Body joints regression using deep convolutional neural networks

Citation of the final article:

Published in its final form at https://doi.org/10.1109/SMC.2016.7844740.

This is the accepted manuscript.

© 2016, IEEE

Downloaded from DRO:
http://hdl.handle.net/10536/DRO/DU:30094569
Body Joints Regression Using Deep Convolutional Neural Networks

Ahmed Abobakr, Mohammed Hossny and Saeid Nahavandi

Institute for Intelligent Systems Research and Innovation (IISRI)
Deakin University, Australia

Abstract—Human pose estimation is a well-known computer vision problem that receives intensive research interest. The reason for such interest is the wide range of applications that the successful estimation of human pose offers. Articulated pose estimation includes real time acquisition, analysis, processing and understanding of high dimensional visual information. Ensemble learning methods operating on hand-engineered features have been commonly used for addressing this task. Deep learning exploits representation learning methods to learn multiple levels of representations from raw input data, alleviating the need to hand-crafted features. Deep convolutional neural networks are achieving the state-of-the-art in visual object recognition, localisation, detection. In this paper, the pose estimation task is formulated as an offset joint regression problem. The 3D joints positions are accurately detected from a single raw depth image using a deep convolutional neural networks model. The presented method relies on the utilisation of the state-of-the-art data generation pipeline to generate large, realistic, and highly varied synthetic set of training images. Analysis and experimental results demonstrate the generalisation performance and the real time successful application of the proposed method.

Index Terms—Kinect, RGB-D sensors, deep learning, ConvNet and pose estimation.

I. INTRODUCTION

Human posture estimation is a key essential step towards achieving better human behaviour understanding and machine interaction. It is a well-known computer vision problem that receives intensive research interest. The reason for such interest is the wide range of applications that the successful estimation of the human pose offers. These applications include but not limited to gaming, fall detection of lonely elders or generally health care, video surveillance, human computer interaction, and robots learning. Human pose estimation is the process of inferring the 2D or 3D positions of human body joints from images or videos. The generic pose estimation pipeline includes real time acquisition, analysis, processing and understanding of high dimensional visual information. Many factors control the accurate estimation of human postures. First, the efficient utilisation of the acquired high dimensional visual information via compact, discriminative, and hierarchical representation. Second, the ability of the learning model to obtain a generalisable mapping function based on these representations. In addition to that, large variations of human poses, occlusions, invisible body parts, noisy data sources, environmental effects such as illumination and cluttering further complicate this task.

Marker-less human pose estimation approaches can be classified into two main categories: body parts and regression methods [2]. The parts-based approaches follow the notion of using the pictorial structure model [3], [4], where, an object is given by a collection of parts with connections between certain pairs of parts. The current state-of-the-art level performance in terms of accuracy and real time applicability, for the human pose estimation problem was reported by Shotton et al. [5], and Buys et al. [6]. They incorporated depth imaging technologies and performed joints localisation via semantic body parts segmentation, also known as pixel-wise classification. In this method, a random decision forest (RDF) [7] model is trained on dense representations obtained using the hand-crafted depth comparison feature extractor (DCF) [5].

The main advantage of the parts-based methods is the ability to model strong pose articulations [8]. However, these methods have limitations. First, body parts based methods rely on local detectors, hence, they are prone to losing track of unobservable body parts due to self-occlusions or cluttered environments, resulting in unrealistic skeletons. Secondly, human poses space grows exponentially with the number of joints that could be modelled [5], therefore, it is infeasible to model all possible interactions between body parts [8]. To address these limitations, regression based methods which infer the human pose through a holistic view have become necessary [8]–[10].
In this paper, we follow the holistic posture estimation approach. The problem is formulated as an offset joints regression from an input raw depth image, and tackled using a deep learning architecture. Deep learning methods attempt to learn multiple levels of representations from raw input data, by composing simple but non-linear modules that each transforms the representation at one level into a representation at a higher, slightly more abstract level [11]. The key aspect of deep learning models is that these levels of representations are trainable, hence, alleviating the need for hand-crafted features [11]. They are the breakthrough for visual object recognition [1], [12], localisation and detection [13], [14]. In particular, the convolutional neural network (ConvNet) [15], is considered the most successful deep learning architecture. The ConvNet is considered an extension to the traditional multilayer perceptron, that is well suited to process data of multiple arrays of different modalities such as images and speech signals. Recently, Toshev et al. [8], incorporated a deep ConvNet regressor for the holistic pose estimation problem from RGB images, and achieved the state-of-the-art performance. However, human pose estimation from depth images is more robust than estimation from 2D images [2].

Traditional RGB approaches are plagued by the difficulty of separating subjects from backgrounds. Thus, they either suffer from poor performance; need uniform backgrounds; or need stationary cameras with static backgrounds [6]. Therefore, a growing trend is to use the depth cameras as they provide much richer geometrical information and facilitate preprocessing tasks such as background subtraction and objects delineation [16], [17].

In this work, we fine-tune the popular AlexNet [1] object recognition model, to work as an offset joint regressor for the articulated human pose estimation problem. The proposed method, as shown in Fig. 1, takes as input a raw depth frame and trained in a supervised mode to produce body joints coordinates. The advantages of this formulation are threefold. First, ConvNets are end-to-end trainable models that can automatically learn hierarchical feature representations. Second, deeper models can learn more abstract concepts from the raw input. Hence, the body joints regressor is able to capture the full context of the input. Third, to benefit from the AlexNets’ well trained low level feature extractors. The main difficulty with deep models is the need for large amounts of training data to achieve desired generalisation performance. Therefore, we extend the implementation of the state-of-the-art synthetic data generation pipeline [6], to enable generating virtually infinite amounts of high quality depth images each with the respective joint positions.

The rest of this paper is organised as follows: Section II presents the synthetic data generation and preprocessing pipeline. Section III, describes the proposed deep ConvNet training method. Experiments and results are discussed in Section IV. Finally, conclusion and future enhancements are highlighted in Section V.

II. DATA GENERATION

Training deep learning architectures is an optimisation problem with respect to millions of parameters. Thus, considering a supervised setting, it requires large amounts of labelled training data to achieve an acceptable generalisation performance and control the effect of overfitting. For computer vision problems, and human pose estimation in particular, collecting training data for people by capturing their movements in front of an RGBD sensor is an expensive process in terms of time and efforts. In addition, acquiring high quality ground truth labels such as joint positions is a challenge. The use of synthetic training datasets has been proven efficient for generalisation [5], [6]. To that end, we extend the state-of-the-art data generation pipeline proposed in [6], to synthesise training data. In this section, the building blocks of data generation and preprocessing pipeline, shown in Fig. 3, are described.

A. Motion Mapping

Typically, motion capture (MoCap) data are mapped onto virtual human models with different anthropometric measures creating virtually infinite amounts of training data. MoCap data is a variation of articulated human poses recorded using marker based motion capture systems for real humans. A wide collection of human behaviours is captured and made available for the research community by CMU Graphics
Fig. 3: Data generation and preparation pipeline. Motion frames of high degree of dissimilarity are mapped onto a 3D human model. The rendering process is parameterised with the camera view angle and distance, and produces depth and label image pairs. Depth maps are further processed by shifting in range (0, 255), rescaling to (227 x 227), and colourisation. Generated label maps are not sufficient for the joint localisation stage, therefore, we label the synthesised depth image using a trained RDF model from [6]. Both label maps are fused together and fed into the skeleton extraction module to obtain the ground truth pose vector for the current depth image.

Lab [18]. After excluding the irrelevant motion sequences such as acrobatics, and maintaining an euclidean distance threshold between consecutive poses, we end up with around 350K poses. During the mapping process, body depth pixels are grouped and assigned to the respective body parts identified by body joints. Body depth pixels are animated by subsequent transformations applied to the body joints.

B. Depth Rendering and Labelling

Animated human models are rendered using a parameterised virtual depth camera. For each model and camera parameters setting, such as viewing angle and distance, we can generate 350K fully labelled training images. Pixels of the ground truth images are labelled, each with the respective body part. Body joint positions are localised by averaging body part pixels. However, not all joints are defined, in particular, elbow and knee joints. Label images are based on the skeleton structure of the CMU dataset, which defines body parts in terms of bones. Buys et al. [6], have succeeded in adjusting the skeleton model of the CMU dataset and adding body part definitions for the elbow and knee joints and trained an RDF model on the new body representation. While their trained model has been released, their approach remains mysterious. To overcome this issue, we utilised their publicly shared RDF model for localising the elbows and knees joints. The evaluated label image is fused with the synthetic one and passed to the skeleton extraction module to produce the final ground truth joints vector. The fusion stage is further depicted in Fig. 4.

C. Depth Encoding and Preprocessing

Learning powerful representations from raw input data is a key aspect of deep learning models, with a great success for RGB modalities. Recently, many approaches [14], [19]–[21] investigated the ability of these models to effectively learn features from depth modalities. Couprie et al. [19], combined the depth data as an additional input channel to the ConvNet model of Farabet et al. [20], forming RGB-D input modality, and achieved good performance for the semantic segmentation task. Gupta et al. [14], proposed the HHA encoding of depth information. In their approach, each depth pixel is represented using three channels, horizontal disparity, height above ground and angle of pixel’s surface normal with the gravity direction. Empirical results demonstrated better performance using the HHA encoding compared to the raw depth data. However, their approach is computationally expensive [21], and may sound analogies to the traditional feature extraction paradigm, which incorporates much prior knowledge about the input. Eitel et al. [21], reported better performance compared to the results in [14], by colourising the depth data.

In this work, similar to [21], depth pixels are represented using three RGB colour channels. The values of the colour components vary according to the distance from the depth camera, and provide more powerful input signal to the ConvNet. Initially, input depth images are cropped and rescaled to 227 x 227 to suit the input dimensionality of AlexNet. Then, depth measurements are shifted, to provide depth invariance, and normalised to be in range (0, 255). Finally, a colour map is applied to produce the final RGB colourised depth image, as shown in Fig. 3 (top right).

III. CONVOLUTIONAL NEURAL NETWORKS

The ConvNet model is a composition of convolutional, pooling, and optionally fully connected layers. These layers are stacked and followed by a classification or regression
Fig. 4: **Label fusion strategy.** On the left, the generated label map based on the skeleton model accompanying the CMU motion sequences. As depicted, it does not model joints like elbows and knees. Therefore, using the trained RDF model of [6], we obtain the label map in the middle which has body parts representing the joints of interest. It also has a limitation of not representing the shoulders. Both label maps are merged to obtain the resulting pose vector.

The idea of fine-tuning also known as transfer learning depends on the degree of similarity between the data used for initialising the original model, and the data for the task at hand. Similar to AlexNet, our data is in the form of RGB modality. Therefore, we can make use of the powerful low level feature extractors that the AlexNet provides, for instance, the first two CONV layers. However, deeper layer of AlexNet are looking for generic concepts from the ImageNet [23] dataset which are not relevant to our task. We overcome this issue by using layer-wise learning rate policy to control the speed of learning for each layer. In particular, a base low learning rate of 0.0001 is used with the first three CONV layers, 0.0005 for the remaining layers, the regression module on top has the highest learning rate of 0.001. All learning rates are decreased with a factor of 10 every 20K iterations. We use the SGD optimisation procedure with a mini-batch size of 256, and a momentum of 0.9. The number of trainable parameters is approximately 57 million. We fine-tuned for 100K iterations which takes about three days on a NVIDIA Titan X GPU.
TABLE I
Percentage of Correct Parts (PCP) Vs Training Iterations

<table>
<thead>
<tr>
<th>Model</th>
<th>Arm</th>
<th>Leg</th>
<th>Avg</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Upper Lower</td>
<td>Upper Lower</td>
<td></td>
</tr>
<tr>
<td>10K</td>
<td>0.75 0.25</td>
<td>0.92 0.85</td>
<td>0.69</td>
</tr>
<tr>
<td>30K</td>
<td>0.76 0.28</td>
<td>0.93 0.85</td>
<td>0.71</td>
</tr>
<tr>
<td>60K</td>
<td>0.76 0.27</td>
<td>0.93 0.86</td>
<td>0.71</td>
</tr>
</tbody>
</table>

PCP accuracy computed on a synthetic test set of 100K images at PCP threshold of 0.5. The model reaches its maximum learning capacity after 30K iterations.

IV. EXPERIMENTAL RESULTS

We have fine-tuned the popular AlexNet deep ConvNet model for the task of human pose estimation. The input to this method is a colorised depth image of dimensionality 227 x 227 x 3 and the output is a real-valued pose vector of joints coordinates. The proposed model is trained on 300K images for 100K iterations and achieves a final validation loss of 0.03 on validation set of 100K images. In this section, we evaluate the generalisation performance of the proposed method. First, the test datasets and evaluation criteria are described. Second, quantitative results are reported on the synthetic test set. Third, as in the mean time, we do not have a high quality ground truth for the real test dataset, we show some inferences as qualitative results.

A. Test Datasets

We use both synthetic and real test data to assess the performance of the proposed method. The synthetic test dataset contains 100K covering three frontal view angles. The real test dataset contains 10K images captured using the Kinect camera, for a single subject. All test images are not included during training.

B. Evaluation Criteria

The two most widely accepted evaluation metrics for the human pose estimation problem are the percentage of correctly estimated body parts (PCP) [24], and the percentage of detected joints (PDJ) [8]. The PCP measures the detection rate of limbs. A limb is considered detected if the distance between its inferred endpoints and the ground truth endpoints is within a fraction of the limb length. This fraction is known as the PCP-threshold and is commonly set for a value between 0.1 and 0.5. The PCP metric has the drawback of penalising shorter limbs, such as lower arms, which are usually harder to detect [8].

PDJ overcomes this limitation by measuring the detection rate of body joints based on the same distance threshold. Typically, a joint is considered detected if the distance between the predicted and the ground truth joint position is within a certain ratio of the torso diameter. We define the torso diameter as the euclidean distance between the right shoulder and the left hip joints.

C. Quantitative Results

Table I, demonstrates the effect of the number of training iterations on the PCP scores. We report PCP results on the synthetic test set for the upper and lower arm and the upper and lower leg at PCP threshold of 0.5. The detection rate increases with the number of training iterations. However, the learning process saturates after 30K training iterations, approximately 25 epochs. Possible reasons for this saturation are either the model reached its maximum learning capacity, or there are no additional discriminative features to capture from the training data. For the rest of experiments, we will be using the 30K iterations model.

The presented results also demonstrate the drawback of the PCP, as it largely penalises the lower arms, which are very challenging to detect. This difficulty is shown in the overall results, as the lower arm limbs have the worse results with large margin. Figuring out a solution to improve the detection accuracy for hands specially, is a work in progress. On the other hand, the model shows strong capabilities to localise the core body limbs by achieving a PCP accuracy of 0.93 for the upper leg. Fig. 5, shows the detection rate of limbs using different PCP thresholds.

The PDJ metric is also evaluated on the synthetic test dataset. Fig. 6, reports the detection rates of the 14 modelled
body joints using a PDJ localisation threshold of 0.5. We also examine the detection rate of joints against range of threshold values, shown in Fig. 7

The presented results conclude that the proposed method has a high detection rates for core body joints such as head, neck, shoulders, hips and knees. However, we still have difficulties in detecting the rapidly changing, small body part joints such as hands and feet. This opens a room for further improvements on both the encoding method, in such a way that strengths the representation of these parts, and the low level feature extractors of the deep learning model.

D. Qualitative Results

Fig. 8 shows example inferences for the proposed method on the real test dataset. Note the high localisation accuracy for core body joints such as shoulders, hips, knees and head across different poses and depth in scene. Further, the anthropometric measures of the test subject are different from the virtual model, which was used during the data synthesis process. None of these real images is included during the training phase.

To obtain these results, first, we performed background modelling using the mode of a set of empty frames, no presence of subjects, as a calibration step. Second, the modelled background is subtracted from input depth frames. Third, post processing operations (erosion, hole filling and blob analysis) are applied to improve the quality of the foreground frames. Finally, jet colour map is applied on the resulting images.

Although the results seem promising, the model did not generalise well to unseen camera view angles. Hence, it requires modelling all the remaining angles during the training phase, which is feasible. Most importantly, as noted during our experiments, deep learning models are very sensitive to noise. This problem is commonly tackled by augmenting different noise models during the training phase. In this work, we alternatively applied the previously mentioned preprocessing operations to ensure noise free input to the model. The third approach is to extend the synthetic data generation pipeline, presented in section II, to simulate different noise models, which is technically possible. However, modelling all existing noise models is intractable. There are several noise models for different input devices, environments and materials. This limitation opens a room for investigating how the current deep learning models are looking through the input data.

V. CONCLUSION

In this paper, we followed the holistic pose estimation approach. A deep ConvNet offset joints regression model is trained to estimate the articulated human pose from an input raw depth image. We extended the state-of-the-art data generation pipeline to allow training on virtually infinite amounts of synthetic training data. We also presented a computationally efficient colorisation technique to provide more powerful input signal to the network, by distributing the depth information over the 3 RGB channels, incorporating minimal prior knowledge. We evaluated the model using the PCP and PDJ performance metrics. Results demonstrate the possibility of transferring knowledge from a generic deep learning model that was trained for visual classification tasks, to perform joints localisation, via training on synthetic data. In addition to that, the proposed method was able to generalise from synthetic training images to real images. Future work includes investigating different encoding techniques and architectural designs to further improve the detection rate of lower limbs. Further, simplifying the data generation pipeline via bypassing the RDF labelling stage to allow generating high quality ground truth joint positions is a work in progress.

ACKNOWLEDGEMENT

This research was fully supported by the Institute for Intelligent Systems Research and Innovation (IISRI) at Deakin University. The motion files used in this research were obtained from mocap.cs.cmu.edu. The database was created with funding from NSF EIA-0196217.
in terms of the low sensitivity to hands and foots positions. None of these real images is included during the training phase.

Fig. 8: Example inferences of the proposed method on real test images. Note the high localisation accuracy for core body joints such as head, neck, shoulders, hips and knees across different poses. Also, these results confirm with the quantitative scores in terms of the low sensitivity to hands and foots positions. None of these real images is included during the training phase.

REFERENCES


