Spatio-temporal Weight of Active Region for Human Activity Recognition

Dong-Gyu Lee[1](✉) and Dong-Ok Won[2]

[1] Department of Artificial Intelligence, Kyungpook National University, Daegu 41566, Republic of Korea
dglee@knu.ac.kr

[2] Department of Artificial Intelligence Convergence, Hallym University, Chuncheon 24252, Republic of Korea

Abstract. Although activity recognition in the video has been widely studied with recent significant advances in deep learning approaches, it is still a challenging task on real-world datasets. Skeleton-based action recognition has gained popularity because of its ability to exploit sophisticated information about human behavior, but the most cost-effective depth sensor still has the limitation that it only captures indoor scenes. In this paper, we propose a framework for human activity recognition based on spatio-temporal weight of active regions by utilizing human a pose estimation algorithm on RGB video. In the proposed framework, the human pose-based joint motion features with body parts are extracted by adopting a publicly available pose estimation algorithm. Semantically important body parts that interact with other objects gain higher weights based on spatio-temporal activation. The local patches from actively interacting joints with weights and full body part image features are also combined in a single framework. Finally, the temporal dynamics are modeled by LSTM features over time. We validate the proposed method on two public datasets: the BIT-Interaction and UT-Interaction datasets, which are widely used for human interaction recognition performance evaluation. Our method showed the effectiveness by outperforming competing methods in quantitative comparisons.

Keywords: Human activity recognition · Human-human interaction · Spatio-temporal weight

1 Introduction

Recognition of human activity is still developing in computer vision, a field with many applications such as video surveillance, human computer interface and automated driving. In previous studies, the bag-of-words approach or preset motion attributes were commonly used in human activity recognition [10,11,24,36]. Recent deep learning-based representation methods such as 3D convolutional neural networks(CNN) [8], two-stream CNN [27], and multi-stream

C. Wallraven et al. (Eds.): ACPR 2021, LNCS 13188, pp. 92–103, 2022.
https://doi.org/10.1007/978-3-031-02375-0_7

CNN [31] have shown promising results for the human activity recognition problem. However, recognizing human activity accurately remains a challenging task, compared to other aspects of computer vision and machine learning. The use of RGB information imposes limitations on extensibility and versatility because it is often influenced by recording conditions, such as illumination, size, resolution, and occlusion.

With the advent of depth sensors such as Microsoft Kinect, Asus Xtion, and Intel RealSense, instead of using RGB camera, action recognition using 3D skeleton sequences has attracted substantial research attention, and many advanced approaches have been proposed [4,9,18,20,32,35]. Human actions can be represented by a combination of movements of skeletal joints in 3D space. In addition, there has also been major advances in skeleton-based human activity recognition researches [2,3,25,29,34]. They models what happens between two or more people based on their joint information. Although the human skeleton can provide sophisticated information about human behavior, most depth sensors are currently limited to indoor applications with close distance; these conditions are necessary to estimate articulated poses accurately. However, Human activity recognition using articulated poses outdoors could have many more practical applications. Therefore, we address such settings: namely, activity recognition problems where articulated poses are estimated from RGB videos.

In recent studies, deep learning-based approaches have achieved excellent results in estimating the human body joints from RGB videos through pose evaluation [6,7,26]. It has become possible to extract accurate multiple human poses with joint information from RGB video in real time. Because pose estimation and action recognition are closely related problems, some studies simultaneously address these two tasks. A multi-task deep learning approach performed joint 2D and 3D pose estimation from still images and human action recognition from video in a single framework [19]. An AND-OR graph-based action recognition approach utilizes hierarchical part composition analysis [33]. Even though the end-to-end approach has advantages for optimization of the task, it has limited extensibility to videos in varying real-world environments. Furthermore, an approach to research involving interactions, rather than single human actions, methodologically distinct; another problem is that requires the large amount of training data.

In this paper, we propose a novel framework for human activity recognition from RGB video based on spatio-temporal weight of active joints. The proposed framework extracts individual human body joints using publicly available pose estimation method, and recognizes human interaction based on joint motion, local path image, and full-body images with spatio-temporal weight of active region. Therefore, the proposed framework selectively focuses on the informative joints in each frame in an unconditioned RGB video. Figure 1 shows that the interaction regions differ in human activity. In the case of a handshake, hand interaction occurs, but a punch can be understood as head and hand interaction, and a hug as interaction between torso and hand.

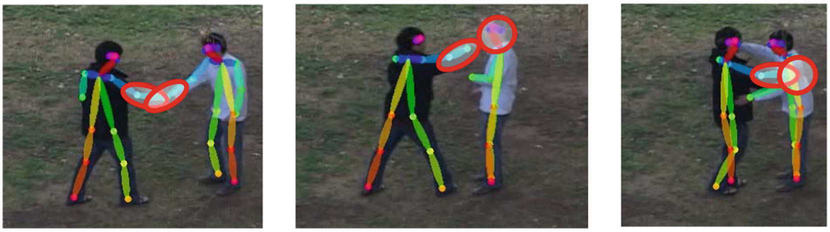


Fig. 1. An example of human body joints with spatio-temporal active region analysis: stretched right hand is interacting three different body part of other person in each activity.

We presents our contributions as follows: first, the proposed framework is based on the RGB video, so it has the benefit that activity recognition can be performed using in the wild without constraints. Second, the spatio-temporal weight of the active region is given to activity relevant motion or poses, that makes the model can focus on important cue of human activity. Third, the experimental result shows the effectiveness of proposed method for the human behavior understanding. This framework allows us to develop a highly extensible application. Furthermore, by not performing separate learning for estimation, detection, and tracking tasks, the proposed framework can be extended to varying datasets in an unconditioned environment.

2 Proposed Method

2.1 Preprocessing

In the most recent studies, video representation through a CNN-based approach has shown good results. We first normalize the RGB pixel data and extract feature vector from images to process input images through CNN. We perform human object detection using Faster-RCNN [22] with the Inception-resnet-v2 network [30]. The detection result provides (x, y) coordinates with height and width. We also perform joint estimation using Part Affinity Fields (PAF) [7] on the same images.

The composition of the estimated joints using PAF is shown in Fig. 2. The PAF provides 18 joints for each human object. In addition, the average of joints 8 coordinate and joint 11 coordinate is designated as point 18 for utilization of the torso information; this is referred to as the hip. For each human subject, we denote each joint as $j_i = \{j_0, ..., j_{18}\}$. The pose estimation in an RGB frame often causes a missing joint. Thus, if the previous n frames have failed to estimate a joint, the value in current frame is used for interpolation and restoration. We use the bounding box to filter out bad results using constraints. First, both the head and torso of each object must be included in the bounding box. If a failure occurs in estimating the head (index 0), the average coordinate of $j_{14}, ..., j_{17}$ is

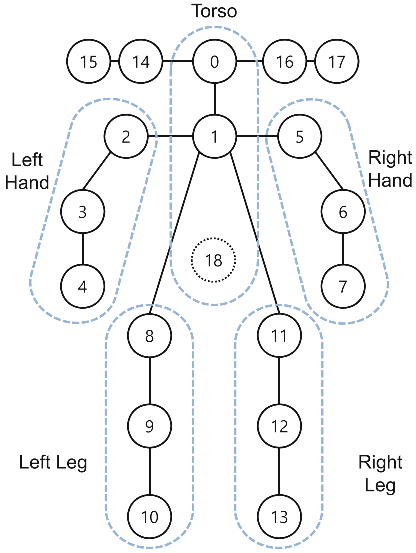


Fig. 2. An illustration of joint indexes from estimated human pose to corresponding five body parts (torso, left hand, right hand, left leg, and right leg).

used as the head position. In this way, noisy objects and poorly estimated joints for interaction can be removed.

In order to consider the local image associated with the body parts in active region, we extract the $(n \times n)$-size image feature from each joint location of index 0 (head position) and 3 (right elbow), 4 (right hand), 9 (right knee), 10 (right foot), 6 (left elbow), 7 (left hand), 12 (left knee), and 13 (left foot). The last fully connected layer of the Inception-resnet-v2 network is used to extract its feature vector, $\mathbf{pf}_j^t$. The input image patches are extracted where the $([x-n/2 : x+n/2], [y-n/2 : y+n/2])$, center is in position (x, y).

2.2 Body Joint Exploitation

We extract four type of joint-based body part features to express the behavior of an individual human. At each time step, for each subject, the 2D coordinates of the 19 body joints are obtained. To consider the characteristics of different behaviors, the motion features were extracted according to the status of the joints. First, we create five body parts using joints from 0 to 18 defined in each frame: right arm p_1 (2, 3, 4), left arm p_2 (5, 6, 7), right leg p_3 (8, 9, 10), left leg p_4 (11, 12, 13), and torso p_5(0, 1, 18). Each number denotes the joint index, and their 2D coordinates are defined as $j_{i^n,(x,y)}$. We then calculate the spatio-temporal weight of five body parts that are created by combining joints and extract the motion feature by using each body part. After defining the five body parts, we calculate the inner angle of each part, θ_{in} as follows:

$$\begin{aligned}\mathbf{a} &= (j_{i^1,x} - j_{i^2,x}, j_{i^1,y} - j_{i^2,y}),\\ \mathbf{b} &= (j_{i^1,x} - j_{i^3,x}, j_{i^1,y} - j_{i^3,y}),\\ \theta &= arccos\left(\frac{\mathbf{a}\cdot\mathbf{b}}{|\mathbf{a}||\mathbf{b}|}\right)\end{aligned} \tag{1}$$

We also calculate the angle between each part using (1). The outer angle θ_{out} denotes the value for connected part, which is calculated using following joint indexes as input in each frame: (1, 2, 3), (1, 5, 6), (1, 8, 9) and (1, 11, 12). The inner angle represents the relative position of the joint inside the body part, and the outer angle represents the shape of the body part. θ_{in} and θ_{out} can express scale-invariant human posture information for the five body parts. This is also an important cue to express the movement of the body parts by changing the position of each joint. For all points in the body parts, the average value of the difference between each previous point and each current point of the sequence, normalized by n length, is used to calculate the motion velocity, v_p^t and acceleration $\hat{v}_j^t$.

2.3 Full-Body Image Representation

We also conduct full-body image-based activity descriptor to capture overall appearance change. The method used here exploits the SCM descriptor used for human interaction recognition [15,16]. Extracting a feature vector from a full-body image has proved useful. Since joint estimation from RGB images includes a failure case, a full-body image can compensate for the missing parts.

From the bounding box of the human object region, we extract weights from the last fully connected layer of the inception-resnet-v2 network. Then we generate a sub-volume for each object $\mathbf{f}_{oi}^t = [p, \delta x, \delta y]$, where p denotes the average of feature vectors in a sub-volume. A series of frame-level image feature vectors of object oi at time t for l consecutive frames, are averaged into a single feature vector. Then, K-means clustering is performed on the training set to generate codewords $\{w_k\}_{k=1}^K$, where k denotes the number of clusters. Each of sub-volume feature $\mathbf{f}_{oi}^t$ is assigned to the corresponding cluster w_k following the BoW paradigm. The index of the corresponding cluster k_{oi}^t is codeword index, which is also the index of the row and column of the descriptor. Here, we should note that, we use the oj_I coordinates from joint estimation to obtain more precise information.

A descriptor using sub-volume features is constructed from each sub-volume of an object $v_{oi}^t = (\mathbf{f}, x, y, k)$. We measure the Euclidean distance between sub-volumes oi and oj. The overall spatial distance between sub-volume oi and the other oj in segment t for $\#pairs$, where $oj \neq oi$, is aggregated as follows:

$$r^t = \frac{1}{2}\sum_{oi}\sum_{oj\neq oi} dist_{oi,oj}^t. \tag{2}$$

The participation ratio of the pair in the segment t is represented using distance difference between sub-volume oi and oj to the global motion activation. The feature scoring function based on sub-volume clustering is calculated as follows:

$$f_p = log\left(\frac{||w_{oi}^t - \mathbf{f}_{oi}^t|| + ||w_{oj}^t - \mathbf{f}_{oj}^t||}{2} + \psi\right). \tag{3}$$

After computing all required values between all sub-volumes, we finally construct the SCM descriptor, as follows:

$$M^t(k_{oi}^t, k_{oj}^t) = \frac{1}{N}\sum_{oi,oi\neq oj}\sum_{1:t}\frac{s_{oi}^t}{\epsilon^t}\frac{r^t}{dist_{oi,oj}^t}f_p(\mathbf{f}_{oi}^t, \mathbf{f}_{oj}^t), \tag{4}$$

where N is the normalization term. The value between oi, oj is assigned to the SCM descriptor using the corresponding cluster index, k_{oi}^t and k_{oj}^t, of each sub-volume. Each of the descriptors is generated for every non-overlapped time step. Therefore, the descriptor is constructed in a cumulative way.

2.4 Spatio-temporal Weight for Classification

In this section, we present the joint based spatio-temporal weight of active region. The basic idea of spatio-temporal weight of active region is the assumption that, when human interaction occurs, the body parts that constitute each action will be of different importance. Spatio-temporal weights of each body part of the person who leads the action and other person have different depends on interactive motions. For example, when person 1 punches person 2, person 1 reaches out to person 2's head and person 2 would be pushed back without motion towards person 1. If person 1 performs a push action, person 2's response will look similar to a punch, but person 1 will reach out to person 2's torso, and two hands will reach out. We try to capture these subtle differences between similar activities, and reflect the difference in the weights. The weight of body parts between persons is calculated as follows:

$$A_{p,t} = S \times \frac{\sum_p^5 |d_{p,t} - d_{p,t-1}|}{|d_{p,t} - d_{p,t-1}|} \tag{5}$$

where d denotes the relative distance between each pair of body parts among the interacting persons. The calculated part weight, $A_{p,t}$ is multiplied by the velocity $wv_p^t = A_{p,t} \times v_p^t$ and acceleration $w\hat{v}_j^t = A_{p,t} \times \hat{v}_j^t$ to determine the weight. The motion feature m_p^t is created by concatenating θ_{in}, θ_{out}, a weighted wv_p^t, and $w\hat{v}_j^t$. The weight is also multiplied by the image patch feature vector from each joint. Since an interacting body part with a high weight plays an important role in the activity, this also gives a high weight to the joint-based image feature extracted from the position of the body part as $\mathbf{wf} = \mathbf{pf} \otimes A_p$.

The overall framework is illustrated in Fig. 3. From a given video, estimated joints are processed through three different streams: joint patch feature extraction, body part motion features with spatio-temporal weight extraction, and

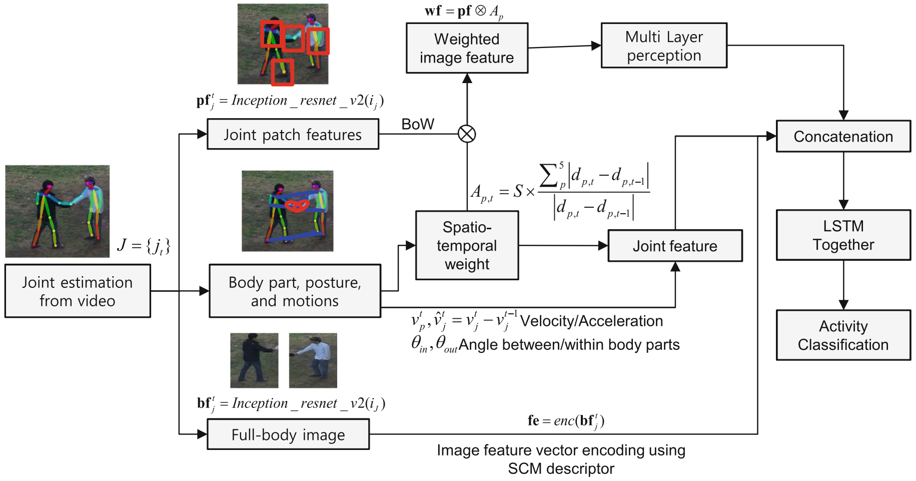

$\mathbf{wf} = \mathbf{pf} \otimes A_p$
Weighted image feature
Multi Layer perception
$\mathbf{pf}_j^t = Inception\_resnet\_v2(i_j)$
Joint patch features
BoW
Concatenation
$A_{p,t} = S \times \frac{\sum_p^5 |d_{p,t} - d_{p,t-1}|}{|d_{p,t} - d_{p,t-1}|}$
$J = \{j_t\}$
Spatio-temporal weight
Joint feature
LSTM Together
Joint estimation from video
Body part, posture, and motions
$v_p^t, \hat{v}_j^t = v_j^t - v_j^{t-1}$ Velocity/Acceleration
$\theta_{in}, \theta_{out}$ Angle between/within body parts
Activity Classification
$\mathbf{bf}_j^t = Inception\_resnet\_v2(i_j)$
$\mathbf{fe} = enc(\mathbf{bf}_j^t)$
Full-body image
Image feature vector encoding using SCM descriptor

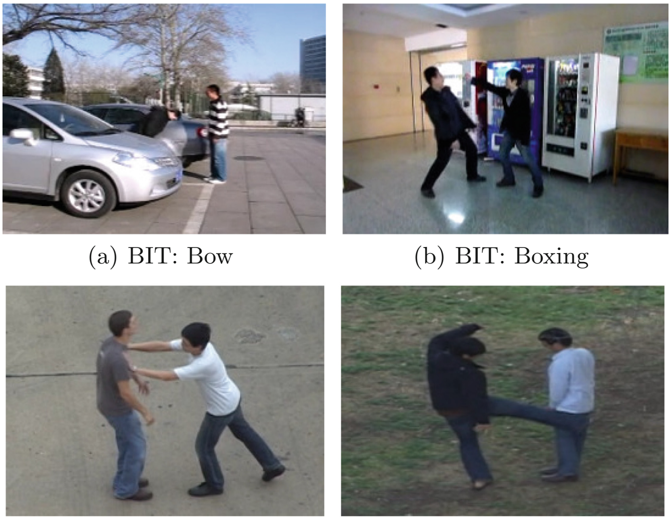
Fig. 3. Illustration of the overall framework combining spatio-temporal weight, joint feature and image feature. The first joint estimation from the video denotes human body joint extraction from RGB input.

full-body image feature extraction. At each step, The generated motion features m_p^t and joint-based weighted image patch features $\mathbf{wf}_p^t$, and SCM descriptor after multi layer perception are concatenated and used as LSTM inputs. The final vector is used as input to LSTM Together, and the activity classification task is the output of the LSTM after processing t segments.

3 Experiment

In this section, we validate the effectiveness of the proposed method on the BIT-Interaction dataset [11] and UT-Interaction dataset [23], which are common and widely used in human interaction recognition research. The performance of the proposed method is shown by comparing the performance with that of the competing methods. In this experiment, the joint estimation was done using PAF [7]. To extract joint patch features and full-body image features, we use the weight of the Inception-resnet-v2 network [30], implemented in Tensorflow [1].

The BIT-Interaction dataset used in the experimental evaluation consists of eight classes of human interactions: bow, boxing, handshake, high-five, hug, kick, pat, and push. Each class contains 50 clips. The videos were captured in a very realistic environment, including partial occlusion, movement, complex background, variying sizes, view point changes, and lighting changes. The sample images of this dataset is shown in Fig. 4(1)–4(b). Both images have the environmental difficulties that are occlusion and complex background. For this dataset, we used a training set with from 1 to 34 index for each class (a total of 272 clips) and the remaining from 35 to 50 index as the test set (128 clips) following official standard in the literature [5,10,13].



(a) BIT: Bow

(b) BIT: Boxing

(c) UT #1: Push

(d) UT #2: Kick

Fig. 4. Sample frames of the BIT-Interaction dataset (a-b), UT-Interaction dataset Set #1 (c) and Set #2 (d).

Table 1. Comparison of the recognition results on the BIT-Interaction dataset

Method	Accuracy (%)
Linear SVM (BoW)	64.06
Dynamic BoW [24]	53.13
MTSSVM [13]	76.56
MSSC [5]	67.97
MMAPM [10]	79.69
Kong *et al.* [12]	90.63
Liu *et al.* [17]	84.37
SCM [16]	88.70
Proposed Method	**92.67**

The experimental results for quantitative comparison on the BIT-Interaction dataset, compared with the competing methods, are shown in Table 1. The table lists the average classification accuracy for eight classes. The proposed method achieved better overall performance over than all the other comparison methods, with 92.67% recognition accuracy for human interaction activity recognition. This result is better than the competing methods. In addition, it shows better performance than SCM-based technique [15], that only considers full-body images. This means that it is better to use the joint-based high-level motion information than to utilize the low-level image features alone.

The UT-Interaction dataset used in the experimental evaluation consist of six classes of human interactions: push, kick, hug, point, punch, and handshake. Each class contains 10 clips for each set. The dataset is composed of two sets of video which were captured in different environments; set #1 and set #2. The set #1 videos were captured in a parking lot background. However, the backgrounds in set #2 of the UT-Interaction dataset consisted of grass and jittering twigs, which could be noise to local patches. We performed leave-one-out cross validation for the performance in the Table 2 and Table 3 as done in previous studies [5,10,16,21,24,28].

Table 2. Comparison of the recognition results on the UT-Interaction dataset (set #1).

Method	Accuracy (%)
Bag-of-Words (BoW)	81.67
Integral BoW [24]	81.70
Dynamic BoW [24]	85.00
SC [5]	76.67
MSSC [5]	83.33
MMAPM [10]	**95.00**
SCM *et al.* [16]	90.22
Mahmood *et al.* [21]	83.50
Slimani *et al.* [28]	90.00
Proposed Method	91.70

Table 3. Comparison of the recognition results on the UT-Interaction dataset (set #2).

Method	Accuracy (%)
Bag-of-Words (BoW)	80.00
Dynamic BoW [24]	70.00
Lan *et al.* [14]	83.33
SC [5]	80.00
MSSC [5]	81.67
MMAPM [10]	86.67
SCM [16]	89.40
Mahmood *et al.* [21]	72.50
Slimani *et al.* [28]	83.90
Proposed Method	**89.70**

Table 2 compares the classification accuracy measured on the UT-Interaction #1. In set #1, the proposed method achieved 91.70 % recognition accuracy. The performance of MMAPM was very high and the proposed method has the second highest performance. On the other hand, our method achieved the highest performance in set #2 as shown in Table 3. This is because set #2 has a noisier background than the set #1, so the proposed method of using human structural characteristics through joint estimation works better than competing methods based on image features only. In real-world scenarios, considering the complexities of environmental change, the proposed method is highly effective.

4 Conclusion and Future Work

Despite numerous studies, it is still challenging and difficult to recognize the complex activity of people in video. However, the complex activity of two or more people interacting with each other requires a higher level of scene understanding than robust image representation. In this study, we showed that robust activity recognition results can be obtained by acquiring joint information of the human that is estimated from RGB videos are informative to understand the human activity. The spatio-temporal weight to actively interacting body parts improve the recognition accuracy than RGB-only methods. This indicates that the relationship between objects plays a key role in complex activity recognition. In addition, the proposed method has high practicality, in the sense that it can overcome the limitations of existing sensors that uses depth information to exploit the skeleton information and increase the possibility of using a common RGB camera. In future research, we intend to expand this work to show robust performance even in interactions involving more people or non-human objects.

Acknowledgment. This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No.2021R1C1C1012590).

References

1. Abadi, M., et al.: TensorFlow: large-scale machine learning on heterogeneous systems (2015). https://www.tensorflow.org/. software available from tensorflow.org
2. Baradel, F., Wolf, C., Mille, J.: Human action recognition: pose-based attention draws focus to hands. In: ICCV Workshop on Hands in Action (2017)
3. Baradel, F., Wolf, C., Mille, J.: Pose-conditioned spatio-temporal attention for human action recognition. arXiv preprint arXiv:1703.10106 (2017)
4. Butepage, J., Black, M.J., Kragic, D., Kjellstrom, H.: Deep representation learning for human motion prediction and classification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 6158–6166 (2017)
5. Cao, Y., et al.: Recognize human activities from partially observed videos. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2658–2665 (2013)
6. Cao, Z., Hidalgo, G., Simon, T., Wei, S.E., Sheikh, Y.: OpenPose: realtime multi-person 2D pose estimation using Part Affinity Fields. In: arXiv preprint arXiv:1812.08008 (2018)

7. Cao, Z., Simon, T., Wei, S.E., Sheikh, Y.: Realtime multi-person 2d pose estimation using part affinity fields. In: CVPR (2017)
8. Ji, S., Xu, W., Yang, M., Yu, K.: 3d convolutional neural networks for human action recognition. IEEE Trans. Pattern Anal. Mach. Intell. **35**(1), 221–231 (2013)
9. Ke, Q., Bennamoun, M., An, S., Sohel, F., Boussaid, F.: A new representation of skeleton sequences for 3d action recognition. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4570–4579. IEEE (2017)
10. Kong, Y., Fu, Y.: Max-margin action prediction machine. IEEE Trans. Pattern Anal. Mach. Intell. **38**(9), 1844–1858 (2016)
11. Kong, Y., Jia, Y., Fu, Y.: Learning human interaction by interactive phrases, pp. 300–313 (2012)
12. Kong, Y., Jia, Y., Fu, Y.: Interactive phrases: semantic descriptions for human interaction recognition. IEEE Trans. Pattern Anal. Mach. Intell. **36**(9), 1775–1788 (2014)
13. Kong, Y., Kit, D., Fu, Y.: A discriminative model with multiple temporal scales for action prediction. In: Proceeding of European Conference on Computer Vision, pp. 596–611 (2014)
14. Lan, T., Chen, T.-C., Savarese, S.: A hierarchical representation for future action prediction. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8691, pp. 689–704. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-10578-9_45
15. Lee, D.G., Lee, S.W.: Human activity prediction based on sub-volume relationship descriptor. In: Proceedings of the IEEE International Conference on Pattern Recognition, pp. 2060–2065 (2016)
16. Lee, D.G., Lee, S.W.: Prediction of partially observed human activity based on pre-trained deep representation. Pattern Recogn. **85**, 198–206 (2019)
17. Liu, J., Kuipers, B., Savarese, S.: Recognizing human actions by attributes. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3337–3344 (2011)
18. Liu, J., Wang, G., Duan, L.Y., Abdiyeva, K., Kot, A.C.: Skeleton-based human action recognition with global context-aware attention lstm networks. IEEE Trans. Image Process. **27**(4), 1586–1599 (2018)
19. Luvizon, D.C., Picard, D., Tabia, H.: 2d/3d pose estimation and action recognition using multitask deep learning. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), vol. 2 (2018)
20. Luvizon, D.C., Tabia, H., Picard, D.: Learning features combination for human action recognition from skeleton sequences. Pattern Recogn. Lett. **99**, 13–20 (2017)
21. Mahmood, M., Jalal, A., Sidduqi, M.: Robust spatio-temporal features for human interaction recognition via artificial neural network. In: International Conference on Frontiers of Information Technology, pp. 218–223. IEEE (2018)
22. Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: towards real-time object detection with region proposal networks. In: Advances in Neural Information Processing Systems, pp. 91–99 (2015)
23. Ryoo, M.S., Aggarwal, J.K.: UT-Interaction Dataset, ICPR contest on Semantic Description of Human Activities (SDHA) (2010). http://cvrc.ece.utexas.edu/SDHA2010/Human_Interaction.html
24. Ryoo, M.S.: Human activity prediction: early recognition of ongoing activities from streaming videos. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1036–1043 (2011)

25. Shahroudy, A., Liu, J., Ng, T.T., Wang, G.: Ntu rgb+ d: a large scale dataset for 3d human activity analysis. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1010–1019 (2016)
26. Simon, T., Joo, H., Matthews, I., Sheikh, Y.: Hand keypoint detection in single images using multiview bootstrapping. In: CVPR (2017)
27. Simonyan, K., Zisserman, A.: Two-stream convolutional networks for action recognition in videos. In: Advances in Neural Information Processing Systems, pp. 568–576 (2014)
28. el houda Slimani, K.N., Benezeth, Y., Souami, F.: Learning bag of spatio-temporal features for human interaction recognition. In: International Conference on Machine Vision, vol. 11433, p. 1143302 (2020)
29. Song, S., Lan, C., Xing, J., Zeng, W., Liu, J.: An end-to-end spatio-temporal attention model for human action recognition from skeleton data. In: AAAI, vol. 1, pp. 4263–4270 (2017)
30. Szegedy, C., Ioffe, S., Vanhoucke, V., Alemi, A.A.: Inception-v4, inception-resnet and the impact of residual connections on learning. In: AAAI, vol. 4, p. 12 (2017)
31. Tu, Z., et al.: Multi-stream CNN: learning representations based on human-related regions for action recognition. Pattern Recogn. **79**, 32–43 (2018)
32. Wang, H., Wang, L.: Modeling temporal dynamics and spatial configurations of actions using two-stream recurrent neural networks. In: e Conference on Computer Vision and Pa ern Recognition (CVPR) (2017)
33. Xiaohan Nie, B., Xiong, C., Zhu, S.C.: Joint action recognition and pose estimation from video. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1293–1301 (2015)
34. Yun, K., Honorio, J., Chattopadhyay, D., Berg, T.L., Samaras, D.: Two-person interaction detection using body-pose features and multiple instance learning. In: 2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 28–35. IEEE (2012)
35. Zhu, W., et al.: Co-occurrence feature learning for skeleton based action recognition using regularized deep LSTM networks. In: AAAI, vol. 2, p. 6 (2016)
36. Zhu, Y., Nayak, N., Gaur, U., Song, B., Roy-Chowdhury, A.: Modeling multi-object interactions using "string of feature graphs." Comput. Vision Image Understanding **117**(10), 1313–1328 (2013)