Activity recognition for ASD children based on joints estimation

Dongxu Gao[1], Zhaojie Ju[1*], Yingfeng Fang[1], Jiangtao Cao[2], Chenguang Yang[3], Honghai Liu[1]
[1]*School of Computing, University of Portsmouth, UK*
Email: zhaojie.ju@port.ac.uk
[2]*School of Information and Control engineering,Liaoning Shihua University, China*
[3]*College of Engineering, Swansea University, UK*

***Abstract*—Human motion recognition is a trending topic and could be applied in many areas, the motion estimation of ASD children is more challenging because of the high uncertainty of their activities, we thus introduced a novel method which is designed for estimating the upper joints and recognising their special motions, we verified the proposed method on our recorded ASD children dataset and adult dataset, the experimental results show the proposed method is effective on the dataset.**

***Index Terms*—Joints estimation, Activity recognition, ASD dataset.**

1. Introduction

The activity recognition usually means to learn about the activities from video sequences and identify similar actions with machine learning method. Human activity recognition is very important in computer vision research area today as it can be applied in many fields including the surveillance system, human-machine interfaces, video indexing, virtual coaching, VR games, patient monitor system[1],and some motion related application [14].

In general, the activity recognition system needs to have the ability to track the human motion [18] and recognise complex human motions from a continuous video sequence or from only a static image. Such a system usually can be classified into two approaches according to the input data in a contactless method, which is known as the computer vision based activity recognition [22], instead of a wearable-based method [13]. As the wearable-based method could limit the human pose and affect the possible motion, we only focus on the vision-based method in this paper.

Different features have been used in activity recognition methods. Michel [19]adopted a tracking method to capture the articulated motion including the 3D position and orientation with two RGB-D cameras. Spatio-temporal and bag-of-word features are used to represent human motion in many works of literature. Semantic features are used to explain the meaning of a motion. For example, it is understandable that a car appears on a road while it is not acceptable for some people that a giraffe appears in a kitchen.

Although lots of researchers work hard on the activity recognition using different methods, many factors, including the diversity of appearances, the variation of the camera angles, background clutter, illumination changes in a scene, and occlusion by other objects, pose a challenge on the performance of the activity recognition. Some research methods were proposed to handle some of these issues or one aspect of them. For example, to handle the illumination changes, depth information based method [25] were used for more accurately estimating the human pose. Multi-view based method [7] was used in the activity recognition system to avoid the negative effect of occlusion.

2. Related work

Activity recognition has been researched for many years and some review papers [3, 17, 26, 30] suggested the features for effectively representing

motions play a key role in this area. Hassaballah [8] provided an overview of image feature range from detection, description to feature matching which are fundamental components for handling computer vision issues. The interest point and local image features contributed to represent object patterns in a static image, but failed to represent features for a dynamic image sequence. Space-time interest point was raised as a response. Such a method [16] performed well for some simple motions such as walking and running.

To improve the human pose estimation on a single image, one way is to extending a static recognition method with utilising a regularisation on the body parts over time by using a probabilistic graphical model [4]. This method typically represents human body parts corresponding to different major body parts such as head, shoulders, elbows and hands. By forming these node parts into a graph, a kinematic method is usually adopted to capture the inter-part relationships. The Pictorial structure model (PSM) [10] allowed the inference to estimate the possible poses over the pose space.

The ordinal pattern is normally seen as the low-level feature. A middle-level feature, which was integrated into an orderlet [29] character, was proposed to represent the relationships among joints and shape information respectively on both skeletons and depth maps. While High-level pose features (HLPF) were introduced for encoding spatial and temporal relations of human skeleton joints [9]. Dense trajectories feature were also been proved of an excellent performance in some activity recognition datasets [27]. In addition, spatio-temporal features have been applied in the activity recognition for representing the action with a dense feature set, while Shi [23] introduced a fast random sampling method on a local part model to speed up the computational efficiency.

Cheron argued that the representation of human pose dominates the performance of the action recognition and introduced a pose-based scheme, which aggregated the descriptor based on the human pose for tracking human body parts [5]. This supervised method extremely relied on the annotation of the human body parts and hand-crafted feature extraction, which needs considerable relative skills and lots of restless work, thus puts lots of burden on the human.

There is little previous work with enough annotations contribute for pose estimation, Johnson and his colleagues [10] proposed a method to estimate the human pose with only inaccurate annotation. Some computer vision related work had proved that the approach is useful, for example, the collaborative LabelMe object annotation system [21] and utility data annotation [24] still benefit when obtaining data from some inexperienced annotators.

Apart from the methods for estimating poses in a single image and the spatio-temporal feature representing methods, a scheme, for continuous motion recognition, based on the static image feature is also important. With accurate skeleton information such as the position and the angles, a skeletal representation is needed for encoding the features with a dynamic time warping. Vemulapalli [25] explored a method through modelling the 3D geometric relationships among body parts with 3D space rotations and translations. To estimate the 3D human pose by optimising the joints over the set of the manifold with a particle-based optimisation algorithm, the low-dimensional manifold [7] was analysed to emphasize the importance of a successful scheme for pose estimation in videos and handle the temporal coupling across time.

3. Method

In this section, we will firstly introduce the method with sampling strategy used for estimating the key joints information, then we introduce a skeletal descriptor method to represent the joints, to make the scheme work more effectively, a Gaussian Process is adopted for mapping between different dimensional space, then an on-line method is utilised for recognizing the real-time activities of the child.

A. Feature model

Human pose feature, especially the body joints, is essential for activity recognition. A deformable mixture-of-parts model is used to represent the body parts for a single image because of the computational efficiency and considerable property [28]. The upper body part is modelled as a set of major joints which are the head, neck, two shoulders, two elbows, and two wrists (or hands). Theses joints contribute significantly the performance of the upper body motions. A pictorial structure

model which uses the tree-type graph with nodes is introduced to represent each joint position and orientation. For some specific camera angles, self occlusion could happen. To handle this issue, a cluster method is used to classify each body part with annotated ground truth T_i for one of the n training images. The problem is formulated as a maximum-likelihood problem through calculating the highest probability:

$$\hat{\Theta} = \underset{\Theta}{argmax} \prod_{i=1}^{N} \max_{j=1}^{k} P(T_i|\Theta_j) \tag{1}$$

There are k pose clusters in total, $P(T_i|\Theta_j)$ is the posterior probability of a particular pose for an image I, which is defined as:

$$P(p|I) \propto f(I|p)f(p) = \prod_{i} f(r_i|l_i) \prod_{(l_i,l_j \in E)} (l_i|l_j) \tag{2}$$

l_i denotes the 2D position and orientation, which is one element of the set $p = \{l_1, l_2, ...l_n\}$,r_i is the corresponding image region, the prior term defines the prior probability of a configuration. This has two main advantages: on one side, it can help to overcome the ambiguous image data, on the other side, it limits the model from the plausible human configurations when the kinematic limits of the body is learned.

In addition, a linear SVM classifier is used for each body parts, the classifier is bootstrapped with some negative samples of other body regions and non-body regions for training. The responses can be computed for each body part is:

$$p(r_i|l_i, \Theta_i) \propto \max_{j=1...n} w_j \Phi(r_i) \tag{3}$$

In which w_j is the weight vector for component j, $Phi(r_i)$ is the feature vector from the image region r_i. The maximum value allows us to determine the appearances mode with the highest confidence.

B. Dense sampling

For a more efficient computation, we use a random sampling strategy for the denser patches, let us look an image with size $n \times m$ for instance, the number of possible sampled patches is n^4 which is explained in [15], besides, it is proved that the performance could be improved with randomly sampled patches for each image [20]. Based on this, reducing the number of sampled points for an individual frame and still maintain an efficient sampling density for representing the features.

C. Descriptor

With the skeleton information obtained, we use a skeletal representation method to represent the body part. The method was proposed in a previous paper[25], which mainly considered a whole body parts, we slightly change the method for represent only the upper body. When a pair of body parts is given, their relative geometry is described as e_m and e_n, which denote the eight joints and oriented rigid body parts respectively, the starting point $(e^n_{m1}(t))$ and end point $(e^n_{m2}(t))$ of each part can be represented in a local coordinate system at time instance t.

$$\begin{bmatrix} e^n_{m1}(t) & e^n_{m2}(t) \\ 1 & 1 \end{bmatrix} = \begin{bmatrix} R_{m,n}(t) & \overrightarrow{d}_{m,n}(t) \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 0 & l_m \\ 0 & 0 \\ 0 & 0 \\ 1 & 1 \end{bmatrix} \tag{4}$$

$$\begin{bmatrix} e^m_{n1}(t) & e^m_{n2}(t) \\ 1 & 1 \end{bmatrix} = \begin{bmatrix} R_{n,m}(t) & \overrightarrow{d}_{n,m}(t) \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 0 & l_n \\ 0 & 0 \\ 0 & 0 \\ 1 & 1 \end{bmatrix} \tag{5}$$

where $R_{m,n}(t)$ and $R_{n,m}(t)$ are the rotations, $\overrightarrow{d}_{m,n}(t)$ and $\overrightarrow{d}_{n,m}(t)$ are the translations, these are measured in the local coordinate system. More detailed information for representing the joints we refer the [25].

D. Feature mapping

From the joint information in motion capture data, we use a Gaussian Process regressions, which is a straightforward extension of Gaussian Mixture Model, to map a low-dimensional space from a high-dimensional space.The equation 7 indicates the back process of the mapping.

$$x = f_a \sim GP(m(y), k(y, y^{'})) \tag{6}$$

$$y = g_a \sim GP(m(x), k(x, x^{'})) \tag{7}$$

f_a denotes the mapping from high-dimensional to low-dimensional space, while g_a denotes the inverse process, where m represents the mean and k denotes the covariance functions. M_a is learned

to model the temporal transitions between effective motions for an action-specific manifold.

$$x_t = M_a(x_{t-1}) \sim GP(m(x-1), k(x_{t-1}, x_{t-1}^{'})) \tag{8}$$

Instead of using a single state space, a set of action-specific manifolds is considered, A_c is defined as a set $\{a_1, a_2, ..., a_{|A|}\}$,which denotes the action classes, where we consider to learn an action-specific manifold for all the classes. As the manifolds only utilised the joint space, the representation of a body pose is determined by $y_a = (r, t, \Theta_a)$, (r, t) is a vector indicates the global orientation and position, Θ denotes the joint angles.

E. Classification

As our aim is to recognise both static motions and dynamic motions, we introduce a scheme which can estimate both motions, for the single image, a pose set regards to a tree-graph which including the 2D coordinates for representing the body parts is defined as:

$$P_s = p^i = (x^i, y^i) \tag{9}$$

Then we formulate the estimation issue as a minimization problem with the cost $C(I, P_s)$:

$$C(I, P_s) := \Sigma_i \phi_i(I, p^i) + \Sigma_{i,u} \varphi_{i,u}(p^i - p^u) \tag{10}$$

F. Real-time activity recognition

For a real-time activity recognition system, it needs to predict a continuous video sequences with reliable scores of different classes. The frame-level score is defined as:

$$R(I_t) = \sum_{a_1=1}^{a_{|A|}} \alpha_m R_m(I_t) \tag{11}$$

$R_m(I_t)$ denotes the response of a orderlet on the frame I_t,while α_m is the corresponding weight, which decides the balance between the positive and negative votes. It is clear that different types of actions have various properties such as the action speed and the durations. These make it difficult to determine the size of a fixed-length window. The temporal smoothness with adaptive smoothing window length is introduced for a reasonable result. The main concept is to maintain a reliable voting score for $t - th$ frame.

$$S(V_t) = max(0, S(V)_{t-1} + R(T_t)) \tag{12}$$

$S(V_t)$ denotes the score at time t, if the value is greater than 0, it means the current action is continuing, on the contrast, if the value is less than 0 or equal to 0, there is no action is happening. Then the value will be reset to 0 and forecasts that a new action will start.

4. Experiments and Discussion

The main aim of this work is to recognise 11 activities of ASD children for assisting the therapist in curing the patients, to verify the efficiency of the proposed activity recognition framework, we test the method on two datasets, one is captured with 8 ASD children, and the other is collected with 15 adult people doing the same motions.

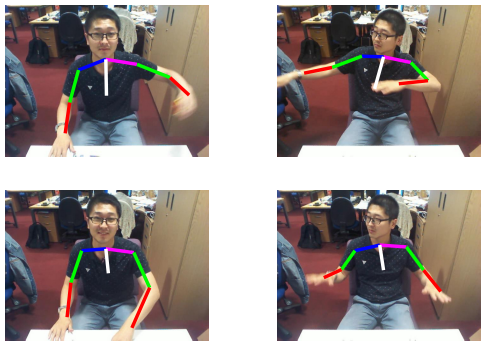

Figure 1: The result samples for representing the joints

As the motions of ASD children are not exactly the same as the normal people, we first labelled the motions from recording dataset and ask the therapist to decide whether the motions are proper or not, all the labelled data used for training in our work is checked by the therapist to make sure the performance is convincing. The uncertainty of motions done by ASD children makes the recording work more difficult, thus makes times of each motion in the dataset are different. To demonstrate the method is effective, we also verify the method on adult motion dataset.

In this section, we report the results on our recorded dataset within only our method as the method is designed only for the specific purpose. For protecting the privacy of the ASD children, we

only show image results from the adult dataset. Figure 1 shows some result samples from our dataset. The joints information is estimated with our method and the table 1 shows the average accuracy of the estimating results for the upper body major joints.

Our dataset is extremely more challenging than the existing dataset as there are more unpredicted factors when recording the ASD children behaviour dataset. The average accuracy is still kept at 84.7%.

TABLE 1: The average accuracy of the joints(%)

joints	accuracy
neck	87.9
shoulders	84.6
elbows	76.6
wrists	78.2
upper body	96.3

The figure 2 indicates the confusion matrix for estimating the motions on both our datasets. The average accuracy for predicting the motions is 85.9%, which can be seen as an acceptable result.

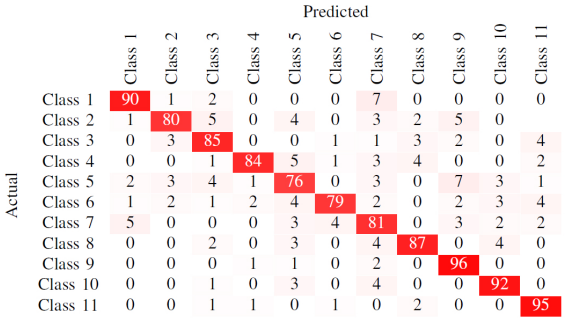

Figure 2: The recognition results

In this paper, we mainly focus on the atomic motions including namely waving the hand, drinking, and moving a toy etc., which are defined by the therapist, these movements indicate a stable coordination pattern among the skeleton joints, each activity normally contains a joint set order [29].For example, when the child is doing the drinking motion, the child first hold the cup from the table, move the cup to his/her mouth, hold on for some seconds and put the cup back to the table. We believe that if skeleton joints especially the wrist, elbow and shoulder are estimated accurately, they will provide us with an effective feature for modelling the motion and recognizing the motion. Thus the accuracy of joints information comes from the very first step for estimating both the continuous motions and some static motions.

We have evaluated our method on both ASD dataset and adult dataset, and have presented the joint estimating results and motion estimation results. The accurate estimation of joint could provide an excellent classification result even using a linear SVM classification method, which implies the importance of the joints estimation for our datasets.

5. Conclusion

In this paper, we propose a novel activity recognition method which is designed especially for recognising the ASD children motion, we run our algorithm on both the ASD dataset and the adult dataset to verify the effectiveness of the proposed method. The experimental results show that our approach performs well on the datasets we collected. Our work can be applied in a real-time system with an integrated gaze estimation method [2] and hand gesture recognition[12][11][6] for assisting therapist to communicate with ASD children, the research confirms that the correct classification of the body parts leads to a significant improvement in estimating the human joints, and the pose estimation can benefit from the accurate human joints. How to estimate the joints from less body annotation or inaccurate body parts annotation will be our next work, we will also compare our method on different challenge datasets to provide a more convincing result in the future work.

Acknowledgments

This work is supported by the EU Seventh Framework Programme (grant no. 611391), China Scholarship Council (grant no. 201508060340), Research Project of State Key Laboratory of Digital Manufacturing Equipment & Technology China (grant no. DMETKF2017003), and National Natural Science Foundation of China (grant no. 51575412).

References

[1] Jk K Aggarwal and Ms S Ryoo, *Human activity analysis*, ACM Comput. Surv. **43** (2011), no. 3, 1–43.

[2] Haibin Cai, Bangli Liu, Jianhua Zhang, Shengyong Chen, and Honghai Liu, *Visual Focus of Attention Estimation Using Eye Center Localization*, IEEE Systems Journal (2015), 1–6.

[3] Yinpeng Chen, Zicheng Liu, and Zhengyou Zhang, *Tensor-based human body modeling*, Proceedings of the ieee computer society conference on computer vision and pattern recognition, 2013, pp. 105–112.

[4] Anoop Cherian, Julien Mairal, Karteek Alahari, and Cordelia Schmid, *Mixing Body-Part Sequences for Human Pose Estimation*, 2014 IEEE Conference on Computer Vision and Pattern Recognition (2014), 2361–2368.

[5] Guilhem Cheron, Ivan Laptev, and Cordelia Schmid, *P-CNN: Pose-based CNN features for action recognition*, Proceedings of the IEEE International Conference on Computer Vision **11-18-Dece** (2016), 3218–3226, available at arXiv:1506.03607v1.

[6] Yinfeng Fang, Dalin Zhou, Kairu Li, and Honghai Liu, *Interface Prostheses with Classifier-Feedback based User Training*, IEEE Transactions on Biomedical Engineering **9294** (2016), no. c, 1–1.

[7] Juergen Gall, Angela Yao, and Luc Van Gool, *2D action recognition serves 3D human pose estimation*, Lecture notes in computer science (including subseries lecture notes in artificial intelligence and lecture notes in bioinformatics), 2010, pp. 425–438.

[8] M. Hassaballah, Aly Amin Abdelmgeid, and Hammam A. Alshazly, *Image features detection, description and matching*, Studies in Computational Intelligence **630** (2016), 11–45.

[9] Hueihan Jhuang, Juergen Gall, Silvia Zuffi, Cordelia Schmid, and Michael J. Black, *Towards understanding action recognition*, Proceedings of the ieee international conference on computer vision, 2013, pp. 3192–3199.

[10] Sam Johnson and Mark Everingham, *Learning Effective Human Pose Estimation from Inaccurate Annotation* (2011).

[11] Zhaojie Ju, Dongxu Gao, Jiangtao Cao, and Honghai Liu, *A novel approach to extract hand gesture feature in depth images*, Multimedia Tools and Applications (2015).

[12] Zhaojie Ju, Xiaofei Ji, Jing Li, Honghai Liu, and Senior Member, *An Integrative Framework of Human Hand Gesture Segmentation for Human Robot Interaction*, IEEE Systems Journal **PP** (2015), no. 99, 1–11.

[13] Preeti Kumari, Lini Mathew, and Poonal Syal, *Increasing trend of wearables and multimodal interface for human activity monitoring: A review*, Biosensors and Bioelectronics **90** (2017), no. December 2016, 298–307.

[14] Nikolaos Kyriazis, Iason Oikonomidis, Paschalis Panteleris, Damien Michel, Ammar Qammaz, Alexandros Makris, Konstantinos Tzevanidis, Petros Douvantzis, Konstantinos Roditakis, and Antonis Argyros, *A Generative Approach to Tracking Hands and Their Interaction with Objects*, Manmachine interactions 4, 2016, pp. 19–28.

[15] Christoph H. Lampert, Matthew B. Blaschko, and Thomas Hofmann, *Beyond sliding windows: Object localization by efficient subwindow search*, 26th ieee conference on computer vision and pattern recognition, cvpr, 2008.

[16] Ivan Laptev, *On space-time interest points*, International journal of computer vision, 2005, pp. 107–123.

[17] Ivan Lillo, Juan Carlos Niebles, and Alvaro Soto, *Sparse Composition of Body Poses and Atomic Actions for Human Activity Recognition in RGB-D Videos*, Image and Vision Computing **59** (2017), 63–75.

[18] An-An Liu, Yu-Ting Su, Wei-Zhi Nie, and Mohan Kankanhalli, *Hierarchical Clustering Multi-task Learning for Joint Human Action Grouping and Recognition.*, IEEE transactions on pattern analysis and machine intelligence **XX** (2016), no. X, 1–14.

[19] Damien Michel, Costas Panagiotakis, and Antonis A Argyros, *Tracking the articulated motion of the human body with two RGBD cameras*, Machine Vision and Applications (2013), 1–14.

[20] Eric Nowak, Frédéric Jurie, and Bill Triggs, *Sampling strategies for bag-of-features image classification*, Lecture notes in computer science (including subseries lecture notes in artificial intelligence and lecture notes in bioinformatics), 2006, pp. 490–503.

[21] Bryan C. Russell, Antonio Torralba, Kevin P. Murphy, and William T. Freeman, *LabelMe: A database and web-based tool for image annotation*, International Journal of Computer Vision **77** (2008), no. 1-3, 157–173.

[22] Amir Shahroudy, Jun Liu, Tian-Tsong Ng, and Gang Wang, *NTU RGB+D: A Large Scale Dataset for 3D Human Activity Analysis*, Cvpr (2016), 1010–1019, available at arXiv:1604.02808v1.

[23] Feng Shi, Emil Petriu, and Robert Laganiere, *Sampling strategies for real-time action recognition*, Proceedings of the ieee computer society conference on computer vision and pattern recognition, 2013, pp. 2595–2602.

[24] Alexander Sorokin and David Forsyth, *Utility data annotation with Amazon Mechanical Turk*, 2008 ieee computer society conference on computer vision and pattern recognition workshops, cvpr workshops, 2008.

[25] Raviteja Vemulapalli, Felipe Arrate, and Rama Chellappa, *Human action recognition by representing 3D skeletons as points in a lie group*, Proceedings of the ieee computer society conference on computer vision and pattern recognition, 2014, pp. 588–595.

[26] Michalis Vrigkas, Christophoros Nikou, and Ioannis A. Kakadiaris, *A Review of Human Activity Recognition Methods*, Frontiers in Robotics and AI **2** (2015), no. November, 28.

[27] Heng Wang, Alexander Kläser, Cordelia Schmid, and Cheng Lin CL Liu, *Dense trajectories and motion boundary descriptors for action recognition*, International Journal of Computer Vision **103** (2013), no. 1, 60–79.

[28] Yi Yang and Deva Ramanan, *Articulated Human Detection with Flexible Mixtures-of-Parts.*, IEEE transactions on pattern analysis and machine intelligence (2012), 1–15.

[29] Gang Yu, Junsong Yuan, and Zicheng Liu, *Propagative hough voting for human activity detection and recognition*, IEEE Transactions on Circuits and Systems for Video Technology **25** (2015), no. 1, 87–98.

[30] Maryam Ziaeefard and Robert Bergevin, *Semantic human activity recognition: A literature review*, Pattern Recognition **48** (2015aug), no. 8, 2329–2345.