segmentation and recognition via motion and shape analysis. *Pattern Recognition Letters*, 33(4):438–445, 2012.

[28] Amr Sharaf, Marwan Torki, Mohamed E Hussein, and Motaz El-Saban. Real-time multi-scale action detection from 3d skeleton data. In *IEEE Winter Conf. Appl. Comput. Vision*, pages 998–1005, 2015.

[29] Zheng Shou, Dongang Wang, and S Chang. Action temporal localization in untrimmed videos via multi-stage cnns. In *Proc. Conf. Comput. Vis. Pattern Recog.*, volume 3, 2016.

[30] Yale Song, David Demirdjian, and Randall Davis. Continuous body and hand gesture recognition for natural human-computer interaction. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 2(1):5, 2012.

[31] George R Terrell and David W Scott. Variable kernel density estimation. *The Annals of Statistics*, pages 1236–1265, 1992.

[32] Raviteja Vemulapalli, Felipe Arrate, and Rama Chellappa. Human action recognition by representing 3d skeletons as points in a lie group. *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, pages 588–595, 2014.

[33] Chenxia Wu, Jiemi Zhang, Silvio Savarese, and Ashutosh Saxena. Watch-n-patch: Unsupervised understanding of actions and relations. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4362–4370, 2015.