

View-invariant Pose Analysis for Human Movement Assessment from RGB Data

Faegheh Sardari, Adeline Paiement, Majid Mirmehdi

▶ **To cite this version:**

Faegheh Sardari, Adeline Paiement, Majid Mirmehdi. View-invariant Pose Analysis for Human Movement Assessment from RGB Data. 20th International Conference on Image Analysis and Processing (ICIAP), Sep 2019, Trento, Italy. 10.1007/978-3-030-30645-8_22 . hal-02171028

HAL Id: hal-02171028
https://hal.science/hal-02171028

Submitted on 15 May 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

View-invariant Pose Analysis for Human Movement Assessment from RGB Data

Faegheh Sardari[1][0000−0002−9134−0427], Adeline Paiement[2][0000−0001−5114−1514], and Majid Mirmehdi[1][0000−0002−6478−1403]

[1] Department of Computer Science, University of Bristol, Bristol, UK
{faegheh.sardari, m.mirmehdi}@bristol.ac.uk
[2] Laboratoire d'Informatique et Systèmes, University of Toulon, Toulon, France
adeline.paiement@univ-tln.fr

Abstract. We propose a CNN regression method to generate high-level, view-invariant features from RGB images which are suitable for human pose estimation and movement quality analysis. The inputs to our network are body joint heatmaps and limb-maps to help our network exploit geometric relationships between different body parts to estimate the features more accurately. A new multiview and multimodal human movement dataset is also introduced part of which is used to evaluate the results of the proposed method. We present comparative experimental results on pose estimation using a manifold-based pose representation built from motion-captured data. We show that the new RGB derived features provide pose estimates of similar or better accuracy than those produced from depth data, even from single views only.

Keywords: Pose Analysis · View-invariant CNN · Health Monitoring ·

1 Introduction

Assessing the quality of human movement is of paramount importance in many areas of human activity, such as sports, health, and surveillance, exemplified by recent works such as [15, 12, 7, 14, 13, 11]. For example, amongst many clinic and home-based tests for patient monitoring in Parkinsons disease [19], a patient's quality of walking or steadiness while standing must be observed, e.g. both soon after prescribing medication and longitudinally across weeks and months as the progression of the disease is assessed. Using computer vision to automate such rehabilitation assessments would eliminate the costs and subjective variability associated with clinicians, and allow the generation of clinical scores that are more consistently and autonomously applied, e.g. [9].

Our motivation is therefore to design a system that allows us to measure both *frame-by-frame and the overall abnormality* in human movement when performing certain actions – with the aim of eventual development of corresponding scores to reflect a measure of (ab)normality. These requirements call for the design of a robust pose estimation method that provides accurate *frame-by-frame*

estimates. Our specific application area is for patient rehabilitation actions, such as walking, sitting-to-standing, and so on [6, 8, 2]. Thus, the pose estimation must be based on robust features obtained from realistic sensor settings for home environments, such as affordable single, or just a small few, RGB cameras.

There is a significant body of work on vision-based human body motion analysis, which for our purposes may be categorised into: (i) traditional methods using high level human pose features [12, 18, 15, 1], and (ii) deep learning approaches [14, 10, 13, 20] that extract features directly from images using CNN networks. The latter may then score a movement's quality directly from such features, or they may use them to provide a body pose estimate, to be used for movement analysis in a later stage. We consider works that rely on wearable technology, such as [8, 16], as out of scope, since we wish to focus on both remote sensing for patient comfort and design methods that may have potential use in other applications, such as sports and surveillance.

Pirsiavash et al. [15] proposed a regression-based method to score sport actions in an Olympic sports dataset, that they also released. They trained an SVM classifier on both low-level edge and velocity features and high-level pose features represented in the frequency domain by the discrete cosine transform. While their method was able to narrow down which segments included higher scoring movements, the performance of their features dropped particularly when encountering self-occlusions. Their method predicts action scores better than human non-experts, but it is far from human expert judgment.

Using 3D joints data to analyse human movements, often generated by RGBD cameras and VICON systems, has picked up pace in recent years, for example in [12, 18, 4, 17]. Not surprisingly, the pose features derived from 3D data are richer and can be leveraged to assess a wider range of movements. However, then the curse of dimensionailty can strike and the application of dimensionality reduction methods, such as PCA or manifold learning, becomes necessary to reduce the redundancy presented in the 3D joints space. In [12], Paiement et al. used skeleton data to model pose information in a reduced dimension manifold for a stairs-climbing rehabilitation analysis application. They then trained a custom-designed statistical model on the pose information gathered from the action video to score the movement's quality on a frame-by-frame basis. Chaaraoui et al. [4] generated a body-joints motion history volume from 3D spatio-temporal skeleton joint features, and reduced the dimension of their volume based on axis projections. They then classified abnormal gait in their own frontal-view dataset using BagOfKeyPoses on their skeletal joints volume.

Deep learning based methods have also been increasingly applied to assess the quality of movement, for example [7, 14, 10, 13]. Crabbe et al. [7] modified the work by Paiement et al. [12] by proposing a CNN regression approach to estimate the high dimensional body pose from depth silhouettes in the same low-dimensional manifold space that was developed for their SPHERE Stairs dataset [12]. AlexNet was applied to perform their pose estimation by mapping depth silhouettes onto the manifold space. The authors discussed that the use of depth silhouettes allowed simplifying the learning task for their deep CNN in the

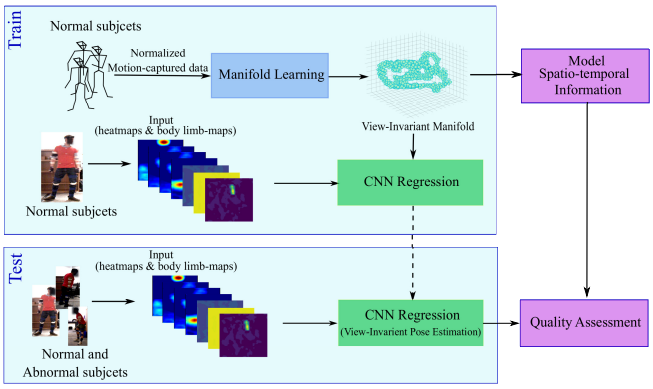


Fig. 1. The overall schema of the proposed approach (including training and testing phases) for normal/abnormal pose estimation.

absence of a large training dataset. However, the extraction of good, accurate-enough silhouettes for movement quality assessment can be a difficult process. Parmar et al. [13] divided a video into 16-frame video clips and averaged the spatiotemporal features from all clips, obtained by applying 3D CNNs [20], to classify sports actions and estimate their score. Li et al. [10] also divided each video into several parts, and extracted their features using 3D CNNs [20]. Then, all features were concatenated and fed into a two-layer convolutional network to predict the action scores. Since such methods extract spatiotemporal features for a whole video, they are better suited to providing a global score rather than analysing human movement in each frame.

In a similar fashion to Paiement et al. [12], Liao et al. [11] verified that dimensionality reduction, implemented through an autoencoder in their case, combined with statistical modelling of a movement's kinematics may provide discriminating pose estimates and movement quality scores for an instance of a movement. They then trained three different types of NNs (CNN, RNN, and HNN) to perform a (whole) movement quality score prediction from sequences of raw VICON skeleton data. Although they effectively used multiview data to generate their model, the data for their testing must also be obtained by their mocap system which makes their method somewhat impractical for participants, especially in more everyday applications, and requires the presence of experts to set up the system. In addition, their fully integrated NN approach extracts spatiotemporal features that do not allow disentangling the pose from the kinematic problems and cannot finely analyze movement on a frame-by-frame basis.

In this paper, we propose a ResNet-based regression method that extracts high-level pose features from body joint heatmaps and body limb-maps from single RGB images of arbitrary viewpoint. A view-invariant manifold obtained from motion-captured 3D joint positions serves as the target pose estimate space for our CNN. A customised statistical model, from [12, 18] is then used to detect and score movement abnormalities on a frame-by-frame basis using our pose es-



Fig. 2. Sample frames from the proposed dataset from four different viewing directions for turn-walk (top row), and limp action (bottom row).

timate. The overall approach is illustrated in Figure 1. The major contributions of our method are its ability to determine and score movement abnormality (in a healthcare setting) from any reasonable viewpoint given an RGB video and without relying on explicit 3D (skeleton or depth) information. Further, we introduce a new fully annotated multiview and multimodal dataset that will be available to the community for the development of health-related or rehabilitation methods. To the best of our knowledge, it is the first time that features extracted from single RGB images are demonstrated to be suitable for movement quality analysis in a healthcare application.

Next, in Section 2, a new multiview and multimodal dataset is introduced, followed by our proposed method in Section 3. Experiments and comparative results are presented in Section 4. Conclusions are in Section 5.

2 SMAD: Sphere Multiview and Multimodal Movement Assessment Dataset

There is increasingly more datasets becoming available for human movement analysis, but there is simply no 'one size fits all' that would be of use across different applications and outcomes. For example, the Olympics sports dataset introduced by Pirsiavash et al. [15], which includes diving and skating actions extracted from Youtube videos, is useful for assessment of overall human movement performance, but would not be of use in rehabilitation movement analysis. Parmar et. al [13] also collected a multiview dataset for the diving action. Paiement et al. [12, 18] captured three single (frontal) view datasets of walking, walking up stairs, and sitting to standing movements, to evaluate their movement quality assessment method for health-related applications. The skeleton, depth, and colour data was captured by a Primesense camera [18], and a physiotherapist manually annotated all frames into normal and abnormal. Vakanski et al. [21] developed a skeletal movement dataset using a VICON system and a Kinect camera for physical rehabilitation exercises involving 10 healthy subjects who performed their exercises in both correct and incorrect fashion. This dataset was then used in [11] as described briefly in the previous section.

We have captured a new multiview human movement dataset that combines, for the first time, motion capture [1], and skeletons, depth and RGB images from one Microsoft Kinect and three Primesense cameras. While the dataset includes different types of actions, here we focus on only the 'turn and walk' action, performed both normally and with 3 types of abnormalities by 19 healthy subjects: turn and walk action (turn-walk), turn and walk with stroke (stroke), turn and walk with short limp (short limp), turn and walk with Parkinson (Parkinson). For the last three actions, the participants were trained by a specialist Physiotherapist. The turn-walk action was repeated five times, while the other actions were performed only once. The actions were videoed from four camera viewing directions for the entirety of each walk – towards one camera and back to the opposite camera, one side view, and one downward view of the scene. Two samples from the dataset are displayed in Figure 2.

3 Proposed Method

To use 3D human body joints to generate a pose space and assess the quality of human movement, dimensionality reduction becomes inevitable to discard the redundant or correlated dimensions. Here, we follow the approach adopted by other works, such as [12, 18, 7], to generate a reduced dimensionality manifold to capture the pose variation in our dataset. However, while these previous works produced a pose manifold from PrimeSense or Kinect skeletons, we use the less noisy VICON skeletons derived from motion capture measurements.

A simple approach to view-invariant manifold learning would be to generate one manifold per view and operate on each independently to exhaustively seek a solution. In [22], Zhao et al. learnt a latent space multiview manifold from several images at once, which may be from different modalities, e.g. RGB or RGB-D, with locality alignment using both a supervised and an unsupervised algorithm. This effectively integrated several individual views of a scene into a single manifold. Motion-capture 3D skeletons combine information from multiple cameras and as such are view-independent. Therefore, we generate a view-invariant manifold by applying Diffusion Maps [5] on our motion capture skeletons, which as a result will allow a reduced dimensionality, view-independent model of an action.

We propose a CNN regression-based method to estimate human pose from single RGB images. The view-invariant pose manifold serves as a target space to our pose estimation method, which is trained from groundtruth poses obtained by projection of motion capture skeletons onto the manifold space. Our CNN is made view-invariant through the combined uses of a view-independent manifold and multiple RGB views in its training set.

Skeleton Data Normalization and Manifold Learning – As in [12, 18, 7], before applying Diffusion Maps, we must normalise our data since different subjects come in various shapes and sizes and they also do not perform actions at the same world coordinates. To normalise for translation, we considered the

[1] We used the Optitrack Flex 3 acquisition system and VICON's NEXUS skeleton building software.

centre of the hip ($p^{(hip\ center)}$) as our coordinate centre and normalised the other joint positions relative to it as:

$$p_t^j = p_t^j - p_t^{(hip\ center)} \ , \tag{1}$$

where t is the frame number, p is the joint position, and $j = \{1, 2, \ldots, J\}$ is the joint number for the $J = 39$ joint positions supplied by our motion capture system. To normalise for scaling, we defined a model skeleton as a template and then used its torso, hand and leg sizes for normalising the data as follows:

$$\begin{aligned} torso_{ratio} &= torso\ size_{template} / torso\ size_{pose} \ , \\ p_t^j &= p_t^j * torso_{ratio}, \quad \text{where } j \in torso \ , \end{aligned} \tag{2}$$

$$\begin{aligned} hand_{ratio} &= hand\ size_{template} / hand\ size_{pose} \ , \\ p_t^j &= p_t^j * hand_{ratio}, \quad \text{where } j \in hand \ , \end{aligned} \tag{3}$$

$$\begin{aligned} leg_{ratio} &= leg\ size_{template} / leg\ size_{pose} \ , \\ p_t^j &= p_t^j * leg_{ratio}, \quad \text{where } j \in leg \ . \end{aligned} \tag{4}$$

To normalise for rotation, we applied Procrustes analysis. This approach was not used for translation and scaling since the center of human body shape is different with the center computed by Procrustes, e.g. different body parts have different scale ratios while the Procrustes analysis method scales the whole shape at once.

Finally, we applied Diffusion Maps [5] to reduce the dimensionality of our data by selecting the manifold's first $N = 5$ dimensions to represent 95% of the total variance of our original data. This exceeds the 3 dimensions used in [7], but our more complex movement requires more dimensions to describe it. While in previous works a Robust Diffusion Map algorithm was used [12, 18, 7], the robust extension was not required in our case because we work with skeletons extracted from motion capture data, which do not suffer from the same level of noise as the Kinect or PrimeSense skeletons used in these previous works.

Proposed Network Architecture – The overall structure of the network is shown in Figure 3. We propose a regression CNN that can exploit the geometric relationship between different body parts to allow the estimation of 3D pose in a reduced-dimensionality manifold space. To this end, and to prevent overfitting on subject appearance during the training of our CNN, we propose to explore body joint heatmaps and a set of 2D vectors which encode the orientation and location of body limbs (as limb-maps) as input, instead of RGB images. This has the added benefit of reducing our data input size. We apply OpenPose [3] to all our images, delimited by the bounding box containing the subject, to generate 26 body joint heatmaps and 52 body limb-maps. All the images are at first zero-padded and resized into 244×244 pixels to remove scale variations.

By injecting priors on position and structure of body parts to the CNN, we are able to estimate the 3D reduced pose of each person in manifold space accurately since we force the CNN to extract the features from our desired regions.

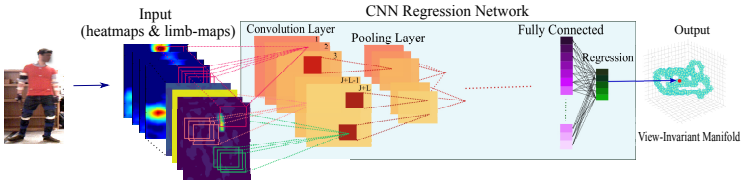


Fig. 3. The overall structure of the network to estimate high-level view-invariant human pose in our view-invariant manifold.

Our input contains $J + L$ channels, (J is the number of body joints, and L is the number of limbs) where each channel describes one body joint or limb, which leads to the size of the kernels in the first convolution layer to be $J + L$.

After [3], for each joint $j \in \{1, \ldots, J\}$, we produce a heatmap $\mathbf{H}_j$ whose value at pixel position p is

$$\mathbf{H}_j(p) = exp(-\frac{\|p - P_j\|_2^2}{\sigma^2}), \tag{5}$$

where P_j is position of joint j and σ determines the spread of the peak.

For each body limb $l \in \{1 \ldots L\}$, we generate a body limb-map $\mathbf{B}_l$, such that, if j_1 and j_2 are body joints defining a limb, then

$$\mathbf{B}_l(p) = \begin{cases} v & \text{if } p \text{ on limb } l \\ 0 & \text{otherwise} \end{cases} \qquad \text{where} \quad v = \frac{p_{j1} - p_{j2}}{\|p_{j1} - p_{j2}\|} \ . \tag{6}$$

To implement our network, we use ResNet and modify its first and last layers. We replace the first layer with a convolutional layer, with a depth of $J + L$, and the last layer with a regression layer with the size of the manifold dimension. Our mean square error (MSE) loss function computes the difference between the groundtruth X and the 3D reduced pose Y estimated by the proposed method,

$$Loss(X, Y) = \frac{1}{N} \sum_{i=1}^{N} \|x_i - y_i\|_2^2 \ . \tag{7}$$

4 Experimental Results

We perform 3 experiments on the turn-walk action. First we show the importance of our heatmap and limb-map in estimating high-level view-invariant human pose on normal subjects. Then, we probe our method's performance, given single and combined of views at training time, to assess the ability of a CNN to attain view invariance for pose estimation. Finally, we perform movement quality classification (into normal and abnormal) using spatio-temporal modeling.

Experimental Setup – Our experiments were performed under Pytorch on a GeForce GTX 750 GPU, training our pose estimation model for the turn-walk action for 15 epochs with a learning rate of 0.001, and batch size of 10.

Table 1. The MSE for pose estimation when training with different inputs

Error	RGB BB	Depth BB	Depth Silhouette [7]	Heatmap	Limb-map	Heatmap & Limb-map
MSE	0.72	0.70	0.72	0.67	0.67	**0.66**

Our training was on 12 subjects at 53544 frames, and testing was on 5 subjects at 21991 frames. For assessing movement quality, we additionally tested on 6 subjects with stroke, 8 subjects with limp, and 12 subjects with Parkinsons.

For evaluation against the closest possible approach, in our first experiment, we compare against Crabbe et al. [7]. Since their dataset does not contain RGB data, the only possible comparative analysis is for us to apply their method using depth silhouettes generated from our data. In addition, [7]'s simple depth silhouette extraction method is not robust to cluttered environments. Instead, we use OpenPose [3] to obtain better depth silhouettes and do region growing from seeds located at joint positions estimated by OpenPose. For robustness, we only use as seeds the non-occluded torso joints. The same depth and contrast normalisation as in [7] was then applied to the silhouette and its background. Also, [7] used Alexnet to train their model, but for a fairer comparison, we apply ResNet for all our experiments. In the following experiments, since it is not simple to extract a depth silhouette, we *also* use the depth bounding box (Depth BB) of the subject to compare against. For accuracy, we compute the MSE between groundtruth pose and estimated human pose in our manifold space.

Comparison of Input Features – We train our network with different types of inputs, i.e. RGB bounding box (RGB BB) of subject, depth BB of subject, depth silhouette similar to Crabbe et al. [7] but extracted using OpenPose [3], body joint heatmaps, body limb-maps, and combined heatmap and limb-maps, from all our views, to assess the performance of the proposed method. Table 1 shows that when trained by using combined heatmap and body limb-maps, the network has the least error in estimating high-level pose. As a result, for the rest of our experiments, we only train the network with heatmaps and body limb-maps. The result from Crabbe et al. [7] using depth silhouettes is poorer than when Depth BB is used potentially due to the general difficulty in accurate silhouette extraction. While in [7] the small size of the dataset required simplifying the learning task for the CNN by extracting the depth silhouette as a preprocessing stage, with our dataset and ResNet architecture this is not the case anymore and the depth BB obtains good results. For this reason, for the rest of the experiments, instead of depth silhouettes, we compare against Crabbe et al.'s work with the simpler Depth BB put through the network.

Assessing Single and Combinations of Views – The first four rows of Table 2 report the pose estimation MSE when we train our method each time using single individual views only. View 1 and View 4 are the opposite camera views, View 2 is the camera view from the side, and View 3 is around 45° above View 1. View 3 provides the best result and will hereafter be used as the basis of all other experiments. Furthermore, for all single views, the proposed method performs better than Depth BB.

Table 2. MSE between estimated pose and groundtruth on single and multiple views.

Train set	Test set	Depth BB	Proposed Method
View 1	View 1	0.73	**0.67**
View 2	View 2	0.76	**0.71**
View 3	View 3	0.70	**0.63**
View 4	View 4	0.73	**0.65**
View 3	View 1	1.13	**1.02**
	View 2	1.42	**1.36**
	View 4	1.10	**1.04**
Average (inc View 3)		1.21	**1.14**
Views 3,4	View 1	1.07	**0.95**
	View 2	1.42	**1.36**
	View 3	1.59	**0.64**
	View 4	0.70	**0.64**
Average		1.19	**0.89**
Views 1,3,4	View 1	0.68	**0.67**
	View 2	1.40	**1.20**
	View 3	1.66	**0.64**
	View 4	0.68	**0.65**
Average		1.10	**0.79**
Views 2,3,4	View 1	1.09	**0.95**
	View 2	**0.73**	0.75
	View 3	1.63	**0.62**
	View 4	0.69	**0.62**
Average		1.03	**0.73**
Views 1,2,3,4	View 1	**0.69**	**0.69**
	View 2	0.75	**0.72**
	View 3	1.73	**0.64**
	View 4	0.69	**0.64**
Average		0.96	**0.67**

Row 5 in Table 2 illustrates that training from a single view only, even if it is the best view, is not sufficient if the test data comes from others views. We only show the case for the best single training view, i.e. View 3, while the results for the other train/test combinations are similar or worse. The remaining rows in Table 2 show the MSE between the estimated pose and groundtruth for different combinations of views for Depth BB and the proposed method. Again, we only show sample combinations that are rooted in View 3, including the combination of all views, while other results remain quite similar. These results indicate that the proposed method can maintain a high accuracy when more views are provided and has learnt well to distinguish between views.

Quality of Movement Assessment – We need to examine if the high-level view-invariant poses extracted by the proposed method are suitable for assessing the quality of human movement. For spatio-temporal analysis of the movement, we apply the framework proposed in [12] which generates two statistical models of normal pose and dynamics. A frame is classified as normal or abnormal depending on how far away from these models it is, based on an empirically determined threshold on log-likelihood. We test on both normal and abnormal sequences which contain all normal (resp. abnormal) frames.

For sequences with abnormal movements, no motion capture skeleton data is available. It is therefore not possible to measure MSE error for pose estimation

Table 3. Frame classification performance for normal and abnormal sequences

	Normal			**Stroke**		**Limp**		**Parkinson**		**All sequences**	
	TN	FP	Specificity	TP	FN	TP	FN	TP	FN	Precision	Recall
Depth BB	3795	4095	0.48	**3793**	**1497**	**5665**	**2663**	**8939**	**3950**	0.81	**0.69**
Proposed	**4780**	**3110**	**0.60**	3540	1750	4592	3736	6699	6190	**0.82**	0.55

Table 4. Percentage of frames classified as normal by the pose/dynamics models

	Normal	**Stroke**	**Limp**	**Parkinson**
Depth BB	79% / 55%	79% / 30%	78% / 34%	80% / 32%
Proposed Method	86% / 65%	78% / 35%	81% / 48%	85% / 51%

due to lack of groundtruth. However, we may still assess our method's performance indirectly through movement quality analysis, which depends directly on the quality of pose estimate, as highlighted in [7].

We note from Table 3 the overall poorer performance of the movement quality assessment method compared to previous uses of it in [12, 18, 7]. This may not necessarily indicate a poor performance of pose estimation, but rather be due in large part to the method being designed for modelling and assessing the quality of single movements, while in our case we consider a more complex action made up of two distinct basic movements (walk-turn). Since improving on this method is not the topic of the present study, we leave this to future works, and we focus on comparing pose estimates from the Depth BB and proposed methods.

Table 3 shows that the specificity for the proposed method to estimate pose of normal sequences is higher at 0.60 than for depth BBs at 0.48, which implies that the estimated poses are close to motion-captured data. Table 4 shows that the movement analysis modelling mostly finds pose to be normal, while the dynamics is particularly abnormal in all abnormal sequences. This is in line with our scenarios where all three abnormality types mostly imply abnormal dynamics with relatively normal poses. The depth BB approach tends to yield more abnormal pose outcomes than ours, in line with the results of previous experiments (Tables 1 & 2). This may contribute to explaining its poorer classification results on normal sequences in Table 3 and its better results on abnormal sequences.

5 Conclusions

We proposed a CNN regression method to extract high-level view-invariant pose and applied it to asses the overall quality of human movement. We also introduced a new multiview, multimodal human movement dataset to evaluate the performance of the proposed method and which we hope will be of use to the rest of the community. The implication of our approach is that a CNN may learn to estimate high-level pose from arbitrary view points. We also demonstrated the superiority of RGB-derived heatmaps and limb-maps as input data for pose estimation, over depth data. For future work, we plan to build on our method to produce a multiview framework that may combine any number of arbitrary view points for a more robust pose estimation.

References

1. Baptista, R., Demisse, G., Aouada, D., Ottersten, B.: Deformation-based abnormal motion detection using 3d skeletons. In: IPTA. pp. 1–6 (2018)
2. Buckley, T., Pitsikoulis, C., Hass, C.: Dynamic postural stability during sit-to-walk transitions in parkinson disease patients. Movement Disorders **23**(9), 1274–1280 (2008)
3. Cao, Z., Simon, T., Wei, S.E., Sheikh, Y.: Realtime multi-person 2d pose estimation using part affinity fields. In: CVPR. pp. 7291–7299 (2017)
4. Chaaraoui, A.A., Padilla-López, J.R., Flórez-Revuelta, F.: Abnormal gait detection with rgb-d devices using joint motion history features. In: FG. vol. 7, pp. 1–6 (2015)
5. Coifman, R.R., Lafon, S.: Diffusion maps. ACHA **21**(1), 5–30 (2006)
6. Comelia, C.L., Stebbins, G.T., Brown-Toms, N., Goetz, C.G.: Physical therapy and parkinson's disease: a controlled clinical trial. Neurology **44**(3), 376–376 (1994)
7. Crabbe, B., Paiement, A., Hannuna, S., Mirmehdi, M.: Skeleton-free body pose estimation from depth images for movement analysis. In: ICCVW. pp. 70–78 (2015)
8. Culhane, K., Oconnor, M., Lyons, D., Lyons, G.: Accelerometers in rehabilitation medicine for older adults. Age and Ageing **34**(6), 556–560 (2005)
9. Li, M.H., Mestre, T.A., Fox, S.H., Taati, B.: Vision-based assessment of parkinsonism and levodopa-induced dyskinesia with pose estimation. Journal of Neuroengineering and Rehabilitation **15**(1), 97 (2018)
10. Li, Y., Chai, X., Chen, X.: End-to-end learning for action quality assessment. In: Pacific Rim Conference on Multimedia. pp. 125–134 (2018)
11. Liao, Y., Vakanski, A., Xian, M.: A deep learning framework for assessment of quality of rehabilitation exercises. arXiv preprint arXiv:1901.10435 (2019)
12. Paiement, A., Tao, L., Hannuna, S., Camplani, M., Damen, D., Mirmehdi, M.: Online quality assessment of human movement from skeleton data. In: BMVC. pp. 153–166 (2014)
13. Parmar, P., Morris, B.T.: What and how well you performed? A multitask learning approach to action quality assessment. arXiv preprint arXiv:1904.04346 (2019)
14. Parmar, P., Tran Morris, B.: Learning to score olympic events. In: CVPRW. pp. 20–28 (2017)
15. Pirsiavash, H., Vondrick, C., Torralba, A.: Assessing the quality of actions. In: ECCV. pp. 556–571 (2014)
16. Raso, I., Hervás, R., Bravo, J.: m-physio: personalized accelerometer-based physical rehabilitation platform. In: MUCSSST. pp. 416–421 (2010)
17. Som, A., Anirudh, R., Wang, Q., Turaga, P.: Riemannian geometric approaches for measuring movement quality. In: CVPRW. pp. 43–50 (2016)
18. Tao, L., Paiement, A., Damen, D., Mirmehdi, M., Hannuna, S., Camplani, M., Burghardt, T., Craddock, I.: A comparative study of pose representation and dynamics modelling for online motion quality assessment. CVIU **148**, 136–152 (2016)
19. Toosizadeh, N., Mohler, J., Parvaneh, S., Sherman, S., Najafi, B.: Motor performance assessment in parkinson's disease: association between objective in-clinic, objective in-home, and subjective/semi-objective measures. PloS **10**(4) (2015)
20. Tran, D., Bourdev, L., Fergus, R., Torresani, L., Paluri, M.: Learning spatiotemporal features with 3d convolutional networks. In: ICCV. pp. 4489–4497 (2015)
21. Vakanski, A., Jun, H.p., Paul, D., Baker, R.: A data set of human body movements for physical rehabilitation exercises. Data **3**(1), 2 (2018)
22. Zhao, Y., You, X., Yu, S., Xu, C., Yuan, W., Jing, X.Y., Zhang, T., Tao, D.: Multi-view manifold learning with locality alignment. PR **78**, 154–166 (2018)