15th IAPR International Conference on Machine Vision Applications (MVA)
Nagoya University, Nagoya, Japan, May 8-12, 2017.

Deep Convolutional Neural Networks for Motion Instability Identification using Kinect

Daniel Leightley
Centre for Military Health Research
King's College London
dleightley@ieee.org

Subhas C. Mukhopadhyay, Hemant Ghayvat
School of Engineering and Advanced Technology
Massey University
s.c.mukhopadhyay@massey.ac.nz

Moi Hoon Yap
School of Computing, Mathematics and Digital Technology
Manchester Metropolitan University
m.yap@mmu.ac.uk

Abstract

Evaluating the execution style of human motion can give insight into the performance and behaviour exhibited by the participant. This could enable support in developing personalised rehabilitation programmes by providing better understanding of motion mechanics and contextual behaviour. However, performing analyses, generating statistical representations and models which are free from external bias, repeatable and robust is a difficult task. In this work, we propose a framework which evaluates clinically valid motions to identify unstable behaviour during performance using Deep Convolutional Neural Networks. The framework is composed of two parts; 1) Instead of using the whole skeleton as input, we divide the human skeleton into five joint groups. For each group, feature encoding is used to represent spatial and temporal domains to permit high-level abstraction and to remove noise these are then represented using distance matrices. 2) The encoded representations are labelled using an automatic labelling method and evaluated using deep learning. Experimental results demonstrates the ability to correctly classify data compared to classical approaches.

1 Introduction

There has been significant interest in digital analysis methods for detection and quantification of human motion for use in electronic health interventions [1]. This, in part, is due to the increased availability of low-cost multi-modality marker-less capturing devices. Sensor technology (*e.g.* Microsoft Kinect) offers new dimensions by harnessing multiple techniques such as feature extraction and encoding. This has been observed in the work of Bigy *et al.* [2], they proposed a technique for recognising posture and Freezing of Gait in those with Parkinsons disease to aid in detecting trips and falls within the home. Yang *et al.* [3] implemented a framework that extracts both depth and colour image data from the Kinect to assess the posture of participants when performing standing balance, the framework allowed for detection of subtle directional changes such as postural sway.

A few studies have sought to validate depth sensor technology and its use in the medical domain [4]. Clark *et al.* [5] captured a cohort of participants performing a series of clinically valid balance tests. Kinect and marker-based Vicon data were captured concurrently, with data from both systems filtered and synchronised. The Kinect was found to be highly robust and accurate when compared to the Vicon capture system. Mentiplay *et al.* [1] assessed the validity of the Kinect in tracking gait compared to 3DMA marker-based camera system. The authors found that while the Kinect is not suitable for tracking lower body kinematic data, only measuring spatio-temporal aspects of gait. These works highlight the clinical feasibility of the Kinect to assess human kinematics.

In this work, we propose a framework for unstable performance using deep learning. The framework acquires motion capture (MoCap) data from a single depth sensor. The skeletal stream is divided into five groups; for each group, feature representation techniques encode the MoCap sequence in the spatial and temporal domains. Deep Convolutional Neural Networks (DCNNs) are used to identify when a participant is unsteady in motion performance. This is then fed back to the clinician to offer greater support in developing a personalised rehabilitation programme.

2 Methods

2.1 Feature Generation and Encoding

Recognising the context and behaviour of human motion is not a straightforward task, more so when identifying instability in motion from a diverse range of participants, environments and modalities. Du *et al.* [6] used a Hierarchical Recurrent Neural Network for action recognition, with the concept of dividing the skeleton into joint groups, based on anatomical significance to the motion sequence. Each joint group is represented by a set of features which are encoded. This encoding shows promise in modelling subtle motion differences. Leightley *et al.* [7] derived sixty-three features representing five joint groups, using the joint group concept, these were trained using Support Vector Machines (SVM) to detect mobility impairment. We utilise the joint groups proposed in [6] and extend the number of features derived in [7] by introducing a set of new features. A summary of the feature groups and encoding methodology is presented in Table 1.

There are several pose-based features and measurements which can be extracted from the skeletal stream [8, 9]. However, there are difficulties in identifying the variables which are capable of describing the motion efficiently. Alongside the features presented in Table

Table 1. Summary of encoded features for each group and the corresponding dimensionality of the group feature vector.

Joint Group	Features	Kinect Joints
$F_{LeftArm}$	Left arm Euler Angle, Euclidean distance between the left shoulder and left hand, x and y axis vectors. $Length = \{1 \ldots, 12\}$	$LeftShoulder$, $LeftElbow$, $LeftWrist$, $LeftHand$
$F_{LeftLeg}$	Left leg Euler Angle, Euclidean distance between the left hip and left foot, x and y axis vectors. $Length = \{1 \ldots, 12\}$	$LeftHip$, $LeftKnee$, $LeftAnkle$, $LeftFoot$
$F_{RightArm}$	Right arm Euler Angle, Euclidean distance between the right shoulder and right hand, x and y axis vectors. $Length = \{1 \ldots, 12\}$	$RightShoulder$, $RightElbow$, $RightWrist$, $RightHand$
$F_{RightLeg}$	Right leg Euler Angle, Euclidean distance between the right hip and right foot, x and y axis vectors. $Length = \{1 \ldots, 12\}$	$RightHip$, $RightKnee$, $RightAnkle$, $RightFoot$
F_{Torso}	Torso Euler Angle relative to the body, Euclidean distance between the spine base and head, Body Movement Zone, Body lean angle (relative to the floor with torso as a reference), Centre-of-Mass (between left shoulder, right shoulder, spine mid), x and y axis vectors. $Length = \{1 \ldots, 16\}$	$SpineBase$, $SpineMid$, $Neck$, $Head$, $SpineShoulder$

1, raw MoCap data, x and y coordinates are extracted to describe the postural change with respect to the central axis. In addition, we utilise features proposed in [7] (Euler Angles, Body Lean Angle and Centre-of-Mass). These features represent the motion characteristics, their associated skeleton form and do not reference the environment to enable invariance. To contribute these, and provide further context to the motions being performed we include the Body Movement Zone (BMZ), which represent the space occupied by the participant.

Body Movement Zone: As in [9], we encode the normalised total space volume occupied by the participant full form skeleton over time. This is computed by identifying the total space covered (or occupied) by the skeleton per frame using standard volume meter-squared calculations. For example, if the participant is stable, with little movement, the BMZ value is small, whereas with large variations in motion such as raising of the arm, due to balance instability the value of the BMZ increases.

2.2 Skeletal coordinate system

Data obtained via a MoCap system such as Vicon or Depth Sensor technology is captured within a predefined action space. To undertaken action analysis and classification it is important to place the participant skeletal structure at the centre the coordinate system to become view-invariant. To achieve this, we utilise the normalisation technique proposed in [10], where the root joint (*Hip Centre*) for each frame is subtracted from all other joints of the frame.

2.3 Labelling

In this approach we utilise supervised learning, which requires data to be labelled prior to training. To label the encoded set of features we employ the digitalised automatic labelling methodology proposed in Leightley *et al.* [7]. We refer the reader to [7] for a detailed explanation of the labelling process. We defined two classes, "good" and "unstable" performance, with the aim to identifying instability associated with the latter class. The labelling can be summarised as follows: Each joint group is combined into a single to represent the frame and motion as a whole for labelling. Each motion was then reviewed to ensure that the class assigned is suitable. The SD measures have been selected based on the literature in the field of human physiology and epidemiology.

2.4 Distance matrices representation

In our work, the aim is to develop a feature set which is representative, informative and suitable to train when using a DCNN. Following on from the success of [11], we represent each encoded frame as a Euclidean distance matrices. This results in a set of distance matrices representing each encoded frame which are then passed to the classifier for training.

2.5 Constructing Deep Convolutional Neural Networks

To enable effective, efficient and representative motion classification, a computational model must be detailed and complex to provide the high performance. Approaches have begun to adopt deep, complex and highly representative models, termed DCNNs [12]. These are represented by the number of layers used during the training phase. There is clear interest is utilising DCNNs for recognition, classification and contextualising human motion represented via MoCap. A DCNN is capable of recognising patterns which contain varying degrees of shift, distortion and noise. We utilise this unique characteristic of DCNN to classify unstable motions from the patterns using derived featured in Table 1 and distance matrices.

To generate the model, we follow the proposal in Ijjina *et al.* [13] and create a 4-layer DCNN model (Figure 1). The architecture is shown in Figure 1 with the layer configuration listed in Table 2. We have used a different layer configuration than that proposed in [13] due to the increased dimensionality of the feature vector. Additionally, we set the Dropout rate at 0.5.

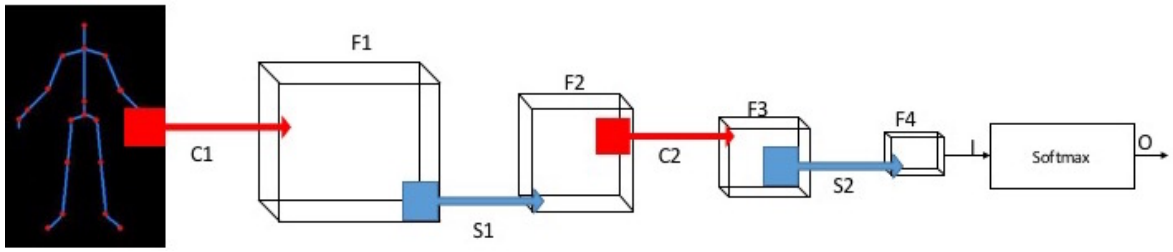

Figure 1. The DCNNs is composed of 4-layers. Each of layer is a combination of convolution, sub sampling and normalization, where the last layer is fully connected. The DCNNs takes a distance matrix as input and propagates using a forward-pass and a backward-pass through the layers activating model neurons at each layer. The output is a DCNN model.

Table 2. Configuration and feature map of the proposed DCNN.

Layer: Configuration		Feature map dimensions	
C1:	4x4 templates	F1:	24x24
S1:	3x3 templates	F2:	12x12
C2:	3x3 templates	F3:	10x10
S2:	3x4 templates	F4:	5x5
I:	300 vector	O:	2 outputs

We train the proposed model using a forward pass and a backward pass approach, similar to that proposed in [13]. Where $C1, C2$ represent the convolution layers, $S1, S2$ represent the sub-sampling layer, $F1, F2, F3, F4$ represent the feature maps generated for each layer and I and O denote the input and output of the network. We also use a Sigmoid transfer function as the activation function in all the neurons and backpropagation algorithm for training.

3 Results: Identification

We evaluated the performance of DCNNs in identifying each motion labelled as "unstable" compared to a range of classical machine learning techniques. We assessed the following approaches with standard cross-validation parameter selection: Support Vector Machines, Random Forests, Boltzmann Machines, Adaptive Boosting, LPBoost, RUSBoost, Total Boost, Bagging and Subspace. For each, a 10-fold cross-validation using the random 'leave-one-out' technique was implemented.

The K3Da Dataset [14] was used in this study, the dataset consists of 500+ motions captured at 30Hz using the Kinect One sensor, these are clinically validated motions based on the Short Physical Performance Battery test [15], it includes skeletal data, RGB data streams and participant demographics. We selected this dataset as it would not have been suitable to evaluate the proposed on gaming and/or action datasets which do not contain clinically relevant motions [16]. The following motions were extracted from the K3Da Dataset: *Chair Rise*, *One-leg Balance (Eyes Open)*, *One-leg Balance (Eyes Closed)* and *Tandem Balance*.

The results for each classifier are summarised in Table 3. We found that DCNNs performs consistently high when compared to other machine learning approaches. The average classification accuracy is 96.20% (with median 96.85%) when compared to ground truth labelling. Amongst the machine learning approaches, SVM has the closest result to DCNNs, with the accuracy of 91.10% (median 90.95%). The poorest performance was obtained by RUSBoost, with accuracy of 63.39% (median 62.35%). These results suggest that DCNNs is capable of identifying unstable motions with minimum error.

Table 3. Machine learning classification rate represented as the median and average for a random 10 iteration execution of the proposed methodology for correctly identifying "unstable" performance and "good" performance.

Iteration:	Median	Mean
SVM	90.95	91.10
RF	90.09	89.66
AdaBoost	82.87	84.26
LPBoost	71.03	72.22
RUSBoost	62.35	63.39
Total Boost	78.92	80.25
Bagging	72.61	73.83
SubSpace KNN	81.29	82.65
GRBM	82.50	81.27
DCNNs	**96.32**	**96.20**

We were able to detect a large number of motions which had been identified as unstable using DCNN (Table 3). Each motion will be discussed hereafter.

Chair Rise each participant started from a seated position. When prompted, they had to stand up so that the legs were fully extended, and then sit down again. This was repeated five times. The classification was highest amongst all motions, we speculate this is due to its unique characteristics in comparison to other motions reducing the inter-/intra class variations.

One-leg Balance (Eyes Open) participants stood with one foot five inches off the floor. They balanced with their eyes open and arms extended horizontally to be parallel with the floor. Classification was robust for this motion, however, inter-class confusion was present between *One-leg Balance (Eyes Closed)* due to large similarities between the motion type and styles of execution.

One-leg Balance (Eyes Closed) participants stood with one foot five inches off the floor. They balanced with their eyes closed and arms extended horizontally to be parallel with the floor. Again, classification was good for this motion. Unlike *One-leg Balance (Eyes Open)*, classification was high due to variation of the motions undertaken by participants.

Tandem Balance each participant placed one foot directly behind the other so that the big toe of the back foot was touching the back heel of the front foot. The arms were fully extended horizontally for a period of 10 seconds. A high sensitivity was achieved, however low scores for specificity and MCC point towards confusion amongst the identifying and correctly associating poor performance.

The results demonstrate that with the proposed framework it is possible to use DCNNs to identify motions which were unstable or executed stable. This offers a vital insight for clinicians to tailor rehabilitation programmes for each individual, focusing on motions which the individual finds difficult.

4 Conclusion

In this work, we propose a method for identifying unstable motions with features extracted from MoCap using Deep Convolutional Neural Networks. Experimental results demonstrate our proposed feature set combined with deep learning provides a high classification accuracy and could provide greater insights for a clinician in developing rehabilitation strategies or to aid in confidence boosting. The ability of DCNNs to recognise motions over other standard machine learning techniques is apparent for our experiments. Further, DCNNs can handle mis-formed skeletal poses provided by the Kinect and accounted for noise in the data. Future work will seek to extend DCNNs to handle a large number of classes, develop a wide range of dynamic features and improve insights to the clinician.

Acknowledgement

This work was supported by Royal Society International Exchanges Scheme (grant number: IE150436).

References

[1] B. F. Mentiplay, L. G. Perraton, K. J. Bower, Y. H. Pua, R. McGaw, S. Heywood, and R. A. Clark, "Gait assessment using the microsoft xbox one kinect: Concurrent validity and inter-day reliability of spatiotemporal and kinematic variables," *Biomechanics*, vol. 48, no. 10, pp. 2166–70, 2015.

[2] A. Amini Maghsoud Bigy, K. Banitsas, A. Badii, and J. Cosmas, "Recognition of postures and freezing of gait in parkinson's disease patients using microsoft kinect sensor," in *IEEE Conference on Neural Engineering*, 2015, pp. 731–734.

[3] Y. Yang, F. Pu, Y. Li, S. Li, Y. Fan, and D. Li, "Reliability and validity of kinect rgb-d sensor for assessing standing balance," *Sensors*, pp. 1633–1638, 2014.

[4] D. Leightley, M. H. Yap, J. Coulson, M. Piasecki, J. Cameron, Y. Barnouin, J. Tobias, and J. S. McPhee, "Postural stability during standing balance and sit-to-stand in master athlete runners compared with non-athletic old and young adults," *Journal of Aging and Physical Activity*, 2016.

[5] R. Clark, Y. H. Pua, K. Fortin, C. Ritchie, K. Webster, L. Denehy, and A. Bryant, "Validity of the microsoft kinect for assessment of postural control," *Gait and Posture*, vol. 36, no. 3, pp. 372 – 377, 2012.

[6] Y. Du, W. Wang, and L. Wang, "Hierarchial recurrent neural network for skeleton based action recognition," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 1110–1118.

[7] D. Leightley, J. S. McPhee, and M. H. Yap, "Automated analysis and quantification of human mobility using a depth sensor," *IEEE Journal of Biomedical and Health Informatics*, 2016.

[8] W. Zhu, C. Lan, J. Xing, W. Zeng, Y. Li, L. Shen, and X. Xie, "Co-occurrence feature learning for skeleton based action recognition using regularized deep LSTM networks," *CoRR*, vol. abs/1603.07772, 2016. [Online]. Available: http://arxiv.org/abs/1603.07772

[9] D. Leightley, M. H. Yap, B. M. Hewitt, and J. S. McPhee, "Sensing behaviour using the kinect: Identifying characteristic features of instability and poor performance during challenging balancing tasks," in *Measuring Behavior 2016*, May 2016.

[10] J. Darby, B. Li, R. Cunningham, and N. Costen, "Object localisation via action recognition," in *IEEE Conference on Pattern Recognition*, Tsukuba, Japan, 2012, pp. 817 –820.

[11] J. Wang, Z. Liu, Y. Wu, and J. Yuan, "Mining actionlet ensemble for action recognition with depth cameras," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2012, pp. 1290–1297.

[12] S. Li, W. Zhang, and A. B. Chan, "Maximum-margin structured learning with deep networks for 3d human pose estimation," in *The IEEE International Conference on Computer Vision (ICCV)*, December 2015.

[13] E. P. Ijjina and C. K. Mohan, "Human action recognition based on mocap information using convolution neural networks," in *Machine Learning and Applications (ICMLA), 2014 13th International Conference on*, Dec 2014, pp. 159–164.

[14] D. Leightley, M. H. Yap, Y. B. J. Coulson, and J. S. Mcphee, "Benchmarking human motion analysis using kinect one: an open source dataset," in *2015 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA)*, Dec 2015, pp. 1–7.

[15] J. Guralnik, E. Simonsick, L. Ferrucci, R. Glynn, L. Berkman, D. Blazer, P. Scherr, and R. Wallace, "A short physical performance battery assessing lower extremity function: Association with self-reported disability and prediction of mortality and nursing home admission," *Gerontology*, vol. 49, no. 2, pp. 85 – 93, March 1994.

[16] M. Firman, "RGBD datasets: Past, present and future," in *Computer Vision and Pattern Recognition WorkShop*, vol. abs/1604.00999, 2016.