

Master's thesis Deep Neural Networks for Dynamic Visual Data

Student Maria Priisalu

Supervisor Prof. Cristian Sminchisescu (LTH)

Examiner Prof. Magnus Oskarsson (LTH)

Deep Model for Human Pose Estimation from Video

POPULAR SCIENCE SUMMARY **Maria Priisalu**

Gaming platforms such as Kinect have mastered the positioning of the human skeleton by using specialized sensors. By removing the need for specialized hardware, pose estimation could be used on all devices with a video-camera. This thesis discusses a possible method for pose estimation from video.

Estimating the distance to objects from a single image can be a hard task. It is a task mastered by those with partially reduced vision in one eye. It is known that the human mind can learn to estimate the distance to objects from a single image. The question is if computers are capable of doing so also?

To answer the question we turned to a mathematical method called Artificial Neural Networks(ANN). As the name suggests the model resembles the human nervous system. The model is capable of learning the relationship between the input and the output. The model's learning capabilities rely on the model architecture and a number of parameters that are tuned during the engineering phase. We attempted to find a model architecture that enables the model to learn the 3D positions of human joints from video.

ANN's are made up of small units called neurons. The neurons are stacked on top of each-other creating layers. A network with a number of layers is called deep.

During recent years deep ANN's have been shown to have good learning capabilities in object recognition from images and activity recognition from videos. In the thesis we adjusted the Oxford Visual Geometry Group's 16 -layer network for image recognition in

a number of ways for pose-estimation. The model was tested and trained on the Human3.6M data-set. The data-set contains videos of 11 actors performing daily tasks, and the exact 3D poses of the actors throughout the videos. In total the dataset contains 3.6 million frames.

In the small-scale tests the number of layers in the model was decreased. It was noted that decreasing the number of layers led to lower accuracy. Three models were built estimating a different number of joints. The model estimating the position of all of the joints received similar accuracy to a model estimating the location of the joints in arms. The model predicting a single joint's position outperformed other models but since it takes more than 15 weeks to build the model it is not useable in practice.

Finally the model predicting the pose of the full skeleton was built on the large scale-dataset. The model received a test-error of around 30cm per joint. The best accuracy received on the given dataset is 13cm. Therefore the proposed model did not receive outstanding accuracy. It however showed that deep ANNs can be applied to pose estimation, and with further work they may outperform other methods.