Feature Descriptors for Gait Analysis from Depth Sensors

Ben Crabbe

August 10, 2015

1 Introduction

Gait analysis plays an important part in the treatment and assessment of a number of medical conditions. Presently gait analysis is usually performed through a combination of visual assessment by an experienced physiotherapist, automated methods such as marker based motion capture, pressure sensitive walkways or accelerometers. It requires patients to travel to a gait assessment laboratory which is far from ideal for patients who have difficulty walking.

This problem, and a range of other healthcare challenges, is being tackled through research and development by the SPHERE (a Sensor Platform for Healthcare in a Residential Environment) group in Bristol. An automatic, in home, gait analysis pipeline has been designed [**?**] which assesses the quality of a subjects movement using inexpensive RGB-D cameras such as the Microsoft Kinect.

Currently this system uses pose information captured stock skeleton tracking software, based on the algorithm of [**?**], to record the residents' body configuration. This skeleton tracking software infers the 3D coordinates of each of the body's relevant joints producing a $n_{joints} \times 3$ dimensional vector. This data is then processed using a non-linear dimensionality reduction, manifold learning method, Diffusion maps [**?**]. This method builds up a 3 dimensional manifold representation of the types of body configurations displayed in a dataset containing footage of the motion being measured. New skeleton data is then projected onto this manifold which, effectively parameterises the motion, removing the redundant information contained in the skeleton data, and enabling comparison of poses. [1] Finally, a statistical model of normal gait is built up from the training data using these pose vectors. New data is compared with this model producing a gait quality score on a frame-by-frame basis.

Since this system uses data driven, machine learning methods to learn both the manifold representation of pose and the model of normal motion, it can be applied to other types of movement quality assessment such a sports movement optimisation or physiotherapy exercise coaching. The system has been applied to a sitting-standing motion, to punching motions in boxing$^{howtositehttp://www.irc-sphere.ac.uk/work-package-2/movem}$ and to people walking upstairs.

One issue currently limiting the effectiveness of this system is the fragility of the skeleton tracking software. This software was designed for controlling entertainment/gaming systems with the user viewed frontally, within a range of 1-4m and at a pitch angle of $\sim 0^{\circ}$. Outside of these conditions skeletons become noisy and unreliable. Typically only a small fraction of data recorded from say a

[1] We will refer to the projected points in this space as the pose vector, and to the skeleton data as the body configuration vector.

camera attached to the ceiling above the stairs is at all usable. Increasing amount of usable data requires more intrusive camera placement which is to be avoided. The skeleton trackers also perform extremely poorly when props are involved in the scene, for example grasping a banister or a ball often leads to erroneous joint positions for that arm. It also struggles to accurately record sitting/standing motions.

The aim of this project is to develop a tailor made system for determining the reduced pose vector directly from RGB-D footage. To be effective this new component should enable the flexibility of the rest of the system by being able to record a wide range of motions. It should also work with an effective accuracy under the kinds of viewing angles produced by practical, unobtrusive, in home camera placements. This requires a data driven approach since the pose representation we wish to infer is not fixed, differing based on the body configurations in training data.

The methodology we find most suited to this task is a convolutional neural network (CNN). CNN's are a supervised learning method for extracting features, e.g. the pose vector, from images. Given training images labelled with the expected output the network learns to map from the input images to the output by adjusting the free parameters in the network. CNNs have been effectively applied to 2D human pose estimation from RGB images [**?**, **?**, **?**, **?**, **?**, **?**, **?**] where the positions of joints in the image plane were inferred. In [**?**] they were also applied to 3D joint position estimation from RGB, where they were shown to have reasonable accuracy from a range of viewing angles when trained with data captured by 4 cameras placed around the subjects. They have also been shown to be effective when applied to RGB-D object detection [**?**], pose estimation [**?**] and recognition [**?**].

For assessing the effectiveness of our solution we will focus on the staircase ascent motion, as this is the motion for which we possess the largest dataset. The accuracy of our predicted pose vectors are measured by computing the mean squared error (MSE) of the produced pose vectors against the labelled ground truth on a testing subset of the SPHERE staircase 2014 dataset [**?**]. We will also monitor the change in overall system performance (how well the measured gait quality score matches the score labelled by a trained physiotherapist).

To the best of my knowledge this project will be the first time that CNNs will be applied to a 3D human pose estimation task on RGB-D images (although this is not quite this since we find the pose representation rather than the full pose, but the network could be expected to perform this task reasonably well). It will also be a novel combination of CNNs and manifold learning methods since what we are essentially doing is simplifying the difficult task of human pose estimation through the dimensionality reduction stage. We believe that this will make the CNN easier to train and more effective overall since it has far less outputs to specify. If this is proved to be the case it could potentially be applied to other tasks that have been attempted with CNNs such as human action recognition.

The added value of this project, aside from the added value of the SPHERE work, is that this system has the potential to be applied to a wide variety of motions simply by retraining each stage. For an example, it could be used to coach basketball players to take free throws. First we would need to record the skeletons of players taking free throws then process them into a reduced pose representation for this motion. It would not be possible to use the Kinect in this scenario as it would fail to recover the skeleton with a ball in its hand. With our system it would be possible to record the motions using traditional marker based motion capture and process that to produce the pose vector. We would then use this data to train the CNN to extract this pose automatically. Then we would train the models

of normal pose and dynamics perhaps using whether the shot was successful or not as indication of whether it is a good or bad pose. This system could then be used on players to assess what specific part of their motions differs with the model of 'successful' motion. This example illustrates how our adaption increases the flexibility of the system over its current state.

2 Kinect Sensor

3 Preprocessing

All data used in this project SPHERE-staircase2014 dataset [**?**])