Feature Descriptors for Gait Analysis from Depth Sensors

Ben Crabbe

August 15, 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 joint position information captured by the OpenNI skeleton tracking software, based on the algorithm of [**?**]. 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 manifold learning method, Diffusion maps [**?**], to reduce the dimensionality of the data. This method builds up a 3D representation of the types of body configurations displayed in a training dataset containing footage of the motion being measured. New skeleton data is then projected onto this manifold. This effectively parameterises the motion, removing the redundant information contained in the skeleton data, and enabling simple 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 quality score for both pose and dynamics 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. Shotton et al's algorithm was designed for controlling entertainment/gaming systems with the user viewed frontally, within a range of 1-4m and at a pitch angle near 0°. Outside of these conditions skeletons become noisy and unreliable. Typically only a small fraction of data recorded

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

from say a camera attached to the ceiling above the stairs is fit for use with the system. Increasing amount of usable data requires more intrusive camera placement which is to be avoided. The skeleton trackers also perform 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 enable the flexibility of the rest of the system we require this new component to exibit the same flexibility as 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 presented 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 extracts progressively higher level features representations leading to the final pose vector. Following training the network is then able to generalise to unseen data, producing an output inferred from the examples it has seen.

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 benefit from depth depth data in the tasks of object detection [?], object pose estimation [?] and object 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. Refered to as the SPHERE staircase 2014 dataset [?], this includes 48 sequences of 12 individuals walking up stairs, captured by a Kinect v1 camera placed at the top of the stairs in a frontal and downward-looking position. It contains three types of abnormal gaits with lower-extremity musculoskeletal conditions, including freezing of gait and using a leading leg, left or right, in going up the stairs. All frames have been manually labelled as normal or abnormal by a qualified physiotherapist. There are 17 sequences of normal walking from 6 individuals and 31 sequences from the remaining 6 subjects with both normal and abnormal walking.

The accuracy of our predicted pose vectors are measured by computing the mean squared error (MSE) of the produced pose vectors against the label values. We also measure 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 our knowledge this project will be the first time that CNNs will be applied to a 3D human pose estimation task on RGB-D images. It will also be a novel combination of CNNs and manifold learning methods since what we are doing is simplifying the difficult task of human pose estimation through the dimensionality reduction stage. We find that this makes 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.

2 Related Work

2.1 Human Pose Estimation

Human pose estimation (HPE) is generally considered the task of measuring in 2D or 3D the joint positions of the human body. It is one of the most researched problems in computer vision due to the difficultly of the problem and the variety of applications such as video surveillance, humancomputer interaction, digital entertainment and sport science as well as medical applications.

This is a difficult task for a number of reasons. Firstly the human body has around 20 degrees of freedom [?], producing a huge space of possible body configurations, many of which will cause some joints to be occluded when viewed from a single camera. Additional difficulties arise from the variety in human appearance and clothing, and from left right ambiguities. Traditional motion capture methods methods rely on markers atached to the subject and multiple cameras to overcome these issues. Whilst such systems can provide highly accurate pose data, their use is restricted to controlled environments using expensive and calibrated recording equipment which renders them unsuitable in many applications.

Monocular visual pose estimation methods (reviewed in [?, ?, ?, ?, ?]) are generally divided into two approaches (e.g. by [?]); model based (or generative) and model-free (or discriminative) approaches. Model based approachs use prior knowledge of human shape and kinematics such as fixed limb lengths and defined joint angle limits to cast the image to pose transformation as a nonlinear optimisation problem or probabilistically in terms of a likelihood function, i.e. given this image data (and sometimes previous frames pose knowledge) what is the most likely valid pose. Model-free approaches instead learn a direct mapping from image data to pose, generally requiring learning/example based methods to achieve this.

With both approaches there are significant issues posed by the high dimensionality of pose data. In model based approaches likelihood functions, which are usually multi-modal and non-Gaussian, require a randomised search [?]. Such searches in 20 dimensions are computationally expensive and often lead to super real time frame rates [?]. In model free approaches training data must account for the highly non-linear mapping between image and pose, which means that the pose space must be densely sampled in the training set. Densely sampling a 20 dimensional space, even only the parts that correspond to valid human motion, whilst also modelling all the invariant aspects such as body shapes, viewing angle etc requires an inordinate amount of data [?, ?].

Although the full pose space is very large and high dimensional it has been shown e.g. in [?, ?] that if considering only the movements in a well defined activity e.g. walking then the pose data can be well represented by a low dimensional latent manifold. In a work closely related to our own Elgammal et al. [?] use a Local Linear Embedding method to generate a 1D manifold representation (embedded within a 3D space)$^{(FIGURE?)}$ of a walking motion from single sequences of silhouette images. They use a Generalised Radial Basis Function interpolation framework [?] (a form of neural network) to learn the nonlinear mappings from the manifold to silhouette image space and from the manifold to full 3D joint positions, which they then invert to extract points on the manifold from silhouettes and 3D pose from points on the manifold. In contrast, our work builds the manifold representation from 3D joint position data, this has the benefit that the same manifold representation applies naturally to multiple subjects, which is not the case in [?] (although they do introduce an solution for this problem in [?]).

It is also unlikely that this method could be used to capture abnormality in gait since defining an image to manifold transformation explicitly from the inverse constrains all input images to the poses contained in the original sequence.$^{(agree?)}$ Conversely Elgammal et al. argue that learning a smooth mapping from examples is an ill-posed problem unless the mapping is constrained since the mapping will be undefined in other parts of the space.$^{(response?)}$

Brand [?], also inferred 3D pose from silhouettes using an intermediate manifold representation. He uses a maximum a posteriori estimation for mapping between the image and manifold space. This uses information across the whole input sequence to find the most likely and consistent solutions in order resolve the ambiguities in the many to many silhouette to pose mapping. A solution of this form is unacceptable in our case as on of the key features of the SPHERE system is online measurement.

2.2 Human Pose Estimation From Depth

With the advent low cost commodity depth sensors challenging aspects of RGB HPE such as the variability in human appearance and scene lighting were greatly simplified. RGB-D also provides richer data for inferring 3D structure; human poses which could be appear identical when projected onto the 2D image plane can be distinguished. Full body HPE methods from single depth images are review in [?]. With the Kinect sensor and its bundled software packages (Kinect SDK) low cost, flexible and reasonably accurate HPE is now available and has been employed in a huge variety of scientific applications [?, ?].

The Kinect SDK and OpenNI skeleton trackers apply some unpublished inter-frame tracking algorithm to the single frame pose measurements of [?]. In this work Shotton et al. leveraged a large motion capture dataset which they re-targetted onto a variety of synthetic body models before rendering as if captured from a Kinect, simulating sensor noise, camera pose, and crop position. Using these generated depth images and a ground truth labelling of each pixel as one of 31 body parts they trained a randomised regression forrest to perform this body part classification at each pixel using simple and computationally efficient pixel wise features. They then use these classifications to infer actual joint position through a simple averaging and mean shift procedure. The whole algorithm operates in real time on the computational resources allowed to them on the Xbox gaming consoles GPU.

Further works

Abit about HPE from multi camera views / marker based. Abit about depth imaging advances. Ubiquitousness of the kinect system. description of shotton etal . mention other methods. difference with our needs and these methods

3 Preprocessing

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