Applied Mathematics, Modeling and Computer Simulation
C. Chen (Ed.)

doi:10.3233/ATDE220094

Three-Dimensional Human Posture Rehabilitation Detection Based on Vibe

Nianfeng LI[a], Yupeng LI[b], Lina LI[a,1], Yan LI[b] and Zhiguo XIAO[a]
[a] *College of Computer Science and Technology, Changchun University, China*
[b] *Graduate School of Changchun University, Changchun University, China*

Abstract. The main manifestation of stroke patients is physical disorder, so physical rehabilitation treatment will also become a very important link. In order to solve this problem, we have researched and developed a physical rehabilitation detection system based on VIBE. System using existing large-scale motion capture data set (AMASS) and unpaired, the laboratory environment points 2D note, by convolution neural network training out a preliminary training model, and then will recover limb disorders of the video input into a network testing, according to the result of parameter judge the body recovery degree of the patients. In addition, the attention mechanism is added in the network to better fit the movements of the body to achieve the real effect. Through the analysis of the test results of rehabilitation video, it is proved that this method has a good effect on the detection of limb rehabilitation.

Keywords. VIBE, Convolutional Neural Network, Motion capture dataset, AMASS, Attention mechanism.

1. Introduction

According to 《The Report on Nutrition and Chronic Diseases of Chinese Residents (2015)》, the number of stroke patients in China is increasing by 9% every year. According to the world Bank, there will be 31.77 million stroke patients in China by 2030. The main manifestation of stroke patients is physical disorders, so physical rehabilitation treatment will also become a very important link. Human THREE-DIMENSIONAL pose estimation refers to the process of restoring key positions of human body in a given picture or video, which is of great significance in describing human body posture and predicting human behavior. However, 3d pose estimation is an ill-posed problem because of the inherent fuzziness in back-projecting the 2d view of an object into 3D space to preserve its structure. As 3D pose people can be in 2 d plane projection in myriad ways, so the mapping from 2D posture to the 3 d pose is not the only based on the key points of human body in the 3 d pose estimation[1] [2], when overlapping parts of the body's key point, it is difficult to accurately distinguish behavior characteristics, and the human body geometry model based on parameterized posture estimation, GAN network [2] was used to carry out confrontational training, which can

[1] Corresponding Author, Lina LI, College of Computer Science and Technology, Changchun University, China; E-mail: liln@ccu.edu.cn.

solve this problem well. VIBE used in this paper is implemented based on the SMPL model.

2. Related Work

A method for 3d posture estimation. At present, there are two main methods for 3d human pose estimation: one is 3d human skeleton key points connected by 3D key points developed in the classic ASM face key point detection algorithm proposed by Cootes in 1995, so as to perform visualization [3]. The other is the parameterized geometric model of human body, commonly used as SMPL model. The 3D generation model of SMPL is extracted from 2D joints [4], and its deformation is usually controlled by a set of poses, and the parameters of poses and shapes need to be estimated.

The method of using parametric model has also experienced a long time of exploration. First of all, although great progress has been made in estimating three-dimensional human posture and shape from a single picture [1][5], it is difficult to obtain the correct motion state of limbs when estimating posture from motion videos, mainly because the time information of motion cannot be obtained from a single picture [1]. Away from a single video in the 3D human body posture to estimate method, the result has never been predicted accurately, the main reason is the lack of true 3D annotation [1], usually methods [6][7] is will indoor 3D data sets with 2D ground real values of ground truth values or pseudo key annotation method combining the video, but this way is very limited, For example: ① the indoor environment is relatively single, and the real outdoor environment has a big gap; ② The number of videos annotated by key points of 2D ground real value or pseudo ground real value cannot reach the number required by deep learning. Therefore, none of the previous methods can accurately fit the posture of the human body [1]. In VIBE, two unpaired information sources are used by training sequence-based generative antagonistic network (GAN) [2]. Here, given a video of a person, we train a time model to predict the parameters of the SMPL body model for each frame, while the motion discriminator attempts to distinguish between the real sequence and the regression sequence. By doing so, the regressors were encouraged to output postures that represented reasonable movement by minimizing the loss of confrontational training, with the discriminator acting as a weak monitor. Motion discriminator uses ground real motion capture (mocap) data to implicitly learn and interpret statics, physics and kinematics of human body in motion [1].

3. Methods

3.1. The Overall Framework

The overall framework for VIBE is shown in figure 1. Given a video of length T, pre-trained CNN is used to extract the features of each frame, and then a time encoder composed of bidirectional gated recursive unit (GRU) is trained, which outputs variables containing information of past and future frames. These features were used to perform regression on the parameters of SMPL limb model at each point [1].

SMPL uses θ to indicate body posture and shape. θ consists of the relative rotations of 23 joints in the global body rotation and axial Angle scheme based on posture and shape. Shape parameters are the first 10 coefficients of PCA shape space. According to the calculation method in previous paper [8][9], gender-neutral shape model is used. Given these parameters, the SMPL model is a differentiable function that outputs a postulated three-dimensional grid.

$$M(\alpha, \beta) \in R^{6890*3}$$

Given a video sequence, VIBE calculates:

$$\widehat{\Theta} = [(\widehat{\theta_1}, \ldots., \widehat{\theta_T}), \hat{\beta}]$$

Assume is the posture parameter of $\widehat{\Theta}$ the time step T, and assume $\hat{\beta}$ is the joint shape of this sequence. Specifically, for each frame, we predict body shape parameters. Then, we apply the average pool to obtain a single shape $(\hat{\beta})$ in the entire input sequence. We refer to the model described so far as time generator G. Assume that the output from G and the sample Θ_{real} from AMASS are provided to the motion discriminator D_M to distinguish the fake example from the real example[1].

3.2. Motion Discriminator

The constraint of a single image is not enough to explain the posture sequence. When the time continuity of motion is ignored, multiple inaccurate postures may be considered effective. Therefore, a motion discriminator D_M is used to judge whether the generated posture sequence corresponds to the real one [1]. Figure 2 shows the structural network of D_M. With this D_M structure, the final output is the true/false probability of each sequence of inputs.

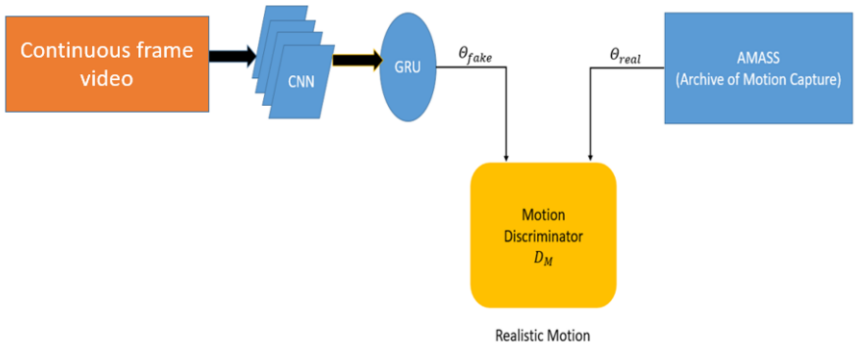

Figure 1. VIBE architecture.

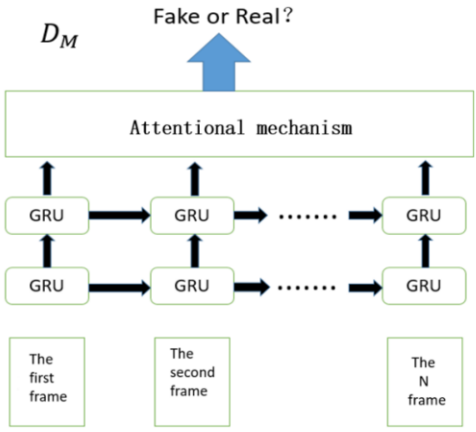

Figure 2. Motion discriminator architectureD_M

3.3. Results Analysis

Resnet-50 network [10] was used as the image encoder to conduct pre-training on the single frame pose and shape estimation task [8][11]. T=16 was used as the sequence length, and the minimum batch size was 32. For the time encoder, a 2-layer GRU with hidden size of 1024 was used. The SMPL regressor has two fully connected layers, each with 1024 neurons. For the attentional mechanism, two MLP layers (each with 1024 neurons) and TANH activation were used to learn attentional weights.

4. Experiment

The overall experimental environment is as follows: The server CPU is Corei9, equipped with two Nvidia GeForce RTX 3090 graphics cards, operating system is CentOS(20.04) system, and python (3.6) is used for development. Data set The existing AMASS data set and our data were trained through GAN network, so as to achieve relatively accurate fitting results.

During the experiment, we will be upper limb disorders of elbow flexion arm action video and kick stand rehabilitation were tested in the video input to the network, as shown in figure 3(a), figure 3(b) and figure 4(a), as shown in figure 4(b), the analysis of test results, in the key position, as well as the edges, through the network, the results from these results better than other networks.

Figure 3(a). Flexural arm rehabilitation test, before test

Figure 3(b). Flexural arm rehabilitation test, after test

Figure 4(a). Standing kick recovery test, before test

Figure 4(b). Standing kick recovery test, after test

In the experiment, we also compared it with other methods [6] As shown in figure 5(a) and figure 5(b), there was deviation when fitting with real body activities, mainly because the former work lacked real 3D annotation for the image, so the final result was not as good as VIBE.

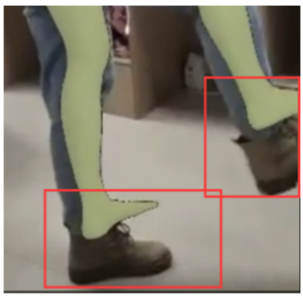

Figure 5(a). Other experimental results

Figure 5(b). Our graph of the result after adding the attentional mechanism

5. Conclusion

For the first time, we applied three-dimensional human posture recognition based on VIBE in limb rehabilitation. On the basis of VIBE, we added attention mechanism to the

physical rehabilitation movement, so that the result could more accurately fit the real human posture. In addition, we conducted comparative tests with other similar 3d human pose recognition methods [12][13], and the results showed that our method performed better.

However, there are still some shortcomings in this method. In accordance with the method in paper [8][11], the body shape model with gender neutral was used in the experiment, so the gender of the patient could not be identified. Secondly, although the body movements can be more accurately fitted after the attention mechanism is added, the human body cannot be recognized when some feature points are blocked. The next step is to prepare these questions so that they can be better applied to real life.

Acknowledgments

The work was supported by the project plan of science and technology development center of the Ministry of Education (No. 2020hyb03002) and the Jilin Science and Technology Development Plan Project (No. 20210201083GX and No. 20200404221YY) and the Natural Science Foundation of Jilin Province (No. YDZJ202101ZYTS191) and the Jilin Provincial Department of Education Plan Project (No. JJKH20210632KJ).

References

[1] Kocabas M, Athanasiou N, Black M J. Vibe: Video inference for human body pose and shape estimation[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020: 5253-5263.
[2] Hossain M R I, Little J J. Exploiting temporal information for 3d human pose estimation[C]//Proceedings of the European Conference on Computer Vision (ECCV). 2018: 68-84.
[3] Bogo F, Kanazawa A, Lassner C, et al. Keep it SMPL: Automatic estimation of 3D human pose and shape from a single image[C]//European conference on computer vision. Springer, Cham, 2016: 561-578.
[4] Sminchisescu C, Triggs B. Estimating articulated human motion with covariance scaled sampling[J]. The International Journal of Robotics Research, 2003, 22(6): 371-391.
[5] Goodfellow I, Pouget-Abadie J, Mirza M, et al. Generative adversarial nets[J]. Advances in neural information processing systems, 2014, 27.
[6] Sun Y, Ye Y, Liu W, et al. Human mesh recovery from monocular images via a skeleton-disentangled representation[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019: 5349-5358.
[7] Kocabas M, Karagoz S, Akbas E. Self-supervised learning of 3d human pose using multi-view geometry[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019: 1077-1086.
[8] Kolotouros N, Pavlakos G, Black M J, et al. Learning to reconstruct 3D human pose and shape via model-fitting in the loop[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019: 2252-2261.
[9] Kanazawa A, Black M J, Jacobs D W, et al. End-to-end recovery of human shape and pose[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 7122-7131.
[10] Arjovsky M , Chintala S , L, et al. Wasserstein generative adversarial networks. 2017.
[11] Kanazawa A, Zhang J Y, Felsen P, et al. Learning 3d human dynamics from video[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019: 5614-5623.
[12] Mehta D, Rhodin H, Casas D, et al. Monocular 3d human pose estimation in the wild using improved cnn supervision[C]//2017 international conference on 3D vision (3DV). IEEE, 2017: 506-516.
[13] Loper M, Mahmood N, Romero J, et al. SMPL: A skinned multi-person linear model[J]. ACM transactions on graphics (TOG), 2015, 34(6): 1-16.