Skeleton-free body pose estimation from depth images for movement analysis

Ben Crabbe, Adeline Paiement, Sion Hannuna, Majid Mirmehdi
Department of Computer Science, University of Bristol
Bristol, BS8 1TH, UK
csatmp@bristol.ac.uk

Abstract

In movement analysis frameworks, body pose may often be adequately represented in a simple, low-dimensional, and high-level space, while full body joints locations constitute excessively redundant and complex information. We propose a method for estimating body pose in such high-level pose spaces directly from a depth image without relying on intermediate skeleton-based steps. Our method is based on a convolutional neural network (CNN) that maps the depth-silhouette of a person to its position in the pose space. We apply our method to a pose representation proposed in [18] that was initially built from skeleton data. We find our estimation of pose to be consistent with the original one and suitable for use in the movement quality assessment framework of [18]. This opens the possibility of a wider application of the movement analysis method to movement types and view-angles that are not supported by the skeleton tracking algorithm.

1. Introduction

Body pose recovery represents a fundamental and extensively researched challenge in computer vision, as its estimation is essential to a large number of tasks, ranging from activity recognition [33, 2, 7] to movement quality analysis [18, 8, 13, 27, 19]. Its actual representation is highly dependent on the task and when single actions or movements are studied, it is usually simplified and tailored to best represent their ranges of variations. Such tailoring of body pose representation typically involves extraction of low-level features from images followed by dimensionality reduction to discard redundancy and retain only relevant information. With the advent of affordable depth cameras [34] and the associated skeleton trackers [25], 3D joint location has become a popular low-level feature [17, 11]. However, skeletons suffer from a number of limitations, notably a restricted range of viewing-angles and a poor tolerance to self-occlusion, which limit considerably the range of movements that can be analysed. This work proposes an alternative to the use of a skeleton tracker in a movement quality assessment framework. We design a system based on a Convolutional Neural Network (CNN), that bridges the gap between depth images and a high-level representation of pose in a reduced dimensional space. We assume that the pose representation and its associated space (hereafter referred to as ”pose representation space”, or for brevity ”pose space”) have already been created during a movement model's learning phase to support a movement analysis task. This pose space may have been built using any number of intermediate steps from depth image to high-level feature, possibly involving skeletons. We propose a direct mapping between the depth image space and the pose space that does not require computation of these intermediate steps during the movement model's testing phase. Such direct mapping offers the possibility of exploiting the movement model from more general conditions than available during the training phase *e.g*. due to the aforementioned skeleton restrictions.

Next in Section 2 we review related work in body pose estimation for movement analysis. We describe our proposed method in Section 3 and present experimental results in Section 4. Section 5 concludes this paper and suggests future works.

2. Background

Pose representation for movement analysis – As mentioned in Section 1, previous works that extract a pose representation from RGB or depth images in order to build a model of movement or activity, typically compute low-level features first, then retain the information relevant to build their model using a dimensionality reduction technique. A common low-level feature is the silhouette of a person, as in [1] where Brand performs action recognition supported by a pose and dynamics space obtained from silhouettes. This space is in effect a Hidden Markov Model (HMM), learnt from silhouette images using entropy minimisation. In [5] a pose space that captures the variations of the silhouette within a given action is obtained by Local Linear Embedding (LLE) [23] of the silhouettes. This work is quite similar to ours, since it then learns a mapping from the visual input (silhouette) to the LLE pose space. However, this mapping is learnt using a Generalized Radial Basis Function (GRBF) interpolation framework [20], and both the pose space and mapping are subject specific as it is derived from a low-level feature (silhouette) that is not subject invariant. This issue is addressed in [6] by devising a fused multi-subjects space and mapping. In our work, we use a non-subject specific pose space and, although we also use silhouettes, our mapping can learn invariance to the person's appearance.

Depth information may be added to the silhouette for increased accuracy and robustness of the pose representation, as in [31] where Uddin *et al.* create a pose space for modelling and recognition of different types of gait, by extracting low-level features from depth silhouettes using Local Directional Pattern (LDP) and then applying a Principal Component Analysis (PCA).

Skeleton data is an increasingly popular source of low-level features, mostly due to its invariance to subject appearance and it recently becoming easily and cheaply available from the Kinect camera and SDK. Vemulapalli *et al.* [32] used them to recognise actions from a Lie group pose space derived from the relative geometry between all pairs of body segments. In [18], Paiement *et al.* analysed the quality of movements using a pose representation built from a skeleton low-level feature embedded in a high-level 3D space computed as a Diffusion Map [3]. This pose representation was passed to a continuous-state HMM to assess quality of movement via a continuous score that assessed deviation from the range of ”normal” movements. This method was further assessed in [29]. We apply our work to Paiement *et al.*'s pose representation and propose in Section 3 a mapping from depth images to this pose space. Our method is evaluated in Section 4.3 in the context of their movement quality assessment method.

Mapping from visual features to pose space – A few methods have been proposed for learning a mapping from visual features to a low-dimensional pose space that is prelearnt using skeleton data. Tangkuampien and Suter in [28] used LLE to learn a mapping between an action specific pose space built using motion capture (MoCap) data and a silhouette based pose space. Both spaces were created from the same movement sequences using Kernel Principle Component Analysis (KPCA) [24]. Rosales *et al.* [22] estimated 3D body pose from multiple views simultaneously while recovering the positions of the cameras. A Specialized Mapping Architecture (SMA), trained from MoCap data, mapped silhouettes in each camera view to a space of 2D poses, generating several hypotheses per image. The set of 2D pose hypotheses, for all camera views, was then used to estimate the most likely positions for the cameras and 3D pose for the body in an Expectation Maximization framework. In [21], Rosales and Sclaroff trained multi-layer perceptron neural networks to map visual features (extracted from silhouettes) to clusters of similar body poses. The clusters were obtained by unsupervised clustering of 2D body joints, and one neural network was trained per cluster. Body pose was estimated by selecting the mapping that produced the most likely hypothesis. Using the recent developments in deep neural networks, several works [30, 10, 16] show that it is now feasible to have a single CNN learn both the visual features best suited and the mapping of all poses. This removes the need for clustering and separate networks leading to a simpler, easily adaptable solution.

Deep neural networks for pose estimation – Toshev and Szegedy [30] and Gkioxari *et al.* [10] used deep neural networks to estimate body pose as 2D skeletons from single RGB images. In [30], a hybrid holistic and local approach first regresses all joint positions using a CNN, before refining the 2D position estimates using a cascade of CNNs that focus on individual joints and small patches around them. In [10], a single R-CNN [9] is used to regress the 2D joint positions, with the possibility to combine detection of the person and classification of their action. Similarly, Li and Chan [16] integrate the regression of 3D joints locations and the individual detections of the joints from single RGB images using a single CNN. The location of each joint is regressed relative the a parent joint, in order to integrate and learn the relationship between the joints.

These methods focus on detecting full-body joints in general cases that include all types of movements, thus learning more information through the CNNs than we do in our movement specific scenario. In many movement analysis applications, the range of body pose variations that need to be recovered is limited, thus full-body joints are considered unnecessarily redundant information. Therefore, although we also base our method on a CNN, we use it to recover a simplified, lower-dimensional pose representation. This allows us to successfully train our system on a significantly reduced amount of data. This low-dimensional pose representation may be mapped to skeleton configurations – again specific to the movement – however computing such mapping is not within the scope of this work. To the best of our knowledge, this is the first time that a CNN is used to recover body pose in a low dimensionality and movement specific pose representation context.

3. Methodology

Our proposed method is outlined in Fig. 1. It maps depth images to a pose space, *e.g.* of [18], by first extracting and pre-processing a depth silhouette of a person, then passing it to a CNN that performs a low-level visual feature extraction followed by a regression to the output space. These two main steps – silhouette extraction and CNN based regression – are detailed next.

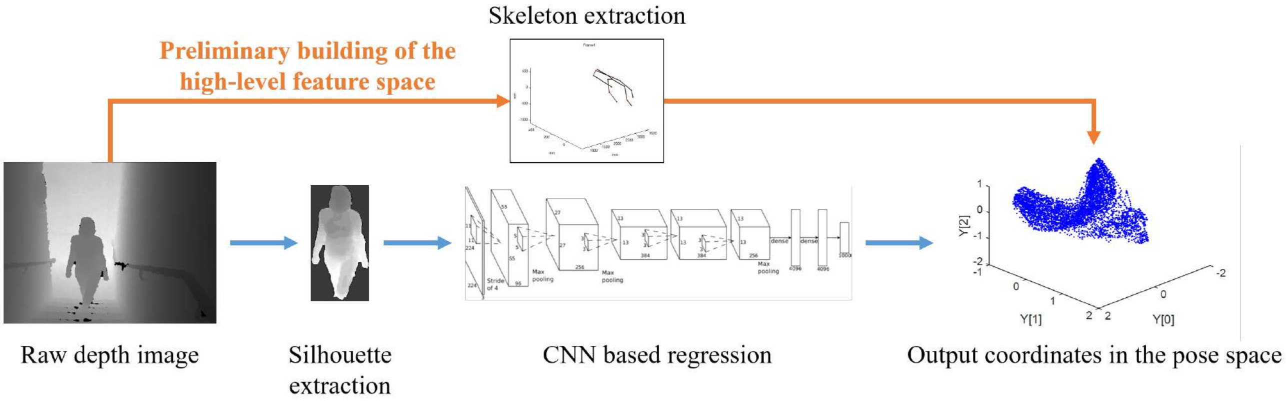

Figure 1: Overview of the proposed method (blue pipeline). The orange pipeline denotes the preliminary building of the pose space, that may be based on skeleton data, and which is a pre-requisite of our method.

3.1. Silhouette extraction and pre-processing

The pre-processing steps are designed to facilitate the regression of the CNN. First, a hole filling algorithm is applied to close the holes of missing depth values that are typically found in Kinect depth images. We use a simple propagation of neighbouring values, iteratively filling holes using the maximum value of a small neighbourhood. This method was found to produce comparable results to more sophisticated methods while being much faster.

Second, the silhouette of a person is extracted from a depth image to focus the network's task on the pose estimation. Although CNN's have demonstrated capacity to detect people in depth images [10], the combined task of detecting people and estimating their pose is more challenging than single pose estimation and it may require a larger amount of training data than available from [18] and the associated dataset [12] that we use in our experiments. We extract silhouettes using the background subtraction method of [35] implemented in the BGSLibrary [26] and we refine them by selecting the largest blob and applying small amounts of erosion and dilation. We use these silhouettes to produce depth-silhouette images that contain depth values inside the person's silhouette and a 0-value background.

Since the CNN requires a normalised input, we crop the depth-silhouette image to a fixed width/height ratio of 0.504, tightly centered around the silhouette, then we rescale the image patch to a size of 227x227 pixels. We also maintain the shoulders (found by considering silhouette width) at a constant height in the image patch. This gives robustness to incomplete silhouettes where, for example, the feet or the top of the head are missing due to occlusion or failure of the background subtraction algorithm. As long as the main body parts are maintained in a fixed location in the image patch, such missing feet or heads where found to have little impact on the accuracy of the final result.

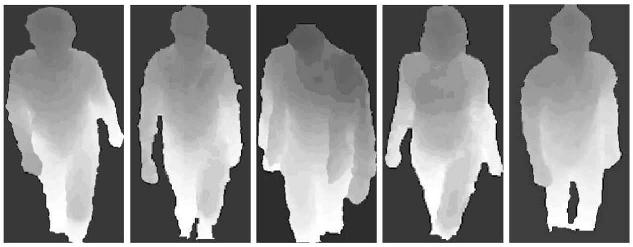

Figure 2: Examples of normalised depth-silhouette patches

We also normalise the depth values to produce a depth silhouette that is independent of the position of the person in the scene and that only denotes pose variations due to movement. Again, the small size of the used dataset would not have allowed the CNN to learn this invariance. Thus, the mean depth value of a small region near the subject's waist that represents the distance of the person to the camera, is subtracted from the depth silhouette.

Our last step is to enhance the contrast of the depth silhouette patch using histogram equalisation. This facilitates exploitation of depth values inside the silhouette by the convolution layers of the CNN, which were found otherwise to focus exclusively on the stronger gradients between the outline of the silhouette and the background.

Examples of normalised depth-silhouette patches are displayed in Fig. 2. They serve as inputs to the CNN for regression.

3.2. Regression

The regression step aims to map the depth-silhouette patches produced in Section 3.1 to the pose space illustrated on the right of Fig. 1. We use a CNN to perform this step, thus benefiting from its inherent ability to extract visual features from the depth-silhouette image patch, which are optimised for the regression task.

Given the low amount of training data available for our experiments, we fine-tune a pre-built network in order to reduce the chance of over-fitting. We use the AlexNet network [15] provided with the Caffe Library [14]. Since this network was originally designed for a classification task from RGB images, we replace its final fully connected layer with a 3-element (still fully connected) layer, which produces the coordinates of a point in the (3D) pose space. A 3-channel image containing 3 duplicated copies of the depth-silhouette patch is used in place of the RGB input image. We found that the usual subtraction of the mean image over the training dataset affected negatively the results so we discarded this step.

Our training dataset contains depth-silhouette image patches and the associated points in the pose space obtained from skeleton data. We augmented this set by flipping the depth images and skeleton joints horizontally. The frames of the movement sequences were shuffled to avoid providing the network with batches of consecutive frames and, therefore, nearly identical poses. We also maximise the amount of training data by cross-validating the network on two left-out subjects.

Training of the CNN was performed using Caffe's Euclidean (L2) loss function

$$\mathcal{L} = \frac{1}{2N} \sum_{i=1}^{N} \|x_i - y_i\|_2^2 \tag{1}$$

where x_i and y_i are the i^{th} coordinates of $\mathbf{x}$ the network's estimated point in the pose space, and $\mathbf{y}$ the ground-truth pose vector, respectively. N is the pose space dimensionality (in our case $N = 3$). The full network is trained from pre-trained weights in the two first convolution layers, and random initial weights in the rest of the network, using the adaptive learning rate method AdaGrad [4]. We found the weight decay did not affect the results significantly and we retained a standard value of 0.005. We trained for 50000 iterations with a batch size of 25, and observed no over-fitting effect.

3.3. Personalisation

In addition to the general network trained on multiple subjects, we also present results of personalised models that are fine-tuned on the data of a specific person. As will be demonstrated in Section 4, this personalisation tends to improve the accuracy of the results, because some dimensions of the used pose space encompass some subject and style specific aspects of pose and movement. Small single-subject training sets are prone to cause over-fitting, thus the number of training iterations needs to be chosen according to the number of pose samples. We found that training the network for 10 epochs (*i.e.* 10 times through the training data) provided the best results.

4. Experimental results

We apply our proposed method to the estimation of body pose in the frame of the pose representation of [18] and its application to movement quality assessment. We evaluate both the accuracy of predictions in the pose space of [18], and their suitability as an input to the movement model and movement quality assessment framework [18].

4.1. Dataset

In our experiments we use the SPHERE-Staircase2014 dataset [12] that was introduced in [18]. This dataset comprises 48 depth video sequences of 12 subjects walking up stairs, captured by an Asus Xmotion RGB-D camera placed at the top of the stairs in a frontal and downward-looking position. The dataset divides into a training set made up of 17 sequences of normal walking from 6 subjects, and a testing set containing 31, both normal and abnormal, sequences from the remaining 6 subjects. Abnormal sequences include sequences that contain one or two temporary freezes of the person, and subjects using always the same leg to walk up a step.

All sequences come with skeleton data, and the skeletons of the training set are used to build the pose space using Diffusion Maps as described in [18]. All other skeletons may be projected into the pose space using the Nyström extension (see [18]), thus providing a ground-truth to assess the accuracy of our pose estimation. We cross-validate the method on two left-out subjects, training the CNN on 10 subjects and validating on the remaining two. For that purpose, we use the sequences (both normal and abnormal) of all subjects as potential training data rather than the original division of the dataset in [18]. Next, for clarity, we refer to the original division of the dataset into training and testing sets as the ”movement model's training/testing set”, while our division is the ”CNN's training/testing set”, or for brevity ”training/testing set”.

4.2. Quantitative accuracy

We measure the accuracy of our pose estimation by computing the estimation error as the Euclidean L2 distance (1) between ground-truth and predicted points in the pose space. Table 1 reports the mean and standard deviation of the errors for both normal and abnormal video sequences of each subject, using either the general or personalised models. Three examples of normal and abnormal sequences are shown in Figs. 3 and 4.

In general we found close agreement between the ground-truth and predicted positions in the pose space. For

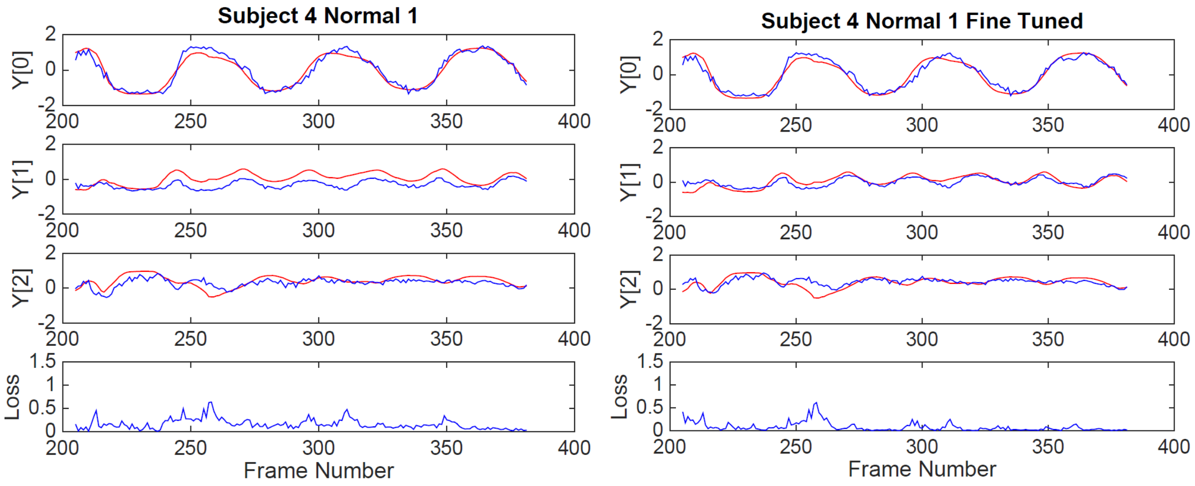

Figure 3: Example of estimated pose using the general (left) and personalised (right) models in the case of a normal movement. The first 3 rows show the 3 dimensions of the pose space, with estimated pose in blue and ground-truth pose computed from skeletons in red. The estimation error is displayed in the bottom row.

Subject	Sequence type	Average mean / std error	
		General model	Personalised
Subject 1	Normal	0.12 / 0.16	0.12 / 0.15
Subject 2	Normal	0.39 / 0.37	0.23 / 0.25
Subject 3	Normal	0.16 / 0.28	0.15 / 0.27
Subject 4	Normal	0.16 / 0.11	0.09 / 0.10
Subject 5	Normal	0.16 / 0.15	0.08 / 0.08
Subject 6	Normal	0.20 / 0.19	0.09 / 0.10
Subject 7	Abnormal	0.12 / 0.14	0.11 / 0.14
Subject 8	Abnormal	0.05 / 0.07	0.08 / 0.10
Subject 9	Normal	0.11 / 0.10	0.07 / 0.09
	Abnormal	0.09 / 0.09	0.08 / 0.09
Subject 10	Abnormal	0.19 / 0.24	0.16 / 0.24
Subject 11	Normal	0.13 / 0.11	0.05 / 0.05
	Abnormal	0.15 / 0.12	0.09 / 0.08
Subject 12	Normal	0.19 / 0.25	0.17 / 0.24
	Abnormal	0.16 / 0.13	0.11 / 0.10
All	**Normal**	**0.18 / 0.23**	**0.12 / 0.17**
	Abnormal	**0.13 / 0.16**	**0.11 / 0.15**
	All	**0.16 / 0.19**	**0.11 / 0.16**

Table 1: Pose estimation error: mean and std of errors between estimated and ground-truth pose space coordinates.

the general models it was observed that in some sequences the predictions of the 2nd and 3rd coordinates would match the form of the ground truth but at a slight offset, as illustrated in the second row of Fig. 3. It was found that the personalised models were able to correct this, leading to an average reduction in error of 0.0436. We believe this is due to these dimensions of the pose space encompassing some of the personal style variations of the individuals.

The accuracy of the skeleton estimation system [25] is limited outside an optimum range of 1.2-3.5 m. The greatest errors of our proposed method were found at the beginning and end of sequences, when they are slightly outside this range. In most of these cases, studying the original skeleton showed there to be a clear miss-measurement by the skeleton tracker [25]. This was also found to be the case in some frames where the subject's trailing leg was occluded by their leading leg. In the majority of these situations our method estimated a pose that appeared consistent when compared with similar images that had correctly labelled skeletons. Thus in such cases we conclude that the estimations of our method are more accurate than that of the skeleton tracker. Instances where our method is actually accountable for large errors (*i.e.* with accurate skeletons) generally occurred in the fringes of the normal range of poses, where there are not enough examples in our training set from which the network can infer. We expect a larger and more evenly distributed training dataset could reduce the size of these errors.

4.3. Application to movement quality assessment

Next, we assess the suitability of our pose estimation for use in the movement analysis method of [18, 29]. This test is necessary to ensure that the level of noise in the estimated values is acceptable and does not hinder the movement analysis. In [18, 29], this analysis aimed at quantifying the quality of movements. Thus, we evaluate our method both on the normal and abnormal sequences of Table1, in order to

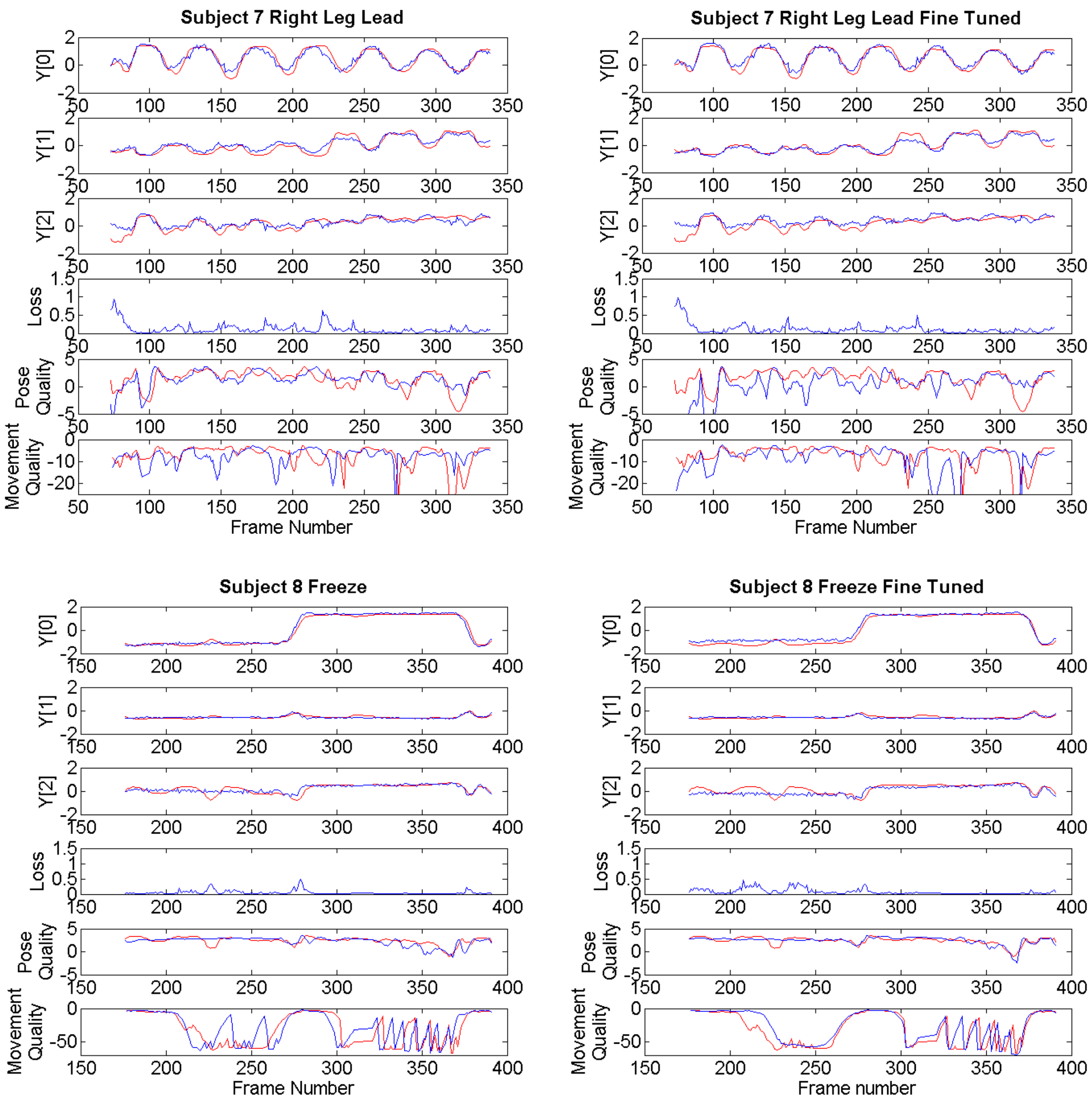

Figure 4: Example of estimated pose using the general (left) and personalised (right) models in the cases of a movement with the "right leg lead" (top) and "freeze" (bottom) abnormalities. The first 3 rows show the 3 dimensions of the pose space, with estimated pose in blue and ground-truth pose computed from skeletons in red. The estimation error is displayed in the 4th row. The two bottom rows display the pose and dynamics scores of the movement quality analysis of [18].

verify its effect on the performance of the movement quality assessment method. Since the subjects of the movement model's training set were used to produce the movement model of [18, 29], we only use subjects of the movement model's testing set (Subjects 7 to 12) to perform this analysis. We apply a temporal smoothing over a 5-frame windows to the estimated coordinates in the pose space, equivalent to the smoothing of the skeleton joints coordinates used in [18].

As in [29], we produce the ROC curve of frame classification accuracy (Fig. 5) and compute the area under the curve (AUC), reported in Table 2. The AUC values ob-

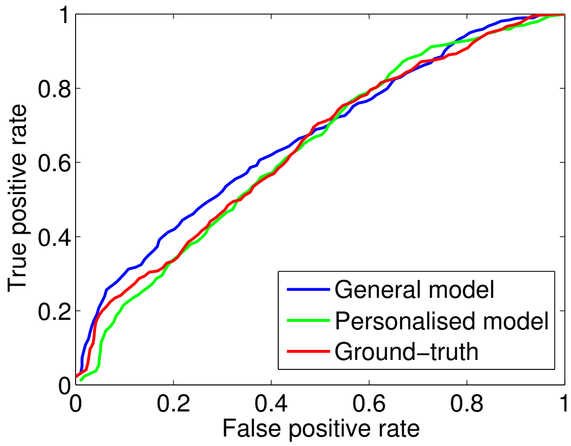

Figure 5: ROC curve of abnormal frame detection when using our estimated and the ground-truth poses. Blue: estimated pose from our general model, green: estimated pose from our personalised model, and red: ground-truth pose.

General model	Personalised model	Ground-truth
0.67	0.63	0.64

Table 2: Area under the ROC curve for abnormal frame detection of [18] using our estimated pose representation.

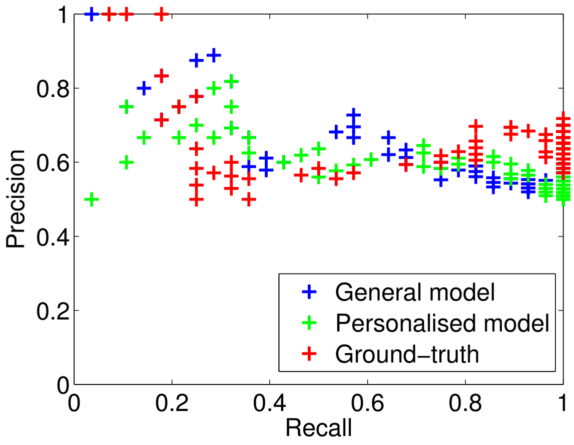

Figure 6: Precision and Recall values of abnormal event detection using our estimated and the ground-truth pose. Blue: estimated pose from our general model, green: estimated pose from our personalised model, and red: ground-truth pose.

tained using the estimated and ground-truth locations in the pose space are consistent, and our estimation did not hinder the movement analysis method of [18]. We also compute the precision and recall values for abnormal event detection when varying the detection threshold, displayed in Fig. 6. Again, the results are similar when using the estimated and original pose representations. We conclude from this test that our method produces pose estimates with an accuracy that is suitable for the movement analysis of [18].

4.4. Timing

We implemented our method on a GeForce GTX 750 GPU and a Linux operating system. Training our network takes just under 7 hours, and testing is performed in real time at near 100 fps. The average forward pass time of the CNN is 9.7 ms.

5. Conclusion and future work

We proposed a method for direct mapping from depth image to a simplified pose representation embedded in a low dimensional space that is suitable for movement analysis tasks. This method is based on depth silhouettes and a CNN to perform regression to the pose space. We applied it to the pose representation of [18] and its application to assessing movement quality. We found that the accuracy of our method was suitable for the movement quality analysis task. This accuracy was attained with much less training data than usually required for pose estimation methods that aim to recover the position of all body joints. We argue that such over-complete representation of pose is not necessary for most movement analysis which typically exploits a simpler and less redundant pose representation. Our method is suitable for such cases and, by providing a direct mapping from depth image to the simplified pose space, it eliminates the need to compute and rely on noisy skeleton data.

The replacement of skeleton based intermediate steps from depth images to a high-level pose representation space presents significant advantages that will be explored in future work. First, it may allow the exploitation of viewing-angles that were not available during the movement model's training phase due to skeleton restrictions. Second, it may enable a more general exploitation of the movement-modelling framework through the handling of movement types impossible to capture using skeleton due to their high level of self-occlusion, such as bending to reach down. Such movement models and their pose representations may be built from other sources of data, *e.g.* MoCap, and then be used with depth images during their testing phase.

Other future work includes further evaluation of our framework, notably against other methods for mapping visual features to a high-level pose space, *e.g.* the LLE based mapping of [28]. The use of MoCap data, as in *e.g.* [21, 22], as opposed to the noisy Kinect skeletons, for building a pose space is a change we expect to improve performance in our system. Similarly rendering synthetic training data from multiple angles as in [25] would be a smart way of improving our viewing angle tolerance.

Acknowledgements

This work was performed under the SPHERE IRC funded by the UK Engineering and Physical Sciences Research Council (EPSRC), Grant EP/K031910/1.

References

[1] M. Brand. Shadow puppetry. *Proceedings of the Seventh IEEE International Conference on Computer Vision*, 2, 1999.

[2] R. Chaudhry, F. Ofli, G. Kurillo, R. Bajcsy, and R. Vidal. Bio-inspired dynamic 3D discriminative skeletal features for human action recognition. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pages 471–478, 2013.

[3] R. R. Coifman and S. Lafon. Diffusion maps. *Applied and Computational Harmonic Analysis*, 21(1):5–30, 2006.

[4] J. Duchi, E. Hazan, and Y. Singer. Adaptive Subgradient Methods for Online Learning and Stochastic Optimization. *Journal of Machine Learning Research*, 12:2121–2159, 2011.

[5] A. Elgammal and C.-S. Lee. Inferring 3D body pose from silhouettes using activity manifold learning. In *Computer Vision and Pattern Recognition (CVPR)*, 2004.

[6] A. Elgammal and C.-S. Lee. Separating style and content on a nonlinear manifold. In *Computer Vision and Pattern Recognition (CVPR)*, 2004.

[7] G. Evangelidis, G. Singh, and R. Horaud. Skeletal Quads : Human Action Recognition Using Joint Quadruples. *Pattern Recognition (ICPR), 2014 22nd International Conference on*, pages 4513 – 4518, 2014.

[8] M. Gabel, R. Gilad-Bachrach, E. Renshaw, and A. Schuster. Full body gait analysis with Kinect. *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS*, pages 1964–1967, 2012.

[9] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Computer Vision and Pattern Recognition (CVPR)*, 2014.

[10] G. Gkioxari, B. Hariharan, R. Girshick, and J. Malik. R-CNNs for Pose Estimation and Action Detection. pages 1–8, 2014.

[11] J. Han, L. Shao, D. Xu, and J. Shotton. Enhanced computer vision with Microsoft Kinect sensor: A review. *IEEE Transactions on Cybernetics*, 43(5):1318–1334, 2013.

[12] S. Hannuna, M. Camplani, L. Tao, A. Paiement, D. Damen, T. Burghardt, and M. Mirmehdi. Depth video and skeleton of people walking up stairs. http://dx.doi.org/10.5523/bris.bgresiy3olk41nilo7k6xpkqf, 2014.

[13] M. C. Hu, C. W. Chen, W. H. Cheng, C. H. Chang, J. H. Lai, and J. L. Wu. Real-Time Human Movement Retrieval and Assessment With Kinect Sensor. *IEEE Transactions on Cybernetics*, 45(4):742–753, 2014.

[14] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. In *ACM Conference on Multimedia*, 2014.

[15] A. Krizhevsky, I. Sutskever, and G. E. Hinton. ImageNet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems (NIPS)*, 2012.

[16] S. Li and A. B. Chan. 3D human pose estimation from monocular images with deep convolutional neural network. In *Asian Conference on Computer Vision (ACCV)*, 2014.

[17] R. Lun and W. Zhao. A Survey of Applications and Human Motion. *International Journal of Pattern Recognition and Artificial Intelligence*, 2015.

[18] A. Paiement, L. Tao, S. Hannuna, M. Camplani, D. Damen, and M. Mirmehdi. Online quality assessment of human movement from skeleton data. In *British Machine Vision Conference (BMVC)*, 2014.

[19] M. Parajuli, D. Sharma, D. Tran, and W. Ma. Senior health monitoring using Kinect. *Communications and Electronics (ICCE), 2012 Fourth International Conference on*, pages 309–312, 2012.

[20] T. Poggio and F. Girosi. Networks for approximation and learning. *Proceedings of the IEEE*, 78:1481–1497, 1900.

[21] R. Rosales and S. Sclaroff. Inferring body pose without tracking body parts. In *Computer Vision and Pattern Recognition (CVPR)*, 2000.

[22] R. Rosales, M. Siddiqui, J. Alon, and S. Sclaroff. Estimating 3D body pose using uncalibrated cameras. In *Computer Vision and Pattern Recognition (CVPR)*, 2001.

[23] S. T. Roweis and L. K. Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290(5500):2323–2326, 2000.

[24] B. Schölkopf, A. J. Smola, and K.-R. Müller. Kernel principal component analysis. *Advances in kernel methods: Support Vector Learning*, pages 327–352, 1999.

[25] J. Shotton, R. Girshick, A. Fitzgibbon, T. Sharp, M. Cook, M. Finocchio, R. Moore, P. Kohli, A. Criminisi, A. Kipman, and A. Blake. Real-time human pose recognitiom in parts from single depth images. *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2011.

[26] A. Sobral. BGSLibrary: An opencv c++ background subtraction library. In *IX Workshop de Visão Computacional (WVC'2013)*, Rio de Janeiro, Brazil, June 2013.

[27] E. E. Stone and M. Skubic. Unobtrusive, continuous, in-home gait measurement using the microsoft kinect. *IEEE Transactions on Biomedical Engineering*, 60:2925–2932, 2013.

[28] T. Tangkuampien and D. Suter. Real-Time Human Pose Inference using Kernel Principal Component Pre-image Approximations. *Procedings of the British Machine Vision Conference 2006*, pages 62.1–62.10, 2006.

[29] L. Tao, A. Paiement, D. Damen, M. Mirmehdi, S. Hannuna, M. Camplani, T. Burghardt, and I. Craddock. A comparative

study of pose representation and dynamics modelling for online motion quality assessment. *Computer Vision and Image Understanding*, Under review, 2015.

[30] A. Toshev and C. Szegedy. DeepPose: Human pose estimation via deep neural networks. In *Computer Vision and Pattern Recognition (CVPR)*, 2014.

[31] M. Z. Uddin, J. T. Kim, and T.-S. Kim. Depth video-based gait recognition for smart home using local directional pattern features and hidden Markov model. *Indoor and Built Environment*, 23:133–140, 2014.

[32] R. Vemulapalli, F. Arrate, and R. Chellappa. Human action recognition by representing 3D skeletons as points in a Lie Group. In *Computer Vision and Pattern Recognition (CVPR)*, 2014.

[33] C. Wang, Y. Wang, and A. L. Yuille. An approach to pose-based action recognition. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 915–922, 2013.

[34] Z. Zhang. Microsoft kinect sensor and its effect. *IEEE Multimedia*, 19:4–10, 2012.

[35] Z. Zivkovic. Improved adaptive gaussian mixture model for background subtraction. In *International Conference on Pattern Recognition (ICPR)*, 2004.