Vision skeleton trajectory based motion assessment system for healthcare rehabilitation

Qinghongya Shi1,2, Hong-Bo Zhang1,2,3*, Li-Jia Dong1,3, Qing Lei1, Tsung-Chih Hsiao1,2

1Department of Computer Science and Technology, Huazhong University, Xiamen, Fujian, People’s Republic of China
2Fujian Key Laboratory of Big Data Intelligence and Security, Huazhong University, Xiamen, Fujian, People’s Republic of China
3Xiamen Key Laboratory of Computer Vision and Pattern Recognition, Huazhong University, Xiamen, Fujian, People’s Republic of China

Abstract: Most existing methods of motion assessment system used the contact sensor, infrared sensor, and depth sensor, and few works provided the solution of digital camera. To solve this problem, the authors propose a new motion assessment system based on light camera. In this work, the motion assessment was regarded as pattern regression problem of skeleton joint trajectory. Firstly, the system uses the camera to capture the image sequences. The pose estimation method is used to obtain the human skeleton from image. Secondly, due to the difference of motion frequency of each person, the length of the image sequences is different and the length of each joint trajectory also will be different. Fourier transform is applied to normalise the joint trajectory and use the coefficients of Fourier transform as the joint trajectory feature. Finally, the regression model is built to assess the motion quality. Some experimental results and discussion on action video data are used to verify the effectiveness of the system.

1 Introduction

With the development of social economy, people’s attention has shifted from eating and wearing to healthy. Exercise rehabilitation and healthcare system is not only needed by professional athletes, but also by the general public after injuries. The emergence of motion rehabilitation system provides a systematic and standardised performance index for exercise rehabilitation and healthcare. It can monitor and manage the patients in time and effectively, and reduce the risk of patients in the process of motion rehabilitation. Motion rehabilitation assessment is very important for exercise rehabilitation and healthcare system. The motion rehabilitation assessment can be divided into two types: contact sensor-based system and non-contact sensor-based system. It is more popular to use a contact sensor to estimate human motion state [1], which extracted human biological data such as arm movement speed, frequency, and strength. The motion rehabilitation system can use these data to assess exercise status and evaluate motion rehabilitation status to provide users more accurate and convenient service.

However, using the contact sensor will affect the movement process of user, which contains slow down the speed of the movement, lower the amplitude of the movement, and increase the resistance of the movement, especially in motion evaluation system. With the development of computer vision technology, more and more approaches which have good performance in computer vision has been extended to various fields, including human pose estimation, action recognition, object tracking, and so on. Especially, deep learning-based methods perform very good results in all the above fields. How to build a non-invasive motion evaluation system based on vision camera has attracted more attention in recent years, the key is to use visual methods to acquire the biological information of human body instead of contact sensor.

To address this problem, some techniques have been proposed to analyse the motion status from videos [2–4], including spatial temporal interesting point (STIP) based methods and skeleton based methods. STIP is the representation of shape and motion, mainly including Hu invariant moments based on motion history and energy, iterative filtering, frame grouping, and Poisson significance. STIP based methods regarded the human motion as a set of key points and encoded the feature of STIP to model human motion.

Other methods to represent human action are human skeleton-based methods which use human pose estimation algorithm to obtain skeleton [5–7]. The action analysis method based on skeleton features mainly uses the changes of human body’s joints to describe the action, which focus on the variation of joints’ position and appearance. However, skeleton features relatively depend on pose estimation and are greatly affected by background occlusion. In addition, skeleton data always provide depth information, which is judged by visual influence. Skeleton based methods represent human action by the change of skeleton and depth information. However, only some public data can provide depth information and the depth sensor is influenced by environmental and high cost.

Building the action analysis system based on vision camera is very important for the practical application of these methods. In this paper, we design a motion assessment system for healthcare rehabilitation based on digital camera. To evaluate the state of the healthcare rehabilitation, the tester is required to complete the specified action. A novel motion assessment method by using visual camera is proposed in this system. The assessment score is corresponded to the status of healthcare rehabilitation. The framework of the proposed method is shown in Fig. 1.

Different from other methods which used depth camera to capture human skeleton, we just use the vision camera and the human skeleton can be extracted by deep learning methods. Firstly, human pose estimation algorithm was used to detect the skeleton joint in real time. Although the assessment system which only uses the camera sensor is more convenient to apply, the problem that its algorithm needs to solve is more complicated. This scene is similar to an open world, and it is difficult to control the posture and angle of the measured people. Due to the difference of the height and posture of the measured people, the difference of the viewpoint and distance of the camera, the skeleton size of the measured people will perform different. It is defined as the skeleton scale change in this work. Some examples are shown in Fig. 2.

On the other hand, different people also have different movements when performing the same action, the joint trajectory of different people will have different length, range, and frequency. In order to overcome the skeleton scale change, in the proposed method, joint trajectory feature was normalised by relative
The field of computer vision has advanced rapidly over the past few years, especially the human pose estimation, while the problem of estimating body skeleton is regarded as human pose estimation. The contributions of our work are twofold:

(i) We propose a systemic framework for motion assessment based on vision camera.
(ii) A pose estimation-based method has been proposed to evaluate motion quality for healthcare rehabilitation and the experimental results have been discussed in this paper.

The remainder of this paper is organised as follows. Section 2 reviews some related works. Section 3 describes the proposed method. Section 4 presents and discusses about our experimental results. Finally, Section 5 concludes the paper.

2 Related work

Excellent performance motion assessment system can accurately judge the recovery of the patient’s body; it will save a lot of time for the doctor and the patient. Recently, the appearance of plentiful approaches, that are capture human motion information from different sensors, is to assess the motion quality. All the above methods promote the progress of motion assessment system. We mainly introduce relative works for motion analysis by vision camera which are summarised in this section. First, the vision based human pose estimation methods are introduced. Then, motion recognition assessment methods are provided.

2.1 Vision based human pose estimation methods

The field of computer vision has advanced rapidly over the past few years, especially the human pose estimation, while the problem of estimating body skeleton is regarded as human pose estimation. A common method is to use depth sensor [8]. Marin-Jimenez et al. [9] proposed 3D human pose estimation from depth data which obtained a depth map containing a person and a set of predefined 3D prototype poses and returns the 3D position of the body joints of the person. Gilbert et al. [10] proposed an approach to fuse multi-viewpoint video with inertial measurement unit sensor data (without optical markers, a complex hardware setup or a full body model) to accurately estimate 3D human pose.

Depth sensor is high cost and low accuracy in outdoor environment, many approaches provided the solution for estimating human skeleton from still image [11]. Felzenszwalb et al. [12] proposed a mixture of multiscale deformable part model to detect and locate object from image. These objects are variable, including not only brightness and perspective, but also the deformation of the object. Uktka and Uematsu [13] used part-segment features for estimating an articulated pose in still images. With advance of deep learning methods in computer vision, many methods have proposed effectively pose and skeleton estimation algorithm based on deep learning theory. Cao et al. [7] proposed OpenPose algorithm to estimate multi-person 2D pose in real time, and is compared with the existing methods; the biggest advantage is that the detection speed is not sensitive to the number of characters, which greatly improves the speed while maintaining the detection accuracy. In [14], the body part proposal is first detected with Resnet, and then classified as different people based on association and spatial information. Body part detector is applied and an incremental optimisation strategy is used to estimate human pose.

2.2 Vision based motion recognition and assessment

Human motion recognition is a hot research topic in computer vision and has been widely used in many application [2, 15]. Especially with the advancement of vision-based pose estimation [16], skeleton sequences have become a key feature of motion recognition. The skeleton based motion recognition methods can be divided into two categories: hand-crafted feature-based methods and deep learning-based methods.

Methods of hand-crafted features: Keceli and Can [17] extracted human action features based on the angle and displacement between each joint and reference joint, instead of calculating absolute skeletons. The relative joint trajectory was modelled by Fourier transform to eliminate the influence of motion difference. Finally, the regression model was built to assess the motion status based on the feature distance between the input video and standard motion video.

The contributions of our work are twofold:

(i) We propose a systemic framework for motion assessment based on vision camera.
(ii) A pose estimation-based method has been proposed to evaluate motion quality for healthcare rehabilitation and the experimental results have been discussed in this paper.

The remainder of this paper is organised as follows. Section 2 reviews some related works. Section 3 describes the proposed method. Section 4 presents and discusses about our experimental results. Finally, Section 5 concludes the paper.
3 Proposed method

3.1 Human skeleton estimation and representation

Compared with RGB information, skeleton information has clear and simple features and is not easily affected by appearance factors. So, we decide to locate the body by using skeletal joints. There are two ways to get human skeleton, one is obtained by joint estimation from RGB image, the other is straight obtained by depth information from RGB-D systems. Skeleton information has clear factors. So, we decide to locate the body by using skeletal joints. Based on OpenPose, the input video can be regarded as the skeleton sequence \( V = \{ S_i, S_2, \ldots, S_n \} \), \( n \) is the number of video frames. \( S_i \) is the human skeleton in each frame. \( S_i = \{ S_{i1}, S_{i2}, \ldots, S_{im} \} \) where \( S_{ij} \) means the \( j \)th joint points in skeleton and \( m \) is the number of joints in skeleton. In OpenPose, \( m \) is set to 18. For each skeleton joint, \( S_{ij} \) can be represented by two-dimensional coordinates in the image space \( S_{ij} = (P_{ij}^x, P_{ij}^y) \).

3.2 Skeleton feature

Due to the difference of each video, such as the number of frames, motion scale, and the skeleton size of different person, the joint feature has been calculated by relative distance with the neck joint (1th joint) as defined in Fig. 3. So, for the skeleton in each frame, we can extract the relative distance \( d_{ij} \), which means the Euclidean distance between the \( j \)th joint and 1th joint in \( i \)th frame. Then, for each joint, the distance feature can be regarded as \( d_{ij} = \{ d_{ij}^1, d_{ij}^2, \ldots, d_{ij}^n \} \), \( i = 1, 2, \ldots, n \), \( i \) is the index of video frame. The Fourier transform have been used to compute the video feature to normalise the different length of video. It also can extract the frequency information of the change of the relative distance. The first ten coefficients of the Fourier transform have been regarded as the joint feature \( t_{d_{ij}} \) in the video. Finally, for each video, the feature contained \( L \) Fourier coefficients vectors \( D = \{ t_{d_{11}}, t_{d_{12}}, \ldots, t_{d_{1n}} \} \), the dimension is \((m-1) \times L\).

3.3 Motion assessment by feature distance regression

We extracted the skeleton information and solved the processing of different videos. Next, we need to put the processed information into the model and let the model evaluate the action. To evaluate the motion status, the support vector regression (SVR) is used to compute the score of the input video. In order to make the regression model convincing and standardised, two main works have been done. One is built a standard dataset for action assessment. The golf video is collected from web and sport dataset. The other is the motion score which has been labelled by professional person.

Supposing that the training set contains a set of labelled training video \( D = \{(D_1, y_1), (D_2, y_2), \ldots, (D_k, y_k)\} \), for which the score of motion status in each image is given. \( y_i \) is equal to the score. \( D_i \in \mathbb{R}^k \), \( k = m - 1 \times L \) is the feature vector of the video. \( l \) is the size of the training set. The regression model \( G(y) \) is trained to find a real-valued function \( f(D) \) of the feature vector. Let \( \phi(D) \) be a feature transformation of \( D, \phi: \mathbb{R}^K \rightarrow \mathbb{R}^L \). Consider the case in which \( f(D) \) is linear in the transformation space, and the target score \( y \) can be modelled by the following equation:

\[
y = G(D) = f(D) + \varepsilon = \phi(D)^T w + \varepsilon \tag{1}
\]

where \( w \in \mathbb{R}^K \) and \( \varepsilon \) is the observation noise. In the proposed method, the SVR is applied to train the model and predict the score for testing video.

To train the model, the loss function is the error between the regression score \( y_p \) and ground true \( y_g \). In SVR, the loss function is defined as follows:

\[
\text{err}(y_p, y_g) = \begin{cases} 
0, & |y_g - \phi(D)^T w - \varepsilon| \leq b \\
|y_g - \phi(D)^T w - \varepsilon| - b, & |y_g - \phi(D)^T w - \varepsilon| > b
\end{cases} \tag{2}
\]

where \( b \) is relaxation variable. The goal of the training process is to compute the parameters which can minimise the loss function. The optimisation of this loss function can be solved based on Lagrangian multiplier.

4 Experimental results and discussion

In this section, we evaluate the effectiveness of our proposed approach. We first introduce the criteria of evaluation that other regression models always use. Then we show the results of our experiments.

4.1 Evaluation criteria

To evaluate the accuracy of motion assessment, the mean squared error (MSE) and mean absolute error (MAE) are used as the evaluation criteria during all experiments:

\[
\text{MSE} = \frac{1}{T} \sum_{i=1}^{T} (\hat{y}_i - y_i)^2 \quad \text{MAE} = \frac{1}{T} \sum_{i=1}^{T} |\hat{y}_i - y_i| \tag{3}
\]

where \( \hat{y}_i \) is the estimation result of the \( i \)th video from the regression model and \( y_i \) is the ground truth of the \( i \)th videos. \( T \) is the number of test dataset. The reason of using MSE and MAE is that single criterion cannot be comparable and MSE is more convenient while MAE is more robust for abnormal point.

In this work, an action assessment dataset is built. There are total 320 videos in the dataset. Each video performs one action by one person. Each video has the assessment score which is labelled from 0 to 1 by the professionals. The higher score means the higher the completion of the action, which means a better rehabilitation. In the experiment, we use \( K \)-fold cross-validation strategy, and \( K \) is set to 10. The dataset was divided into ten portions, each one...
In this paper, we proposed a new method to evaluate human motion for healthcare rehabilitation. The motion assessment problem has been regarded as the motion status regression problem. It can be addressed by skeleton based motion feature and SVR. To verify the performance of the proposed method, an action assessment dataset is built for the assessment of healthcare rehabilitation. The experimental result showed that the proposed method is effective.

6 Acknowledgments

The authors thank the anonymous reviewers for their valuable and insightful comments on an earlier version of this manuscript. This work is supported by the National Natural Science Foundation of China (no. 61871196), the National Natural Science Foundation of Fujian Province of China (no. 2019J01082), the Promotion Program for Young and Middle-aged Teacher in Science and Technology Research of Huaqiao University (ZQN-YX601), and Subsidized Project for Postgraduates’ Innovative Fund in Scientific Research of Huaqiao University (no. 18014083014).

7 References

[1] Huang, L.S., Chu, Y.Z., Zhang, X.F.: ‘Recognition of table tennis strokes based on acceleration sensor’. Transducer Microsyst. Technol., 2016, 35, (1), pp. 56–61
[2] Zhang, H.B., Zhang, Y.X., Zhong, B.N., et al.: ‘A comprehensive survey of vision-based human action recognition methods’. Sensors, 2019, 19, (5), p. 1005
[3] Müller, B., Jig, W., Giese, M.A., et al.: ‘Validation of enhanced kinect sensor based motion capturing for gait assessment’. Plas ONE, 2017, 12, (4), p. e0175813
[4] Zia, A., Sharma, Y., Bhatadupa, V., et al.: ‘Video and accelerometer-based motion analysis for automated surgical skills assessment’. Int. J. Comput. Assist. Radiol. Surg., 2018, 13, (3), pp. 443–455
[5] Güler, R.A., Neverova, N., Kokkinos, I.: ‘Densepose: dense human pose estimation in the wild’. Proc. IEEE Int. Conf. on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 2018, pp. 7297–7306
[6] Fang, H.S., Xie, S., Tai, Y.W.: ‘RMPE: regional multi-person pose estimation’. Proc. IEEE Int. Conf. on Computer Vision, Venice, Italy, 2017, pp. 2334–2343
[7] Cao, Z., Simon, T., Wei, S.E., et al.: ‘Realtime multi-person 2d pose estimation using part affinity fields’. Proc. IEEE Int. Conf. on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 2017, pp. 7291–7299
[8] Shotton, J., Girshick, R., Fitzgibbon, A., et al.: ‘Efficient human pose estimation from single depth images’. IEEE Trans. Pattern Anal., 2013, 35, (12), pp. 2821–2840
[9] Marín-Jiménez, M.J., Romero-Ramírez, F.J., Muñoz-Salinas, R., et al.: ‘3d human pose estimation from depth maps using a deep combination of poses’. J. Vis. Commun. Image Represent., 2018, 55, pp. 627–639
[10] Gilbert, A., Trumble, M., Malleson, C., et al.: ‘Fusing visual and inertial sensors with semantics for 3D human pose estimation’. Int. J. Comput. Vis., 2019, 127, (4), pp. 381–397
[11] Zhang, H.B., Lei, Q., Zhong, B.N., et al.: ‘A survey on human pose estimation’. Intell. Autom. Soft Comput., 2015, 22, (3), pp. 483–489
[12] Felzenszwalb, P.F., Girshick, R.B., McAllester, D., et al.: ‘Object detection with discriminatively trained part-based models’. IEEE Trans. Pattern Anal., 2010, 32, (9), pp. 1227–1245
[13] Ukti, N., Uematsu, Y.: ‘Semi- and weakly-supervised human pose estimation’. Comput. Vis. Image Underst., 2018, 170, pp. 67–78
[14] Insafutdinov, E., Pishchulin, L., Andres, B., et al.: ‘Deepercut: A deeper, stronger, and faster multi-person pose estimation model’. European Conf. on Computer Vision, Cham, 2016, pp. 34–50
[15] Wang, J., Chen, Y., Hao, S., et al.: ‘Deep learning for sensor-based activity recognition: a survey’. Pattern Recognit. Lett., 2019, 119, pp. 3–11
[16] Presti, L.L., Cascia, M.L.: ‘3D Skeleton-based human action classification: a survey’. Pattern Recognit., 2016, 53, (C), pp. 130–147
[17] Keceli, A., Can, A.: ‘Recognition of basic human actions using depth information’. Int. J. Pattern Recognit. Artif. Intell., 2014, 28, (2), p. 1450004
[18] Yang, X., Tian, Y.L.: ‘Effective 3D action recognition using EigenJoints’. J. Vis. Commun. Image Represent., 2014, 25, (4), pp. 2–11
[19] Yan, S., Xiong, Y., Lin, D.: ‘Spatial temporal graph convolutional networks for Skeleton-based action recognition’. Thirty-Second AAAI Conf. on Artificial Intelligence, New Orleans, LA, USA, 2018
[20] Zhu, Y., Lin, Z., Newsam, S., et al.: ‘Hidden two-stream convolutional networks for action recognition’. Asian Conf. on Computer Vision, Cham, 2018, pp. 363–378
[21] Springer, S., Yogev, S.: ‘Validity of the kinect for gait assessment: a focused review’. Sensors, 2016, 16, (2), p. 194