Human Action Recognition from RGBD Videos based on Retina Model and Local Binary Pattern Features

Rawya Al-Akam
Active Vision Group, AGAS
Institute for Computational Visualistics
University of Koblenz-Landau
Universitatsstr. 1
56070 Koblenz, Germany
rawya@uni-koblenz.de

Salah Al-Darraji
Department of Computer Science
University of Basrah
Basrah, Iraq
aldarraji@uobasrah.edu.iq

Dietrich Paulus
Active Vision Group, AGAS
Institute for Computational Visualistics,
University of Koblenz-Landau
Universitatsstr. 1
56070 Koblenz, Germany
paulus@uni-koblenz.de

ABSTRACT

Human action recognition from the videos is one of the most attractive topics in computer vision during the last decades due to wide applications development. This research has mainly focused on learning and recognizing actions from RGB and Depth videos (RGBD). RGBD is a powerful source of data providing the aligned depth information which has great ability to improve the performance of different problems in image understanding and video processing. In this work, a novel system for human action recognition is proposed to extract distinctive spatio and temporal feature vectors for presenting the spatio-temporal evolutions from a set of training and testing video sequences of different actions. The feature vectors are computed in two steps: The **First** step is the motion detection from all video frames by using spatio-temporal retina model. This model gives a good structuring of video data by removing the noise and illumination variation and is used to detect potentially salient areas, these areas represent the motion information of the moving object in each frame of video sequences. In the **Second** step, because of human motion can be seen as a type of texture pattern, the local binary pattern descriptor (LBP) is used to extract features from the spatio-temporal salient areas and formulated them as a histogram to make the bag of feature vectors. To evaluate the performance of the proposed method, the k-means clustering, and Random Forest classification is applied on the bag of feature vectors. This approach is demonstrated that our system achieves superior performance in comparison with the state-of-the-art and all experimental results are depending on two public RGBD datasets.

Keywords

Action Recognition, RGBD videos, Local Binary Pattern, Retina Model, Random Forest.

1 INTRODUCTION

Human action recognition from videos is a very important research topic in image processing, computer vision, and pattern recognition. The human action recognition systems analyze the image sequences or videos by using different methods to predict the type of action and characterize the behavior of persons. This field has represented a challenge because it works with performance issues based on low illumination, perspective effects, and occlusions. It also can handle variations of people, scenes, and motion characteristics [AA13]. An efficient recognition of actions is used in many applications such as video surveillance, robotics human-computer interaction, gesture recognition, behavior analysis, and a variety of systems that involve interactions between persons and computers. All these application domains have own demands, but generally, algorithms have the ability for detecting and recognizing several actions in real time. The designed algorithm should be able to handle different forms of environment and all variations in performing actions because of the different appearance and movement of people [KZP11]. In this paper, we adopt the ideas of spatio-temporal analysis and global features extraction. Global features have been used to characterize textures informa-

tion from body motions on RGBD videos. Our approach represents the human action recognition in four steps: The **First** is high frequency spatio and temporal noise removal and the detection of motion by using a spatio-temporal output from the retina channels. The **Second** uses local binary pattern (LBP) descriptor on different retinal channels such as retina Parvo (parvocellular), Magno (magnocellula) and the combination of both channels Parvo-Magno from both RGB and depth images. The **Third** has computed the histogram from LBP of each frame in the videos and combined all histogram values into a bag of feature vectors, the bag of features algorithm which encodes all the descriptors extracted from each video into a single code. And Finally, in the **Fourth** step, we use k-means clustering and Random Forest (RF) for classifying the different action from videos. The general steps of our system is illustrated in Figure 1, which that represents the structure for our action recognition method.

The rest of the paper is represented as follow, section 2 describes the related work done in this area. Section 3 explain in detail the overview of the proposed method. Section 4 represents the experimentation and results, and finally in Section 5 provides the conclusion.

2 LITERATURE REVIEW

In this literature, the human action recognition is demonstrated from different video actions by using different methods of computer vision and machine learning techniques.

There are several methods used to detect the moving object and extract the important features information from videos, one of the most efficient techniques which is used to detect the motion is Optical Flow, that is used to segment a moving object from video frames background and track it. The moving areas through the video frames can be used to detect the area of interest (motion area) [A07, LD08].

Another research study [BD01] motion energy images (MEI) and motion history images (MHI) are used for the first time as temporal templates to represent human actions and the recognition was done by using seven Hu-moments method. The motion and salient event detection also can be detected by using the Retina model. In this model, a double spatio-temporal filtering used and a good structuring of video data occur as a noise and illumination variation removal and static and dynamic contour enhancement [CsDH10]. The human action recognition is improved by using different techniques of the global features from the moving object. The human action can be represented as a histogram of oriented gradient (HOG) of motion history image (MHI) [HHLH11], the MHI is computed with differential images from successive frames of video sequences, and the HOG features are computed from these motion area. After that, the HOG features are supplied to a support vector machine (SVM) for action classification. In the other work [KZP11] a histogram of the local binary pattern (LBP) was extracted from MHI and MEI as temporal templates to represent human action and the Hidden Markov Models (HMMs) is used to represent and recognize a temporal behavior of the action. The human action recognition can also be described by using the dynamic texture feature descriptors on spatio-temporal domains [CKZP08] and this features are used for human detection to extract LBP-TOP features in spatio-temporal domains from image data, these features are used to detect human bounding volumes and to describe human movements.

3 OVERVIEW OF THE PROPOSED METHOD

This section describes the steps of the proposed system. It details the computation of feature vector from the video sequences and the recognition of the video action. Section 3.1 introduces the pre-processing that is applied to the input RGB and depth videos. Section 3.2 gives a brief description of Bag-of-Features extraction. Section 3.3 explains the k-means clustering. Section 3.4 explains the random forest classification method which is used to compute the recognition accuracy.

3.1 Pre-processing

As the first step in this system is pre-processed to the input data as shown in Figure 2. This data is represented by RGB and depth videos which contain object appearance, shape and motion characteristics. In this work, the video sequence is converted into frames and in turn into lower resolution images of 100×100, for reducing the computational complexity of the system. The depth maps data captured by the Kinect camera are often noisy due to imperfections related to the Kinect infrared light reflections [AaP17]. To remove unwanted signals (noise) in order to preserve the required details and to eliminate the unmatched edges from the depth images, lighting variation, changes in background clutters and so on, the spatial-temporal bilateral filtering and Gaussian filtering are used for pre-processing. This process is done before feature extraction.

3.2 Bag of Features Extraction

The Bag-of-Features (BoFs) [WRLD13] is the most popular technique of feature representation in videos action for recognizing the different human actions. The global feature vectors are computed from the spatio-temporal domain by depending on motion detection and texture descriptor methods. The feature vectors are computed in two steps: Motion detection from spatio and temporal images by using retina model and feature extraction from motion area based on local binary pattern (LBP) descriptor as illustrated in the following.

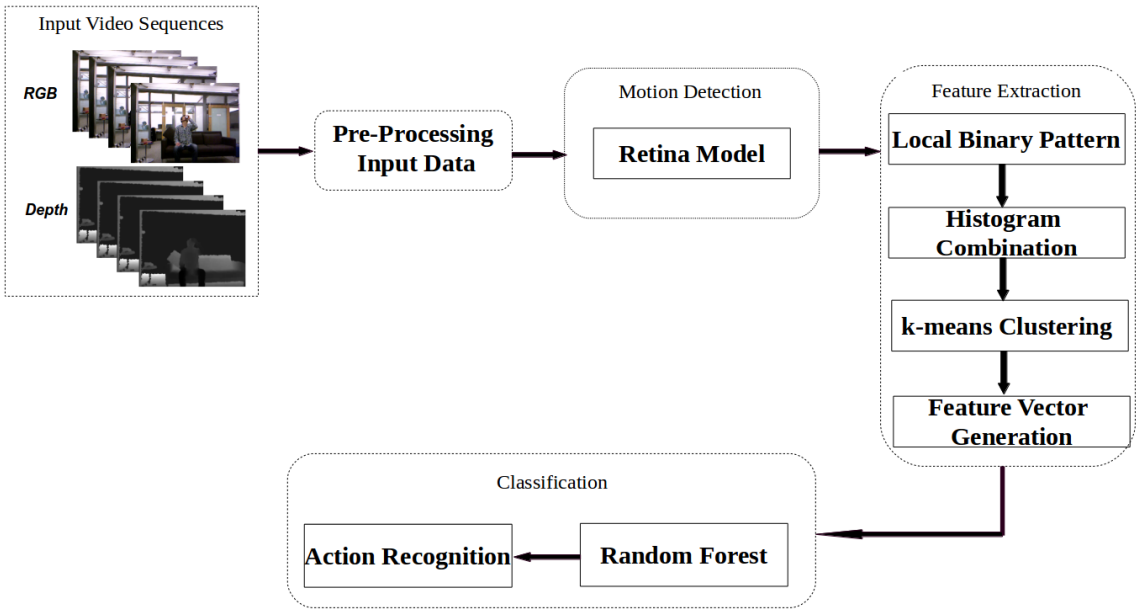

Figure 1: General structure steps of our approach for action recognition using RGB and depth video. Pre-Processing to the input video data, Motion detection, Feature extraction, and Classification.

Figure 2: Pre-processing to the input video data.

3.2.1 Motion Detection using Retina Model

The retina is a non separable spatio-temporal filter model. In order to detect the moving object in each video sequences, the human retina [1] model [CsDH10] is applied to the input data. This model is able to whiten the image spectrum and could remove the high frequency spatio and temporal noise from the images, thus providing enhanced signals for the following processing stages. The retina model decorrelation of details information of spatio and temporal by providing two output video channels, as illustrated below [SBL14]:

- The parvocellular channel (parvo), it is mainly active in the foveal retina area and provide an accurate color vision for visual details with reduced spatio-temporal high frequency noise. Furthermore, objects moving on the retina projection are blurred. The parvo retina output represented in Figure 3.
- The magnocellular channel (magno), it is fundamentally active in the retina environmental vision and send signals related to change events (motion, moving events, etc.). Also it help in improving the visual scene context and object classification from the benefits of local contrast and noise removal. The magno retina output represented in Figure 4.

Figure 3: The retina parvocellular channel (parvo), Left: RGB image and right: Depth image.

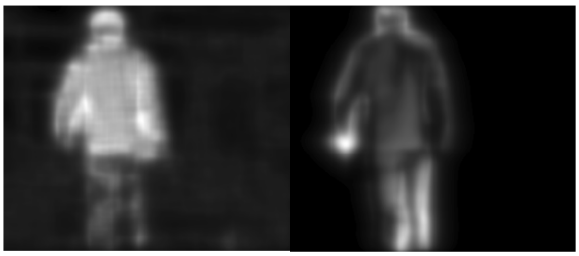

Figure 4: The retina magnocellular channel (magno), Left: RGB image and right: Depth image.

[1] https://docs.opencv.org/3.2.0/d2/d94/bioinspired_retina.html

In this work, the output of the retina model, which represents the motion area, is used to compute the global feature vectors.

3.2.2 *Features Extraction using Local Binary Pattern*

After detecting the motion area, the texture Local Binary Pattern (LBP) descriptor is used to summarize the local structures of the image [Tub17]. The illustration of the original LBP operator is shown in Figure 5, where a LBP operator is calculated by thresholding the differences among the gray value of the center pixel and the neighborhood in a 3×3 grid. Each pixel in the frame is compared with its eight neighbors. The resulting eight values are then considered as an 8-bit binary number. A binary number is obtained by concatenating all these binary codes in a clockwise direction starting from the top-left pixel and its corresponding decimal value is used for labeling. The derived binary numbers are referred to as Local Binary Patterns or LBP codes. The original LBP at the location (x_c, y_c) can be derived from this formula equation (1) as in [CKZP08], which proposed to use elliptic sampling for the x_t and y_t planes:

$$LBP_{x_c,y_c} = \sum_{p=0}^{P-1} s(g_p - g_c)2^p, \quad s(x) = \begin{cases} 1, x \geqslant 0 \\ 0, x < 0 \end{cases} \qquad (1)$$

where g_c is the gray value of the center pixel (xc, yc) and g_p are the gray values at the P sampling points.

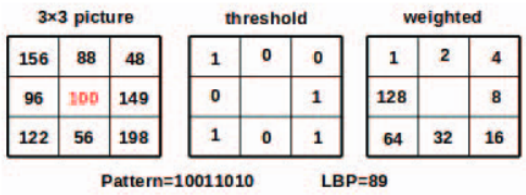

Figure 5: Illustration of the original LBP [VHS15].

The histogram of the encoded motion area is obtained by applying the LBP operator and then used as a texture descriptor for that area in the images. These LBP histograms are combined to represent the spatio-temporal feature vectors from all images in a video.

3.3 k-means Clustering

Clustering is the technique aims to divide the data into groups and each group is constructed by similar data. In simple words, the aim is to separate groups with similar type and assign them into clusters. k-means clustering is a type of unsupervised learning [Yao13], that used with unlabeled data (i.e., data without defined categories or groups) because it has a high efficiency on the data partition, especially in the large dataset. The goal of this clustering method is to find groups in the given data, with the number of groups represented by the variable k. This method is working iteratively to assign each data point to one of k groups based on the features that are provided. Data points are clustered based on feature similarity. The output results of the K-means clustering algorithm are: (1) The centroids of the k clusters, which can be used to label new data, and (2) Labels for the training data (each data point is assigned to a single cluster). In this work, after extracting all LBP features from all RGBD videos. The k-means algorithm is used to generate the dictionary from LPB feature vectors (which is called Bag of features (BOF)) and it was applied on all BoF of training videos sequences, the k represented the dictionary size. The centroids of each cluster are combined to make a dictionary. In this method, we got the best result with a value of k = 400 as a dictionary size. After that, each feature description of the video frame is compared with each centroid of the cluster in the dictionary using Euclidean distance measure (e). Then, the difference (e) was checked, if it is small or features values is close to a certain cluster, the count of that index is increased. Similarly, the other feature description of video frames are also compared and the counts of the respective indices are increased of which the feature description values are closest to which cluster[AaP17, AaP18]. These steps are computed from all the feature vectors of training and testing dataset.

3.4 Action Recognition with Random Forest

To recognize the human actions, the classifiers is needed. The Random Forest (RF) classifiers are used in this work, because of the RF can handle thousands of input variables and large dataset. Moreover, the Random Forest a good performance and outperforms many other machine learning classification algorithms for action recognition [Nab17]. Random Forest was introduced by Breiman [Bre01] as a set of decision trees. For each decision tree in this forest behave like a weak classifier and combined together to compose a strong classifier. During training stages, nodes in the trees are split by randomized selection of features. This selection decreases the error rate in the forest by decreasing the correlation among trees in the forest. Finally, each random tree in the forest grows and predicts the input test data class label. The importance of variables is estimated at the end of training stage [AA13].

4 TEST RESULTS

For our action recognition experiments, we chose to use the MSR DailyActivity 3D Dataset [2] and Online RGBD Action dataset (ORGBD) [3].

The MSR DailyActivity 3D Dataset is a daily activity dataset was recorded by Microsoft and the Northwestern University in 2012, this dataset is captured by a Kinect device and focused on daily activities in the living room, there is a sofa in the scene and the camera was fixed in front of it [Wan12]. This dataset contains 16 actions and 10 subjects; each subject performs each activity in two different poses: *drinking, eating, read a book, call cell phone, writing on a paper, using laptop, using vacuum cleaner, cheer up, sitting, still, tossing paper, playing game, laying down on sofa, walking, playing guitar, stand up, and sit down*. The total number of the activity videos are 320 samples. Some example activities are shown in Figure 6 [AaP18].

The Online RGBD Action dataset (ORGBD) [YLY15] targets for human action recognition (human-object interaction) based on RGBD video data, they are recorded by the Kinect device. Each action was performed by 16 subjects for two times. This dataset contains seven types of actions which captured in the living room: *drinking, eating, using a laptop, picking up a phone, reading phone (sending SMS), reading a book, and using a remote.* as shown in Figure 7 [AaP18]. We compare our approach with the state-of-the-art methods on the same environment test setting, where half of the subjects are used as training data and the rest of the subjects are used as test data.

Figure 6: Sample frames of MSR-Daily Activity 3D Dataset.

The proposed method for computing a texture features by using LBP texture descriptor which applied on retina

[2] http://www.uow.edu.au/w̃anqing/#MSRAction3DDatasets

[3] https://sites.google.com/site/skicyyu/orgbd

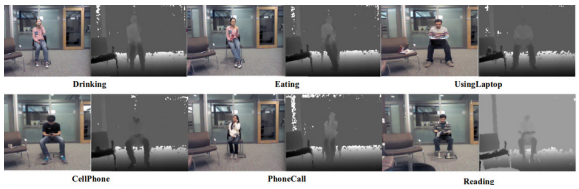

Figure 7: Sample frames of Online RGBD Action Dataset.

motion detection model from spatio-temporal domains. In this work, the experimental results is done on different retina channels as shown in Table 1, and we conclude that the best recognition results are from the LBP features on Magno retina filter (motion detection) channel in compare with other LBP on Parvo and Combination of Parvo-Magno retinal channels. Table 2 and Table 3 show the comparison of accuracy results of our system test and the other state-of-the-art which the used different methods of the MSR-DailyAction 3D datasets and ORGBD Dataset respectively. In our experiments, to compute the feature vectors values three important steps is done: motion detection by using spatio-temporal retina model and the texture feature descriptor LBP is applied on retina output from both RGB and depth channels and finally the histogram from LBP are computed and combined all histogram values to form the bag of feature (BOF). The feature vector size is computed as $2^8 * 2 = 512$ from both RGB and depth channels. To compute the recognition accuracy from our system, k-means clustering and Random Forest (RF) classifier are computed from all feature vector values of different actions and gave the good accuracy rates on both types of datasets.

Retina Channels	MSR3D	ORGBD
Parvo+LBP	83.37%	93.13%
Magno+LBP	**90.21%**	**96.86%**
Parvo+Magno+LBP	81.11%	92.86%

Table 1: Our comparison of recognition accuracy using different retina channels on MSR-Daily Activity 3D (MSR3D) and Online RGBD (ORGBD) Datasets.

Methods	Accuracy
CHAR [ZZSS16]	54.7%
Discriminative Orderlet [YLY15]	60.1%
Feature covariance [PTDZ17]	65.00%
Moving Pose [ZLS13]	73.80%
Parvo+LBP	**83.37%**
Magno+LBP	**90.21%**
Parvo+Magno+LBP	**81.11%**

Table 2: Comparison of recognition accuracy with other methods on MSR-DailyActivity 3D Dataset.

Methods	Accuracy
HOSM [DLCZ16]	49.5%
Orderlet+SVM [YLY15]	68.7%
Orderlet+ boosting [YLY15]	71.4%
Human-Object Interaction[MDDB15]	75.8%
Parvo+LBP	**93.13%**
Magno+LBP	**96.86%**
Parvo+Magno+LBP	**92.86%**

Table 3: Comparison of recognition accuracy with other methods on ORGBD Dataset.

5 CONCLUSION AND FUTURE WORKS

We have presented a new method for human action recognition based on Retina model and local binary pattern (LBP) descriptor. Its main idea is to capture spatio-temporal relation of moving object depending on the spatio-temporal filtering retina model and applied the LBP texture feature extractor on the moving object from each image in video actions to compute a bag of important feature information. These feature values are tested using random forest classification machine learning methods. Our system is tested on two different public RGBD dataset and its achieved superior performance in comparison with the state-of-the-art approaches. These datasets are MSR Daily Activity 3D and Online RGBD (ORGBD) and the recognition accuracy on this dataset reached to 90.21% and 96.86% respectively. For the future work, we will apply a convolution neural network on our system.

ACKNOWLEDGEMENTS

This work was partially supported by Ministry of Higher Education and Scientific Research (MHESR), Iraq, and University of Koblenz-Landau, Germany.

6 REFERENCES

[A07] Deborah A. An Adaptive Optical Flow Technique for Person Tracking Systems. *Pattern Recognition Letters*, 25:43–53, 2007.

[AA13] Ilktan Ar and Yusuf Sinan Akgul. Action recognition using random forest prediction with combined pose-based and motion-based features. *ELECO 2013 - 8th International Conference on Electrical and Electronics Engineering*, pages 315–319, 2013.

[AaP17] Rawya Al-akam and Dietrich Paulus. RGBD Human Action Recognition using Multi-Features Combination and K-Nearest Neighbors Classification. *(IJACSA) International Journal of Advanced Computer Science and Applications*, 8(10):383–389, 2017.

[AaP18] Rawya Al-akam and Dietrich Paulus. Local and Global feature Descriptors Combination from RGB-Depth Videos for Human Action Recognition. In *7th International Conference on Pattern Recognition Applications and Methods (ICPRAM)*, number Icpram, pages 265–272, 2018.

[BD01] Aaron F Bobick and James W Davis. The recognition of human movement using temporal templates. *IEEE Trans. Pattern Anal. Mach. Intell.*, 23(3):257–267, 2001.

[Bre01] Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

[CKZP08] P.A. Crook, V. Kellokumpu, G. Zhao, and M. Pietikainen. Human Activity Recognition Using a Dynamic Texture Based Method. *Procedings of the British Machine Vision Conference 2008*, pages 88.1–88.10, 2008.

[CsDH10] ABenoit A Caplier sinpgfr, B Durette, and J Herault. Using Human Visual System Modeling for Bio-Inspired Low Level Image Processing. *Computer Vision and Image Understanding.*, 114, no. 7:758 – 773, 2010.

[DLCZ16] Wenwen Ding, Kai Liu, Fei Cheng, and Jin Zhang. Learning hierarchical spatio-temporal pattern for human activity prediction. *Journal of Visual Communication and Image Representation*, 35:103–111, 2016.

[HHLH11] Chin Pan Huang, Chaur Heh Hsieh, Kuan Ting Lai, and Wei Yang Huang. Human action recognition using histogram of oriented gradient of motion history image. *Proceedings - 2011 International Conference on Instrumentation, Measurement, Computer, Communication and Control, IMCCC 2011*, pages 353–356, 2011.

[KZP11] Vili Kellokumpu, Guoying Zhao, and Matti Pietikinen. Recognition of human actions using texture descriptors. *Machine Vision and Applications*, 22(5):767–780, 2011.

[LD08] Z. Lin and Larry S. Davis. A Pose-Invariant Descriptor for Human Detection and Segmentation. *Computer Vision–ECCV 2008*, 5305:423–436, 2008.

[MDDB15] Meng Meng, Hassen Drira, Mohamed Daoudi, and Jacques Boonaert. Human-object interaction recognition by learning the distances between the object and the skeleton joints. *2015 11th IEEE Int. Conf. Work. Autom. Face Gesture Recog-*

nit., pages 1–6, 2015.

[Nab17] Mohsen Nabian. A Comparative Study on Machine Learning Classification Models for Activity Recognition. *Journal of Information Technology & Software Engineering*, 07(04):4–8, 2017.

[PTDZ17] Alexandre Perez, Hedi Tabia, David Declercq, and Alain Zanotti. Feature covariance for human action recognition. *2016 6th International Conference on Image Processing Theory, Tools and Applications, IPTA 2016*, 2017.

[SBL14] Sabin Strat, Alexandre Benoit, and Patrick Lambert. Retina enhanced bag of words descriptors for video classification. In *European Signal Processing Conference*, 09 2014.

[Tub17] Tuba Milan Simian Dana Tuba, Eva. Support Vector Machine Optimized by Firefly Algorithm for Emphysema Classification in Lung Tissue CT Images. In *25. International Conference in Central Europe on Computer Graphics, Visualization and Computer Vision WSCG*, 2017.

[VHS15] Domonkos Varga, László Havasi, and Tamás Szirányi. Pedestrian detection in surveillance videos based on CS-LBP feature. *2015 International Conference on Models and Technologies for Intelligent Transportation Systems, MT-ITS 2015*, (June):413–417, 2015.

[Wan12] Junsong Wang Jiang Liu Zicheng Wu Ying Yuan Junsong Wang, Author. Mining Actionlet Ensemble for Action Recognition with Depth Cameras. In *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1290–1297, 2012.

[WRLD13] Jun Wan, Q Ruan, W Li, and S Deng. One-shot learning gesture recognition from RGB-D data using bag of features. *Journal of Machine Learning Research*, 14:2549–2582, 2013.

[Yao13] Yang Yu Yongqing Xu Hong Lv Weiming Li Zhao Chen Xiaoyun Yao, Yukai Liu. K-SVM: An effective SVM algorithm based on K-means clustering. *Journal of Computers (Finland)*, 8(10):2632–2639, 2013.

[YLY15] Gang Yu, Zicheng Liu, and Junsong Yuan. Discriminative orderlet mining for real-time recognition of human-object interaction. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 9007:50–65, 2015.

[ZLS13] Mihai Zanfir, Marius Leordeanu, and Cristian Sminchisescu. The Moving Pose: An Efficient 3D Kinematics Descriptor for Low-Latency Action Recognition and Detection. *The IEEE International Conference on Computer Vision (ICCV)*, pages 2752–2759, 2013.

[ZZSS16] Guangming Zhu, Liang Zhang, Peiyi Shen, and Juan Song. An Online Continuous Human Action Recognition Algorithm Based on the Kinect Sensor. *Sensors (Basel, Switzerland)*, 16(2):161, 2016.