To appear in the proceedings of 2019 18th International Conference on Computer Analysis of Images and Patterns

Two-stage RGB-based Action Detection using Augmented 3D Poses

Konstantinos Papadopoulos, Enjie Ghorbel, Renato Baptista, Djamila Aouada, and Björn Ottersten

Interdisciplinary Centre for Security, Reliability and Trust (SnT)
University of Luxembourg, Luxembourg
{konstantinos.papadopoulos, enjie.ghorbel, renato.baptista, djamila.aouada, bjorn.ottersten}@uni.lu

Abstract. In this paper, a novel approach for action detection from RGB sequences is proposed. This concept takes advantage of the recent development of CNNs to estimate 3D human poses from a monocular camera. To show the validity of our method, we propose a 3D skeleton-based two-stage action detection approach. For localizing actions in unsegmented sequences, Relative Joint Position (RJP) and Histogram Of Displacements (HOD) are used as inputs to a k-nearest neighbor binary classifier in order to define action segments. Afterwards, to recognize the localized action proposals, a compact Long Short-Term Memory (LSTM) network with a de-noising expansion unit is employed. Compared to previous RGB-based methods, our approach offers robustness to radial motion, view-invariance and low computational complexity. Results on the Online Action Detection dataset show that our method outperforms earlier RGB-based approaches.

Keywords: Action detection · LSTM · pose estimation · action proposals.

1 Introduction

Action detection remains a very challenging topic in the field of computer vision and pattern recognition. The main goal of this research topic is to *localize* and *recognize* actions in untrimmed videos.

Numerous action detection approaches using a monocular camera have been proposed in the literature [19,6,10,25]. Nevertheless, due to the use of effective RGB-based human motion descriptors in a large spatio-temporal volume, these approaches are usually demanding in terms of computational time. As a result, they can be barely adapted to real-world applications such as security and video surveillance [1], healthcare [2,3], human-computer interaction [24], etc. Furthermore, they perform poorly in the presence of radial motion (defined as the perpendicular motion to the image plane), since this information is not encoded in 2D descriptors.

In order to face those challenges, many researchers have exploited the availability of 3D skeletons provided by RGB-D sensors [13,7,4,14]. This high-level representation has the advantage of being compact, largely discriminative and capturing both lateral and radial motion. However, RGB-D cameras present two major limitations in real-life scenarios: First, the acquisition of acceptable depth images and skeletons is only possible under specific lighting conditions. For instance, these devices are not suitable in outdoor settings. Second, acceptable estimation of depth maps is restricted within a very small range.

Inspired by the effectiveness of RGB-D based approaches, we propose to augment 2D data with a third dimension. This is achieved thanks to the tremendous advances in Convolutional Neural Network (CNN)-based approaches, which have made the estimation of relatively accurate 3D skeletons from a monocular video [17,20,28] possible. These estimated compact representations are used in both temporal localization of actions and action classification stages. As in [19], during the action localization phase, hand-crafted features are extracted at each instance using a temporal sliding window. The descriptors used in this stage are chosen to be both spatial with the specific use of *Relative Joint Positions* (RJP) [26], and spatio-temporal with the use of *Histogram of Oriented Displacements* (HOD) [9]. Each frame is classified by a k-Nearest Neighbor (kNN) classifier as *action* or *non-action* forming temporal segments of interest. During the second phase, action recognition is carried out using a Long Short-Term Memory (LSTM) network on the detected action segments. For the estimation of 3D poses, a state-of-the-art CNN-based 3D pose estimator [17] is chosen because of its real-time performance and the encoded temporal coherence. To validate the proposed concept, experiments are conducted on the challenging Online Action Detection dataset [13]. The obtained results show that our approach outperforms other RGB-based methods.

In summary, the contributions of this work are twofold. First, we introduce a novel framework for RGB-based action detection using 3D pose estimation which overcomes the issues of view-variation and radial motion. We argue that this concept can be combined with any 3D skeleton-based action detection approach. Second, a two-stage method for 3D skeleton-based action detection is proposed which uses hand-crafted features in the action localization stage and deep features in the action recognition stage. This two-stage approach is able to offer noise-free action proposals, by removing background frames in the detection stage. Therefore, the recognition part becomes more reliable, regardless of the method used.

The structure of this paper is organized as follows: In Section 2, the background and the motivation of our work are presented. Section 3 offers a detailed description of the proposed method. The experiments are given in Section 4. Finally, Section 5 concludes the paper and discusses future directions and extensions of this work.

2 Background

Given a long sequence of N frames $\mathbf{R} = \{\mathbf{R}_1, ..., \mathbf{R}_N\}$ of continuous activities, the goal of action detection is to find an action label l for each frame at an instance t, such that:

$$l(t) = \mathrm{G}(\mathbf{R}_t), t \in [1, ..., N], \tag{1}$$

where l denotes one of the M actions of interest $\{a_1, ..., a_M\}$ or a background activity a_0 and G is the function which labels the frame $\mathbf{R}_t$.

In general, there are two categories of approaches for finding G; *single-stage* [13,14,4] and *multi-stage* [19,21]. Single-stage approaches are usually able to operate in an online manner, whereas multi-stage ones separate the detection from the recognition step in order to generate mostly noise-free action segments.

In this work, we follow a multi-stage strategy, since they have been shown to be more reliable [19,22]. This strategy, instead of directly estimating G, decomposes the problem into the estimation of two functions G_1 and G_2, such that $\mathrm{G} = \mathrm{G}_2 \circ \mathrm{G}_1$. While G_1 allows the localization of actions via a binary classification, G_2 assigns a label to each frame containing an action. Papadopoulos et al. [19] proposed a two-stage RGB-based action detection concept in which, for finding G_1 during the first stage, 2D skeleton sequences are estimated by a state-of-the-art deep pose detector [18]. To achieve this, a function $f_{2D}(\cdot)$ is estimated to compute J joint positions $\mathbf{P^{2D}}(t) = \{\mathbf{P}_1^{\mathbf{2D}}(t), ..., \mathbf{P}_i^{\mathbf{2D}}(t), ..., \mathbf{P}_J^{\mathbf{2D}}(t)\}$, with $\mathbf{P}_i^{\mathbf{2D}}(t) = (x_i, y_i)$ at time instance t such that:

$$\mathbf{P^{2D}}(t) = f_{2D}(\mathbf{R}_t). \tag{2}$$

Instead of the RGB image $\mathbf{R}_t$, a 2D skeleton at an instance t is used as an input to G_1 to label each frame as action a_+ or non-action a_0, as defined below:

$$e(t) = \mathrm{G}_1((\mathbf{P^{2D}}(t)), \quad e(t) \in \{a_0, a_+\}. \tag{3}$$

During the second stage, the final label l is computed using the estimated 'action' and 'non-action' labels from G_1 and by extracting motion features from RGB images using iDT [27]:

$$l(t) = \mathrm{G}_2(\mathrm{G}_1(\mathbf{P^{2D}}(t)), \mathbf{R}_T), \quad l \in [a_1, ..., a_M]. \tag{4}$$

Despite offering a compact spatio-temporal representation of actions and being relatively fast to compute, the pose estimation approach of (2) lacks 3D information and is dependent on features extracted from RGB images. This results in two issues: the first one is the sensitivity to view variation. Skeletons extracted by the pre-trained model in [18] are solely 2D, therefore any change in body orientation could potentially affect the classification performance. The second challenge is the limited capabilities of describing radial motion.

Taking these observations into consideration, we propose to further strengthen this concept by utilizing 3D skeleton information. For that purpose, we exploit the recent advances in deep learning which have enabled the development of a wide range of 3D pose estimators from monocular cameras [17,16,20]. These

models find a function $f_{3D}(\cdot)$ which estimates, similarly to (2), the 3D pose $\mathbf{P^{3D}}(t)$ at each frame $\mathbf{R_t}$ as:

$$\mathbf{P^{3D}}(t) = f_{3D}(\mathbf{R_t}). \tag{5}$$

By incorporating the third dimension, a two-stage action detection method becomes dependent on the estimated 3D poses as follows:

$$l(t) = \mathrm{G}_2(\mathrm{G}_1(\mathbf{P^{3D}}(t)), \mathbf{P^{3D}}(t)); \tag{6}$$

thus, poses can be aligned using simple linear transformations, resulting in view-invariant representations. In addition, using 3D data, radial motion can be described more effectively.

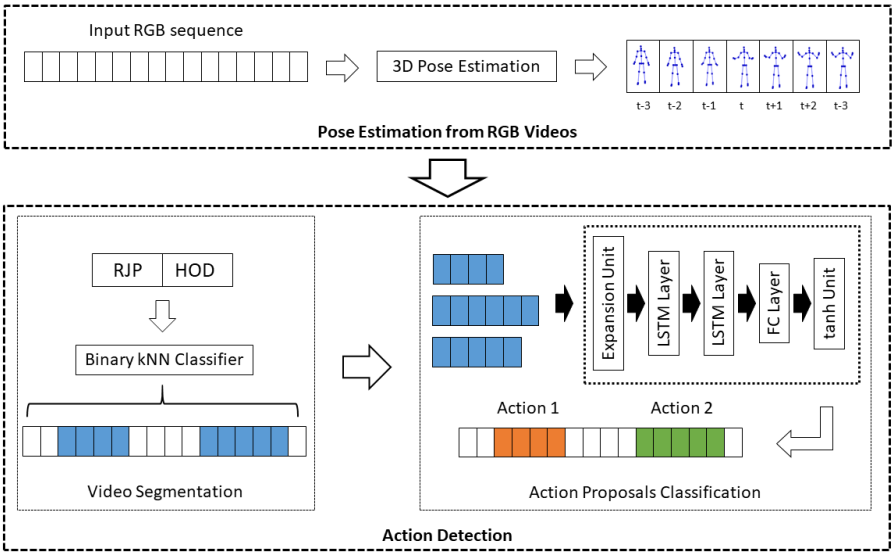

Fig. 1. The proposed model for 2D action detection. Initially, a 3D pose estimator extracts skeleton sequences from RGB frames which are then used for feature generation. In Video Segmentation stage, each frame is classified as action-of-interest or non-action, forming the video segments. These segments are labeled in Action Proposals Classification stage, using a LSTM-based network.

3 Proposed Approach

In this section, we propose to use the estimated 3D skeleton sequences P^{3D} as in (6) for carrying out action detection. To overcome view-point variability, a pre-processing of skeleton alignment is performed by estimating a linear transformation matrix between the absolute coordinate system and a local coordinate

system defined using the spine and the hip joints of the skeleton as in [5,8]. To validate the use of 3D skeletons estimated from RGB images, a 3D skeleton-based approach for action detection is introduced. As shown in Fig. 1, we propose a two-stage approach: during the first stage, the video sequence is segmented into temporal regions of interest and during the second stage the generated video segments are recognized.

3.1 Action Localization

During the first stage, RJP [26] and HOD [9] descriptors are computed in a sliding temporal window around the current frame. The RJP descriptor computes the pairwise relative distances δ^{ij} between the J estimated joints, as shown below:

$$\delta^{ij} = \mathbf{P}_i^{\mathbf{3D}} - \mathbf{P}_j^{\mathbf{3D}}, \quad i, j \in [1, ..., J]. \tag{7}$$

In addition, we use HOD descriptor to describe the orientation of the joint motion in a temporal window around the current frame. To do that, for each pair of Cartesian planes *xy, yz, xz*, a displacement angle θ is computed between consecutive frames as shown below (for the pair xy):

$$\theta_j^{xy} = arctan\left(\frac{d(\mathbf{P}_{j_y}^{\mathbf{3D}})}{d(\mathbf{P}_{j_x}^{\mathbf{3D}})}\right), \tag{8}$$

where $d(\mathbf{P}_{j_x}^{\mathbf{3D}})$ is the accumulated displacement of joint j in x coordinate in a temporal window around the current frame. The final feature vector is a concatenation of all histogram representations of $\theta_j^{xz}, \theta_j^{xy}, \theta_j^{yz}$ for $j \in [1, ..., J]$ in the quantitized 2D space for every pair of Cartesian planes. Both RJP and HOD features are concatenated for each frame, before being injected in the kNN classifier for the action localization. The classifier labels each frame as *action* or *non-action* and window-based patching is applied for filling any gaps in the *action* regions.

3.2 Action Recognition

After the generation of action proposals, the recognition stage takes place. Instead of relying on computationally expensive RGB-based descriptors which use a sizable spatio-temporal volume, such as iDT [27], we propose a compact yet efficient LSTM network using 3D skeletons as input. This end-to-end network allows to simultaneously learn features from sequences and classify actions. As shown in Fig. 1, it is composed of a data expansion unit, a compact LSTM unit with two layers, a fully connected layer and a tanh softmax unit. The expansion unit increases the dimensions of the sequences in order to decouple the noisy factors from the informative ones in the pose sequence. The expansion unit for an input pose $\mathbf{P}^{\mathbf{3D}}$ is defined as follows:

$$\tilde{\mathbf{P}}^{\mathbf{3D}} = tanh(W\mathbf{P}^{\mathbf{3D}} + b). \tag{9}$$

By employing an LSTM unit of two hidden layers, we achieve a favorable balance between performance and compactness. LSTM networks are an advanced Recurrent Neural Network (RNN) architecture [12] which mitigates the problem of the vanishing gradient effect [11], and is capable of handling long-term dependencies. Such architectures have been proven to be effective in action classification [15,23]. Finally, a fully connected layer and a tanh softmax unit are utilized for the label prediction, as shown in Fig. 1.

4 Experiments

In this section, we present the details of our implementation, the experimental settings as well as the obtained results. To evaluate our method, we test it on the challenging Online Action Detection dataset [13].

4.1 Implementation Details

For the extraction of 3D skeletons from RGB sequences, we use the state-of-the-art VNect pose estimator [17]. VNect has two advantages over alternative 3D pose estimators: real-time performance and temporally-coherent skeletons. To achieve real-time performance, VNect is designed using Residual Networks (ResNet). The temporal coherence is ensured by the combination of 2D and 3D joint positions in a unified optimization framework. Temporal smoothing is also applied in order to establish stability and robustness. The available pre-trained model generates 3D poses of 21 joints.

For the localization of action proposals, we empirically choose a sliding window size of 11 and 21 frames for, respectively, computing the RJP and HOD descriptors. Furthermore, we choose to quantitize the 4D space using 8 bins for the computation of HOD features and we train our kNN classifier using 25 nearest neighbors.

In order to maintain a real-time performance without sacrificing the level of abstraction, we empirically use 2 LSTM layers and 256 hidden units per layer. The batch size is fixed to 2, because of the small size of the dataset. Moreover, we use a learning rate of 0.0002 and 200 epochs for the training part.

4.2 Dataset and Experimental Setup

We evaluate the proposed approach on the Online Action Detection Dataset [13] (Fig. 2). This dataset consists of 59 long sequences of 10 continuously performed actions captured by a Kinect v2 device. Thus, this dataset provides RGB and depth modalities along with 3D skeleton sequences. In each sequence, a subject is continuously performing the following activities: *no-activity, drinking, eating, writing, opening cupboard, opening oven, washing hands, sweeping, gargling, throwing trash* and *wiping*. For our experiments, we follow the cross-splitting protocol proposed in [13]. All the reported experiments were conducted on an Intel Xeon E5-1650v3 CPU, clocked at 3.5 GHz. For the evaluation of performance,

Fig. 2. Sample activities from the Online Action Detection dataset [13]

F1-score is used as realized by [13]. F1-score conveys the balance between the precision and the recall and is widely used for the evaluation of action detection concepts.

4.3 Comparison with RGB-based Approaches

Our approach denoted as *CNN-Skel-LSTM* is compared to two RGB-based action detection methods, namely *iDT-SW* [22] and *Skeleton-iDT* [19]. While *iDT-SW* refers to a single-stage method using iDT, *Skeleton-iDT* indicates a double-stage approach using 2D CNN-based skeletons. In Table 1, the obtained per-class and average F1-scores are presented. Similar to Skeleton-iDT, our approach outperforms iDT-SW [22] since the two-stage concept is more effective in trimming action segments.

Our concept also performs better than Skeleton-iDT [19] approach in cases when there is intra-class variability of pose orientation. Recognition results are boosted in 'Drinking' and 'Eating' classes thanks to the use of 3D data, which offer significant view-invariance capabilities. In addition, our approach performs adequately when the subject is not facing the camera (e.g. 'Opening Oven'), by exploiting the radial motion description. However, in 'Opening Cupboard' class, the performance is low, due to the inaccurate estimation of 3D poses. Moreover, in 'Washing Hands' class, the motion is limited only in a small local area where the subject stays still and iDT approach is proven to be more effective in describing such motion.

4.4 CNN-based Skeletons vs. Kinect-based Skeletons

In order to determine the accuracy of the VNect-generated 3D poses for the task of action detection, we compare our approach against *RGBD-Skel-LSTM* case, in which the provided RGB-D skeleton sequences are utilized. The reported (see Table 1) average F1-score of this approach is marginally higher than the

Table 1. F1-score performance of our proposed approach against literature.

	iDT-SW [22]	**Skel-iDT** [19]	**CNN-Skel-LSTM**	**RGBD-Skel-LSTM**
Drinking	0.350	0.218	0.580	0.330
Eating	0.353	0.404	0.601	0.627
Writing	0.582	0.619	0.685	0.646
Opening cupboard	0.453	0.499	0.347	0.536
Opening oven	0.294	0.581	0.624	0.650
Washing hands	0.591	0.759	0.652	0.577
Sweeping	0.467	0.430	0.513	0.673
Gargling	0.505	0.550	0.434	0.723
Throwing trash	0.425	0.573	0.594	0.446
Wiping	0.647	0.802	0.710	0.783
Average	**0.467**	**0.543**	**0.574**	**0.599**

CNN-Skel-LSTM case. The RGBD-based skeleton sequences are more accurately estimated and more informative than the CNN-based ones (25 joints instead of 21). However, the reported results prove the potential of our approach, which goes in pair with the advances of the CNN-based 3D pose estimators.

4.5 Computational Time

One strong feature of the proposed approach is the fast execution time. Indeed, the first stage (segmentation) requires an average execution time of 0.125 second per frame (8 fps) to generate features, while the second stage requires an average of 0.061 second per frame (16.39 fps) to recognize the action. All these measurements were obtained on a single CPU-based implementation and can be further improved.

5 Conclusion

In this paper, we propose an action detection approach which utilizes estimated 3D poses from RGB data. Our concept solves effectively the challenges of radial motion, view dependency and computational complexity. For testing the proposed concept, a novel 3D-based action detection method is introduced. Our method uses estimated 3D skeletons for both determining the temporal regions of interest in a long video sequence and recognizing them, showing, at the same time, competitive detection performance (refer to Table 1). A future direction of our research is to investigate and develop CNN-based descriptors robust to noisy pose estimates, since, in our concept, noisy 3D poses degraded the recognition performance.

6 Acknowledgements

This work was funded by the European Union's Horizon 2020 research and innovation project STARR under grant agreement No.689947, and by the National Research Fund (FNR), Luxembourg, under the project C15/IS/10415355/3D-ACT/Björn Ottersten.

References

1. Baptista, R., Antunes, M., Aouada, D., Ottersten, B.: Anticipating suspicious actions using a small dataset of action templates. In: 13th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISAPP) (2018)
2. Baptista, R., Antunes, M., Shabayek, A.E.R., Aouada, D., Ottersten, B.: Flexible feedback system for posture monitoring and correction. In: Image Information Processing (ICIIP), 2017 Fourth International Conference on. pp. 1–6. IEEE (2017)
3. Baptista, R., Goncalves Almeida Antunes, M., Aouada, D., Ottersten, B.: Video-based feedback for assisting physical activity. In: 12th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISAPP) (2017)
4. Boulahia, S.Y., Anquetil, E., Multon, F., Kulpa, R.: Cudi3d: Curvilinear displacement based approach for online 3d action detection. Computer Vision and Image Understanding (2018)
5. Demisse, G.G., Papadopoulos, K., Aouada, D., Ottersten, B.: Pose encoding for robust skeleton-based action recognition. CVPRW: Visual Understanding of Humans in Crowd Scene, Salt Lake City, Utah, June 18-22, 2018 (2018)
6. Gaidon, A., Harchaoui, Z., Schmid, C.: Actom sequence models for efficient action detection. In: Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on. pp. 3201–3208. IEEE (2011)
7. Garcia-Hernando, G., Kim, T.K.: Transition forests: Learning discriminative temporal transitions for action recognition and detection. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 432–440 (2017)
8. Ghorbel, E., Boonaert, J., Boutteau, R., Lecoeuche, S., Savatier, X.: An extension of kernel learning methods using a modified log-euclidean distance for fast and accurate skeleton-based human action recognition. Computer Vision and Image Understanding (2018)
9. Gowayyed, M.A., Torki, M., Hussein, M.E., El-Saban, M.: Histogram of oriented displacements (hod): Describing trajectories of human joints for action recognition. In: IJCAI. vol. 13, pp. 1351–1357 (2013)
10. Hoai, M., De la Torre, F.: Max-margin early event detectors. International Journal of Computer Vision **107**(2), 191–202 (2014)
11. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural computation **9**(8), 1735–1780 (1997)
12. Kawakami, K.: Supervised Sequence Labelling with Recurrent Neural Networks. Ph.D. thesis, PhD thesis. Ph. D. thesis, Technical University of Munich (2008)
13. Li, Y., Lan, C., Xing, J., Zeng, W., Yuan, C., Liu, J.: Online human action detection using joint classification-regression recurrent neural networks. In: European Conference on Computer Vision. pp. 203–220. Springer (2016)

14. Liu, C., Li, Y., Hu, Y., Liu, J.: Online action detection and forecast via multitask deep recurrent neural networks. In: Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on. pp. 1702–1706. IEEE (2017)
15. Liu, J., Wang, G., Hu, P., Duan, L.Y., Kot, A.C.: Global context-aware attention lstm networks for 3d action recognition. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR). vol. 7, p. 43 (2017)
16. Mehta, D., Rhodin, H., Casas, D., Sotnychenko, O., Xu, W., Theobalt, C.: Monocular 3d human pose estimation using transfer learning and improved cnn supervision. arxiv preprint. arXiv preprint arXiv:1611.09813 **1**(3), 5 (2016)
17. Mehta, D., Sridhar, S., Sotnychenko, O., Rhodin, H., Shafiei, M., Seidel, H.P., Xu, W., Casas, D., Theobalt, C.: Vnect: Real-time 3d human pose estimation with a single rgb camera. vol. 36 (2017), `http://gvv.mpi-inf.mpg.de/projects/VNect/`
18. Newell, A., Yang, K., Deng, J.: Stacked hourglass networks for human pose estimation. In: European Conference on Computer Vision. Springer (2016)
19. Papadopoulos, K., Antunes, M., Aouada, D., Ottersten, B.: A revisit of action detection using improved trajectories. In: IEEE International Conference on Acoustics, Speech and Signal Processing, Calgary, Alberta, Canada (2018)
20. Pavlakos, G., Zhou, X., Derpanis, K.G., Daniilidis, K.: Coarse-to-fine volumetric prediction for single-image 3d human pose. In: Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on. pp. 1263–1272. IEEE (2017)
21. Shou, Z., Wang, D., Chang, S.F.: Temporal action localization in untrimmed videos via multi-stage cnns. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1049–1058 (2016)
22. Shu, Z., Yun, K., Samaras, D.: Action detection with improved dense trajectories and sliding window. In: Workshop at the European Conference on Computer Vision. pp. 541–551. Springer (2014)
23. Song, S., Lan, C., Xing, J., Zeng, W., Liu, J.: Spatio-temporal attention-based lstm networks for 3d action recognition and detection. IEEE Transactions on Image Processing **27**(7), 3459–3471 (2018)
24. Song, Y., Demirdjian, D., Davis, R.: Continuous body and hand gesture recognition for natural human-computer interaction. ACM Transactions on Interactive Intelligent Systems (TiiS) **2**(1), 5 (2012)
25. Sun, C., Shetty, S., Sukthankar, R., Nevatia, R.: Temporal localization of fine-grained actions in videos by domain transfer from web images. In: Proceedings of the 23rd ACM international conference on Multimedia. pp. 371–380. ACM (2015)
26. Vemulapalli, R., Arrate, F., Chellappa, R.: Human action recognition by representing 3d skeletons as points in a lie group. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 588–595 (2014)
27. Wang, H., Schmid, C.: Action recognition with improved trajectories. In: Proceedings of the IEEE international conference on computer vision (2013)
28. Yang, W., Ouyang, W., Wang, X., Ren, J., Li, H., Wang, X.: 3d human pose estimation in the wild by adversarial learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. vol. 1 (2018)