Proceedings of the 2018 IEEE
International Conference on Robotics and Biomimetics
December 12-15, 2018, Kuala Lumpur, Malaysia

A child caring robot for the dangerous behavior detection based on the object recognition and human action recognition*

Qiang Nie[1], Xin Wang[1], Jiangliu Wang[1], Manlin Wang[1], Yunhui Liu[1]

***Abstract*— In this paper, a child caring robot is developed for detecting some dangerous behavior performed by child in the domestic environment based on the human action recognition and object recognition technologies. A human behavior is an interactive process between human and objects. Therefore, three factors need to be considered: the engaged objects, human actions and the relationship between human and the engaged objects. In our application scenario, a correlative filter is proposed to improve the stability of the object recognition with human interference. For the human action recognition, a new motion encoding method by using the Euclidean distant matrix (EDM) between joints is introduced and a convolutional neural network is utilized. Evaluation on the Northwester-UCLA dataset verified the effectiveness of this method when action categories are small. The proposed action recognition method is simple and efficient, which is crucial for online behavior detection. Extensive experiments in the real physical world for detecting the behavior of eating allergic fruit and touching/playing with electrical socket have achieved good performance.**

I. INTRODUCTION

With the development of robotics technology, robots have started to enter people's daily life and coexist with ordinary people other than professionals. Among them, service robot or intelligent system, which has been used in the business environment serving as a guide or the home environment serving as a companion for the elders or children, has attracted a lot of attention. And people are expecting more and more from them. For example, considering the aging society, people would like to develop an elderly caring robot to help take care of the elders. In this task, the robot may need to recognize and track a certain elder, understand his or her language, as well as to understand the gestures, actions or behavior in case the sound is too small and vague. However, understanding human behavior is an under-explored task. A behavior, which is a dynamic process of the interaction between objects and human, is consist of the engaged objects, human actions, the semantic relationship between the engaged objects and human. Complex spatiotemporal information needs to be considered in such a process. Attribute to the development of deep learning technologies, object recognition and detection have achieved great progress and have high accuracy in real world applications. However, the recognition of human action is still an unsolved problem.

*This work was supported by the T-Stone Robotics Insititute of the Chinese University of Hong Kong

[1]Qiang Nie, Xin Wang, Jiangliu Wang, Manlin Wang, and Yunhui Liu is with Faculty of Engineering, the Department of Mechanical and Automation Engineering, The Chinese University of Hong Kong, Shatin, N.T., Hong Kong, China qnie@mae.cuhk.edu.hk

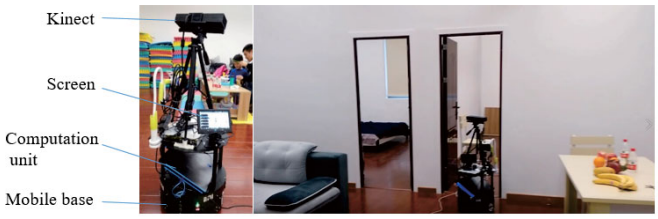

Fig. 1. The child caring robot (left) and our home environment lab (right)

Human action recognition has been researched for decades and many previous works were concentrate on finding a proper way to represent the human action from the RGB images or depth data. However, these data always has a big change of presentation in the images when observed from different viewpoints. Thereby, human action recognition based on the skeleton data attracts more and more efforts. Skeleton data can explicitly describe the human body structure and the variation of human pose. As Johansson [1] who spent decades on exploring the problem of how human eyes tracking moving objects mentioned, the skeleton with constant bone lengths is one of the most effective ways to represent human motions.

To represent human action, two kinds of information need to be considered, the spatial structure of the human body and the variation of the body structure along with time. In the past, researchers used the coordinates of joints in each frame as a feature of human action and concatenated them into a vector. However, such a feature is highly dimensioned and suspicious to the viewpoints. Later, people started to use the displacement of joints [2], [3], [4], [5], which can be calculated between different joints within one frame or between the same joints located in the different frames. The former displacement of joints contains the spatial information and the latter displacement contains the temporal information. But concatenating the displacements into a one-dimensional vector is still not the best way to represent the human action as the spatial-temporal information is not well structured and represented.

Recently, deep learning method has been proved to be effective in many areas, especially in image recognition. Therefore, some researchers [6], [7], [8], [9], [10], [11], [12], [13], [14] introduced deep learning method into action recognition. Among these methods, methods based on the convolutional neural network [6], [7], [12], [13] reported better performance. These methods visualize human action with images and transfer the problem of human action recognition

978-1-7281-0377-8/18/$31.00 © 2018 IEEE

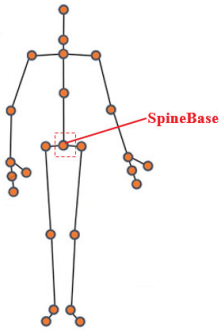

Fig. 2. The Kinect skeleton model

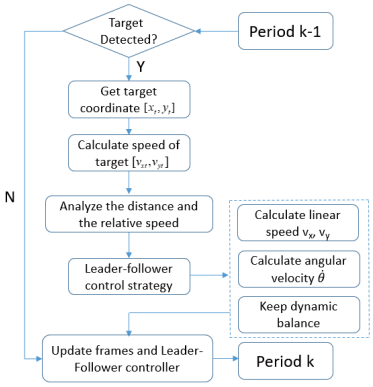

Fig. 3. The Leader-Follower strategy for human tracking

into a problem of image classification. Motion visualization is a technique to encode motion features as some visible graphics, such as 2D color images or gray images. In these images, the motion features within a frame are encoded into a column or row and another axis of the encoded image is the time or frame axis. Motion visualization method is able to make use of the good performance of CNNs in image recognition. However, as the features in each frame are concatenated into a row or column, this method is strong in representing the temporal variation but the special structure information of the human body can hardly be utilized. Thus, how to balance the spatial structure information and temporal variation to generate a more descriptive motion image is still underexplored.

In this paper, we developed a child caring robot to explore the problem of human behavior detection based on the human action recognition and object recognition technologies. This robot is used for detecting some dangerous behavior performed by child in the domestic environment. A new motion encoding method is introduced by using the Euclidean distant matrix (EDM). Evaluation on Northwester-UCLA dataset verified the effectiveness of this method when action categories are small. The proposed method is simple and efficient, which is important for online behavior detection. Extensive experiments in the real physical world for detecting the behavior of eating allergic fruit and touching/playing with electrical socket achieved good performance.

II. Overview of the Child Caring Robot System

Our child caring robot mainly consists of three parts: a wheeled mobile base, a 3D sensor for observing and a computation unit for information processing and decision making, as shown in Fig. 1. The 3D sensor we used is Kinect which can provide the RGB data, the depth data and the skeleton data for our system. These data will then be used by computation unit for target tracking, object recognition, human action recognition and dangerous behavior judging. Based on the information of sensor, the system has to decide whether the target person is tracked and whether a dangerous behavior is happening. If a dangerous behavior is detected, the robot will give an alarm to inform the parents or other supervisors.

A. Human tracking strategy

To guarantee the robot located at a proper observing position, the robot has to track the position of the target child and keep an appropriate distance with the child considering the safety issue and the effective working distance of 3D sensor. Our tracking strategy is to keep the target locates in the middle of the field of view (FoV) and within a distance range.The position of joint SpineBase, as shown in Fig. 2, is utilized as the 3D position of human body for target tracking. We assume that the target person will not leave the ground. Thus, only the position in the horizontal plane need to be considered. As the effective working distance of Kinect is about 1m $\sim$ 3.5m, the distance between our robot and the target child is selected to be 2m $\sim$ 2.5m. If the child moves too fast and the robot lost tracking of the child, the robot will keep rotating itself in the same place until the child was found again. Once the child is observed, we adopt a Leader-Follower strategy as [15] to track the child. The control strategy is shown in Fig. 3.

B. Computation unit

The working principle of the computation unit is shown in Fig. 4. From the 3D sensor, three data streams are obtained, the RGB stream, the depth stream, and the skeleton stream. The RGB image is used for recognition of the target person and some dangerous objects, such as the allergic fruit for the child and electrical sockets. By mapping the recognized object from the RGB image to the depth image, the 3D position of an object can be calculated from the corresponding point cloud. The position of the human body can be obtained from the skeleton data. If a person is the target person and the 3D position of the person is known, our robot can track the target person based on the aforementioned moving strategy. On another aspect, the skeleton data is also used to extract the motion features of the human body. These features will be encoded into visible images for action recognition. Based on the results of action and the object recognition, the dangerous behavior performed by the child can be detected.

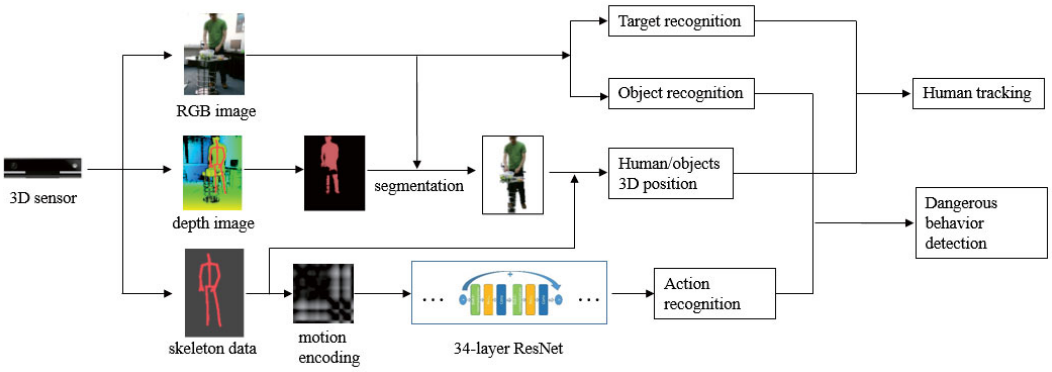

Fig. 4. The working principle of the computation unit which is used to analyze the input data obtained from the 3D sensor and make decisions

III. Dangerous objects detection

In object detection module, we utilized YOLONet2 [16] network structure which has an advantage in the processing speed. This network has 19 full convolution layers and uses the global average pooling layer for prediction. Batch normalization is utilized to speed up the convergence rate. As our training data is limited, we applied a series of methods, such as random cropping and color shifting to expand the training samples. As the original network is designed for classification, the fully connected layers and the last convolution layer is removed when doing object detection. Three 3×3 convolutional layers with 1024 filters each followed by a final 1×1 convolutional layer with 125 output channels are added.

However, there are two challenges in our application scenario. Taking the detection of electrical sockets as an example, most of the time, the electrical socket appears very small in the image, as shown in Fig. 5. If we use the whole image as input, the sockets can hardly be detected. Another challenge is the occlusion caused by the human interference. Occlusion will result in two kinds of failure cases of dangerous behavior detection. One case is missing of sockets when they are completely occluded by the human body. In this case, behavior of touching or playing with socket cannot be detected. The second case is the wrong calculation of the socket's position when they are partially occluded by the human body. For example, in the right image of Fig. 5, the socket in the black box is partially occluded by a human leg. If we map the region of the socket in the RGB image to the depth image, the corresponding region in the depth image will contain both points of the leg and the socket, which causes the calculated position of the socket locates near the leg rather than the wall far behind the human body. In this case, error detection of behavior may happen.

To solve the challenge of small resolution of sockets in the whole image mentioned above, we split the original image into four small sub-images with 1/5 overlap. The socket detection algorithm is implemented parallelly in each of the sub-image and non-maximal suppression is used to deal with redundant detection bounding boxes. As for the detection of objects occluded by the human body, we propose to use a correlative filter to predict and track the position of occluded objects. If the detected image patch has high similarity with the template patch but the position is different, then the detected object is treated as a moving object. If the similarity between the target image patch and the template patch is low and the position varies a lot, the object is occluded and the position recorded before the occlusion happens will be used. Otherwise, the object can be detected as in the static environment. The template patches are selected as all the bounding box regions of detected objects and will be updated every 2 seconds in our application.

Fig. 5. The electrical socket presented in our image

IV. Human action recognition

In this part, we will introduce a skeleton-based human action recognition method based on motion visualization technology and the convolutional neural network (CNN). With the success that CNN has achieved in the image recognition and classification, researchers are thinking about whether the advantages of CNN can be utilized for human action recognition. Motion visualization, which encodes the motion features of human body into images, is one of those explores that have been proved to be effective. Therefore, the problem of human action recognition is transformed into a problem of image classification. The key issue here is to decide what features to encode and how to encode these features into an image. Most of the current visualization method cannot represent the special structure information of the human body. According to our research experience on this topic, the structure information is much more important than the dynamic variation of joint. Hence, we proposed to use the EDM matrix which defined as a matrix of the

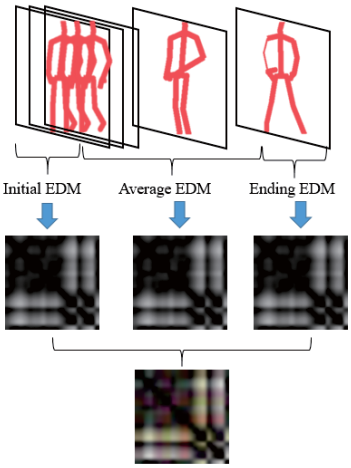

Fig. 6. The proposed approach to encode a human action into a color EDM image

pairwise Euclidean distances between joints as denoted in (1), where λ is used to normalize the distance value within the range $[0, 1]$. p_i and p_j is the position of joint i and j. N is the total number of considered joints. EDM has been widely used in many areas, such as modal analysis and structure representation. In [17], a method of recovering the 3D human pose from a single RGB image is proposed based on the joint EDMs. It has been verified that richer information about pairwise correlations between body joints can be captured by the EDM. More importantly, it is coordinate-free, invariant to rotation, translation, and reflection.

$$EDM_{i,j\in N}(i,j) = \frac{1}{\lambda}\|p_i - p_j\|_2 \tag{1}$$

An EDM can be converted into a gray image by using the equation $\mathcal{I}(i,j) = 255EDM(i,j)$, where $\mathcal{I}(i,j)$ is the value of pixel (i,j) in the encoded image $\mathcal{I}$. In this way, the human pose in each frame can be encoded into an EDM image and an action sample corresponding to a series of EDM images. However, the CNN can only input an image once a time and a single EDM contains no temporal information. Thus, we split the EDM images into three parts, as shown in Fig. 6. The first a few frames are selected to calculate the initial EDM which is an average result of the selected frames. So is the ending EDM which is calculated by using the last several frames in an action sequence. The remain frames are averaged together to generate the average EDM. The obtained 3 kinds of EDMs corresponding to the initial pose, the average pose and the ending pose of an action. The average operation here encodes the temporal information and can help the calculated EDM more robust to noisy skeletons. Further, these three kinds of gray EDM images are synthesized into a single color EDM image as denoted in (2).

$$\mathcal{I}_{RGB}(i,j) = [\mathcal{I}_{ini}(i,j), \mathcal{I}_{avg}(i,j), \mathcal{I}_{end}(i,j)] \tag{2}$$

With the proposed method, each human action can be mapped to an RGB image that contains both the structure information and temporal information. These images are feed into a CNN for action classification. We selected the ResNet [18] to extract high-level features and learn the action patterns. Using f denotes the function of the whole neural network, the training process of the CNN is to find proper parameters θ that has the highest accuracy of predicting the action labels $\hat{y} = f(\mathcal{I}, \theta)$.

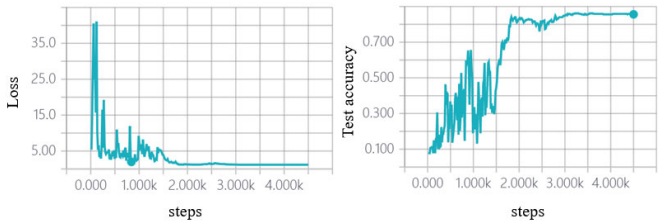

Fig. 7. Variation curve of loss and testing accuracy during training

In our application, we need to recognize particular actions: eating fruit and touching/playing with socket. As the action categories are relatively small, we trained our data together with the Northwestern-UCLA dataset. This dataset contains 10 action categories: pick up with one hand, pick up with two hands, drop trash, walk around, sit down, stand up, donning, doffing, throw, carry. Totally, 1494 motion sequences are collected from 3 different viewpoints. Our network is trained with the cost function defined as the summation of cross-entropy between training data and the model distribution and parameter regularization, as given by:

$$L = -\mathbb{E}_{\mathcal{I},y\sim\hat{p}_{data}} \log p_{model}(y|\mathcal{I},\theta) + R(\theta) \tag{3}$$

where $\mathbb{E}$ denotes the expectation, $p_{model}(y|\mathcal{I},\theta)$ means the distribution of the model and $R(\theta)$ is the regularization part of trainable parameters. The training process is shown in Fig. 7. The testing accuracy converged after about 3000 steps. To evaluate the proposed method, we compared the performance of the proposed method with some state-of-the-art methods based on the Northwestern-UCLA dataset, as shown in Table I.

TABLE I
RESULTS ON NORTHWESTERN-UCLA DATASET(CROSS-VIEW PROTOCOL[19])

Feature	Method	Accuracy(%)
Hand-crafted	HOJ3D 2012[20]	54.50
	LARP 2014[21]	74.20
	TLDS 2018[22]	74.6
RNN	HBRNN-L 2015[9]	78.52
	TS-LSTM 2017[10]	89.22
	Denoised-LSTM 2018[23]	80.25
	Multi-task RNN 2018[24]	87.3
CNN	**Ours**	**86.13**

As we can see from the Table I, our method achieved a competitive accuracy compared with the state-of-the-art methods. Though the performance of our method is not the best among the reviewed methods, our method is quite simple and efficient, which is critical for online action detection in

real scenarios. The proposed method is strong in describing the structure relationship of the human joints but weak to present the temporal variation. However, compared with most methods reviewed in Table I, our method still can achieve a better performance, which verified the importance of the body structure information in human action recognition.

V. Dangerous Behavior Detection

In our application scenario, we considered two kinds of dangerous behavior that are likely performed by the child. The first behavior is eating allergic food and the second is touching/playing with electrical socket in the domestic environment. These two behavior is common for young children since they are lack of ability to control themselves and curious about everything. To depict a behavior, three factors are needed to be considered: the engaged objects, human action, and the relationship between human and the engaged objects. That is to say, a behavior can be defined as (4), where $\mathcal{O}, \mathcal{A}, \mathcal{R}$ denote the engaged objects, human action and the relationship between human and the engaged objects. In the former sections, we have explored the problems of object recognition and human action recognition.

$$\mathcal{B} = (\mathcal{O}, \mathcal{A}, \mathcal{R}) \tag{4}$$

The training data used in human action recognition is well trimmed. While in behavior detection, all the events happen continuously. To solve this problem, we used a temporal sliding window to select the frames that considered in one detection step. Only the frames in the temporal window are utilized for action recognition. The relationship between human and the engaged objects is defined as the Euclidean distance between the human hand and the engaged objects. Based on these definitions, a behavior can be well described. Take the behavior of eating allergic food for example, this behavior can be defined as:

$$\mathcal{B}_{EAF} = (\mathcal{O}_{allergicfruit}, \mathcal{A}_{eating}, D_{hand\&\mathcal{O}} \leq d) \tag{5}$$

Within the temporal window T, if the three factors mentioned in (5) are detected, the behavior of eating allergic food is detected.

VI. Experiments in the Real Scenarios

We test the behavior detection algorithm in our home environment lab, as shown in Fig. 1. This lab is decorated like an ordinary home as possible as we can. Some kinds of fruit are put on the dining table and several power strips or sockets are distributed beside the sofa, at the corner, behind the table and on the wall above the TV stand. The child is encouraged to walk freely and touch anything in the lab. Mango is assumed to be the allergic food for the testers. Considering the Kinect can run at a speed of 20 frames per second in our application, the width of the temporal window is selected as 3s which equivalent to 60 frames. The sliding step of the temporal window is 0.5s, i.e., 10 frames. An Alienware notebook with Intel i7 CPU, GTX 1060 GPU, and 16 GB memory is used as the computation unit.

Fig. 8. Results of dangerous objects recognition

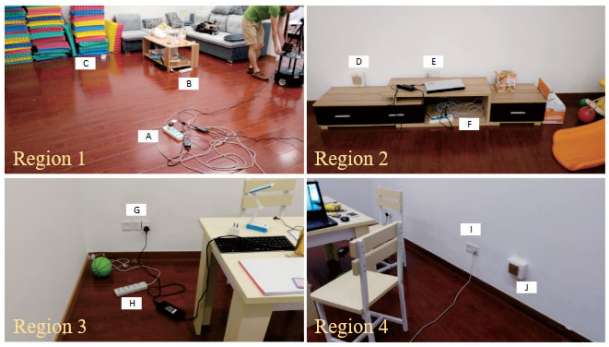

Fig. 9. Distribution of electrical sockets

A. Results of object recognition

At first, we test the performance of our child caring robot in recognizing dangerous objects that are defined as mango and electrical socket in our application scenarios. As shown in Fig. 8, our robot can detect the dangerous objects stably from different viewpoints, with different background, and under different light conditions.

B. Detection of dangerous behavior

As eating mango is a relatively simple behavior, we choose the behavior of touching/playing with socket to evaluate the performance of our robot in detecting dangerous behaviors. In the test, the home environment lab is divided into four different regions and two or three electrical sockets are placed in each region, as shown in Fig. 9. The tester is required to touch each socket 5 times from different angles. Thereby, occlusions happened from time to time. The success rate of each region is recorded for different test subjects, as shown in Table. II.

TABLE II
The Detection Results of the Behivor of Playing with/Touching Socket

Area	subject 1	subject 2	subject 3	subject 4	average
1	75%	80%	80%	70%	76.25%
2	86.7%	67%	86.7%	93.3%	83.41%
3	70%	90%	50%	100%	77.5%
4	100%	90%	90%	80%	90%
average	82.93%	81.75%	76.68%	85.83%	81.8%

In our experiments, an overall success rate of 81.8% is achieved. One of the reasons that the success rate fluctuate

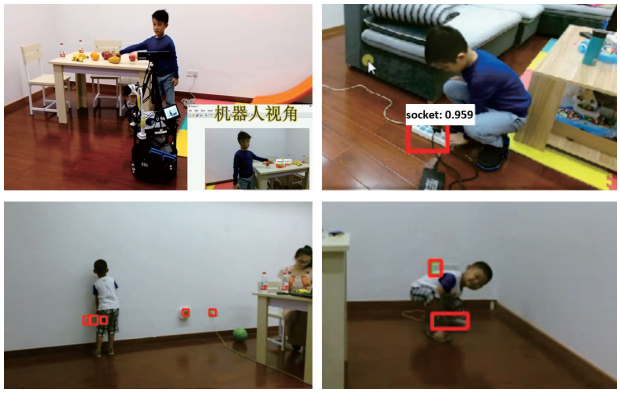

Fig. 10. Some successful detection samples of the dangerous behavior performed by child

among different subjects is that we didn't constrain the touching angle.To make sure the behavior is performed naturally, the tester is encouraged to touch the socket from any angle they want. So in some cases, the electrical socket is occluded seriously by the tester, which is a common phenomenon when the service robot applied in our daily life. Another reason that causes the detection failure is the moving speed of the human. The Kinect cannot capture the skeleton data when a person moves too fast. Without accurate skeleton data, it's hard for our robot to recognize the action performed by human and the relationship between the human and the engaged objects. Generally, a good performance is achieved for the detection of dangerous behavior performed by the child, which verified the effectiveness of our efforts in solving problems of object recognition and human action recognition in the highly dynamic home environment.

CONCLUSIONS

This paper presents a caring robot used for detecting dangerous behavior performed by the child based on artificial intelligent technologies. In particular, the methods for object and human action recognition by using the convolutional neural network are discussed. These different technologies are combined together for the child's behavior detection in the real world. Although some decent performance has been achieved, there are still many problems need to be solved to bring a service robot to our daily life. For example, the approach to keep tracking of human and achieve accurate pose estimation in such a highly dynamic environment. Developing a more robust algorithm to occlusion and viewpoint variation is another problem that needs more efforts.

REFERENCES

[1] G. Johansson, "Visual perception of biological motion and a model for its analysis," *Perception & psychophysics*, vol. 14, no. 2, pp. 201–211, 1973.

[2] H. Rahmani, A. Mahmood, D. Q. Huynh, and A. Mian, "Real time action recognition using histograms of depth gradients and random decision forests," in *Applications of Computer Vision (WACV), 2014 IEEE Winter Conference on*. IEEE, 2014, pp. 626–633.

[3] J. Wang, Z. Liu, and Y. Wu, "Learning actionlet ensemble for 3d human action recognition," in *Human Action Recognition with Depth Cameras*. Springer, 2014, pp. 11–40.

[4] X. Yang and Y. L. Tian, "Eigenjoints-based action recognition using naive-bayes-nearest-neighbor," in *Computer vision and pattern recognition workshops (CVPRW), 2012 IEEE computer society conference on*. IEEE, 2012, pp. 14–19.

[5] X. Yang and Y. Tian, "Effective 3d action recognition using eigenjoints," *Journal of Visual Communication and Image Representation*, vol. 25, no. 1, pp. 2–11, 2014.

[6] M. Liu, H. Liu, and C. Chen, "Enhanced skeleton visualization for view invariant human action recognition," *Pattern Recognition*, vol. 68, pp. 346–362, 2017.

[7] P. Wang, Z. Li, Y. Hou, and W. Li, "Action recognition based on joint trajectory maps using convolutional neural networks," in *Proceedings of the 2016 ACM on Multimedia Conference*. ACM, 2016, pp. 102–106.

[8] Y. Du, Y. Fu, and L. Wang, "Skeleton based action recognition with convolutional neural network," in *Pattern Recognition (ACPR), 2015 3rd IAPR Asian Conference on*. IEEE, 2015, pp. 579–583.

[9] Y. Du, W. Wang, and L. Wang, "Hierarchical recurrent neural network for skeleton based action recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1110–1118.

[10] I. Lee, D. Kim, S. Kang, and S. Lee, "Ensemble deep learning for skeleton-based action recognition using temporal sliding lstm networks," in *2017 IEEE International Conference on Computer Vision (ICCV)*. IEEE, 2017, pp. 1012–1020.

[11] C. Li, Y. Hou, P. Wang, and W. Li, "Joint distance maps based action recognition with convolutional neural networks," *IEEE Signal Processing Letters*, vol. 24, no. 5, pp. 624–628, 2017.

[12] Q. Ke, S. An, M. Bennamoun, F. Sohel, and F. Boussaid, "Skeletonnet: Mining deep part features for 3-d action recognition," *IEEE signal processing letters*, vol. 24, no. 6, pp. 731–735, 2017.

[13] Q. Ke, M. Bennamoun, S. An, F. Sohel, and F. Boussaid, "A new representation of skeleton sequences for 3d action recognition," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2017, pp. 4570–4579.

[14] C. Li, Q. Zhong, D. Xie, and S. Pu, "Skeleton-based action recognition with convolutional neural networks," in *Multimedia & Expo Workshops (ICMEW), 2017 IEEE International Conference on*. IEEE, 2017, pp. 597–600.

[15] W. Ye, Z. Li, C. Yang, J. Sun, C.-Y. Su, and R. Lu, "Vision-based human tracking control of a wheeled inverted pendulum robot," *IEEE transactions on cybernetics*, vol. 46, no. 11, pp. 2423–2434, 2016.

[16] J. Redmon and A. Farhadi, "Yolo9000: better, faster, stronger," *arXiv preprint*, 2017.

[17] F. Moreno-Noguer, "3d human pose estimation from a single image via distance matrix regression," in *Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on*. IEEE, 2017, pp. 1561–1570.

[18] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[19] J. Wang, X. Nie, Y. Xia, Y. Wu, and S.-C. Zhu, "Cross-view action modeling, learning and recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 2649–2656.

[20] L. Xia, C.-C. Chen, and J. Aggarwal, "View invariant human action recognition using histograms of 3d joints," in *Computer Vision and Pattern Recognition Workshops (CVPRW), 2012 IEEE Computer Society Conference on*. IEEE, 2012, pp. 20–27.

[21] R. Vemulapalli, F. Arrate, and R. Chellappa, "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*, 2014, pp. 588–595.

[22] W. Ding, K. Liu, E. Belyaev, and F. Cheng, "Tensor-based linear dynamical systems for action recognition from 3d skeletons," *Pattern Recognition*, vol. 77, pp. 75–86, 2018.

[23] G. G. Demisse, K. Papadopoulos, D. Aouada, and B. Ottersten, "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.

[24] H. Wang and L. Wang, "Learning content and style: Joint action recognition and person identification from human skeletons," *Pattern Recognition*, vol. 81, pp. 23–35, 2018.