A Deep Learning Based Method For 3D Human Pose Estimation From 2D Fisheye Images

Ching-Chun Chen* **Chia-Min Wu*** **I-Chao Shen*** **Bing-Yu Chen‡**
*†National Taiwan University
*{r04944003, a2301133, jdily}@cmlab.csie.ntu.edu.tw †robin@ntu.edu.tw

ABSTRACT
We propose a deep learning based method to directly estimate the human joint positions in 3D space from 2D fisheye images captured in an egocentric manner. The core of our method is a novel network architecture based on Inception-v3 [4], featuring the asymmtric convolutional filter size, the long short-term memory module, and the anthropomorphic weights on the training loss. We demonstrate our method outperform state-of-the-art method under different tasks. Our method can be helpful to develop useful deep learning network for human-machine interaction and VR/AR applications.

Author Keywords
Fisheye Image; 3D Human Pose Estimation; Egocentric View; Convolutional Neural Networks; Anthropomorphic Weights

ACM Classification Keywords
H.5.m. Information Interfaces and Presentation (e.g. HCI): Miscellaneous; I.2.6. ARTIFICIAL INTELLIGENCE: Learning

INTRODUCTION
The virtual reality HMDs (Head-Mounted Display) on the market all adopt tracking systems in an outside-in manner. Such tracking systems need the users to keep their bodies inside a tracking area defined in advance. These outside-in systems can only provide functional 6 DoFs tracking accuracy for the HMDs and the controllers in an indoor scene, also lack of high accuracy full-body pose tracking. To deal with these challenges, we design a tracking systems in an inside-out manner combined with fullbody tracking. Unlike outside-in tracking systems, inside-out tracking systems enable users go places beyond the limited space. Although leverage built-in depth and RGB cameras might be the most direct method to do this, neither of them have enough FoV to cover the fullbody motion. In this work, we instead using fisheye RGB camera with wider FoV (Figure 1), combined with a deep learning model to perform fullbody tracking. Unlike previous solutions provided by Microsoft Kinect [3], which use Random Decision Forest to track fullbody using RGBD camera, our method use only fisheye RGB camera, and achieve better tracking performance.

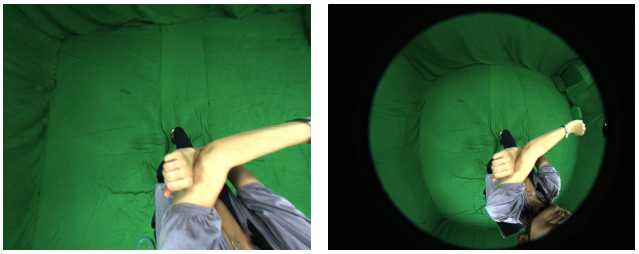

Figure 1. Compared with perspective cameras, fisheye cameras have much wider field-of-view. The picture on the right side is one of the fisheye images from the dataset provided by EgoCap [2]. The picture on the left side is the perspective version of the picture on the right side. We use the provided calibration information to remap the fisheye image into the perspective image.

METHOD
Our deep learning based method is designed to be able to estimate human 3D poses from only 2D fisheye images in an end-to-end fashion. Following, we will introduce the major contributions of our proposed model.

Asymmtric Filter Size
Our model is based on Inception-v3 [4] proposed by Google. Following the concept of original Inception-v3 model, we can transform an $n \times n$ filter into two parallel or sequential convolutional operations with $1 \times n$ filter size and $n \times 1$ filter size respectively. And we found out this asymmetric filter shape makes great performance gain compared to symmetric filters. The main reason is unlike normal images captured by standard perspective cameras, straight lines in the real world will become curves in fisheye images when they are captured by fisheye cameras.

Long Short-Term Memory
Long Short-Term Memory [1] is built with multi-gates structure and cell states focuses on solving the gradient vanishing problem and the gradient explosion problem between the hidden layers at different timestamps.

As we observed, the training and testing images of predicting human poses in 3D space are usually captured as video frames. We then combined our model with LSTM which enable us to obtain better performance with this data with strong temporal coherence.

IUI'18 Companion March 7–11, 2018, Tokyo, Japan

ACM ISBN 978-1-4503-5571-1/18/03…$15.00
DOI: https://doi.org/10.1145/3180308.3180344

ANTHROPOMORPHIC LOSS

When predicting joint positions of the human body in 3D space from 2D images, the torso joints and the head joints are usually easier to be predicted than other human body joints. The reason is that there is not much difference in 3D positions of the torso joints or the head joints between different human poses especially when the images are captured in an egocentric manner. On the other hand, the 3D positions of the joints of human limbs often vary a lot with different human poses.

To deal with this challenge, we propose an novel loss function that takes Anthropomorphism into consideration. Anthropomorphism is the attribution of human traits, emotions, or intentions to non-human entities. We realize this concept in our deep learning model as predefined weight parameters for limb joints of the human body (we divide the whole body joints into limb and non-limb set). For joints in the limb set, we multiply the original distance loss value with the predefined weight to amplify it's effect to the final aggregated loss value during training process.

EXPERIMENTAL RESULT

Experimental Environment

Here, we report our results on two different sequences: 750 2D fisheye image sets of gesturing recognition and 250 2D fisheye image sets of walking action [2]. In the image set, 60 image sets in the *"Gesture"* sequence and 30 image sets in the *"Walk"* sequence are reserved to be the testing datasets. The remaining image sets as the training datasets will be used to train our model. In the following experiments, we define the mean value of average error distance per body joint (Euclidean distance in millimeter between the ground truth and the prediction) across all the testing samples as the predicting accuracy in one epoch. In the end, we compare the performance of different training settings with each other according to the highest prediction accuracy.

Evaluation

Our work can reconstruct the 3D skeletal joints of the human body from the *"Gesture"* sequence with the error distance value at 13.10 mm and from the *"Walk"* sequence with the error distance value at 19.58 mm. For comparison, we use the same training sets and test sets in EgoCap [2], where the error distance values which are both at 70 mm and evaluated on the *"Gesture"* sequence and the *"Walk"* sequence (Figure 2 and Figure 3). Our work has the upper body joint accuracy and the lower body joint accuracy with the similar error distance values at 12.8 mm (see Figure 2) on the *"Gesture"* sequence. The upper body joint accuracy is higher than the lower body joint accuracy with the error distance values at 14.70 mm and 24.05 mm (see Figure 3) on the *"Walk"* sequence. In summary, our method is able to greatly outperform the state-of-the-art EgoCap [2].

CONCLUSION AND FUTURE WORK

In this work, we proposed a deep learning model that effectively predict joint positions of the human body in 3D space from the 2D fisheye images captured in an egocentric manner. Our model is able to outperform the state-of-the-art method

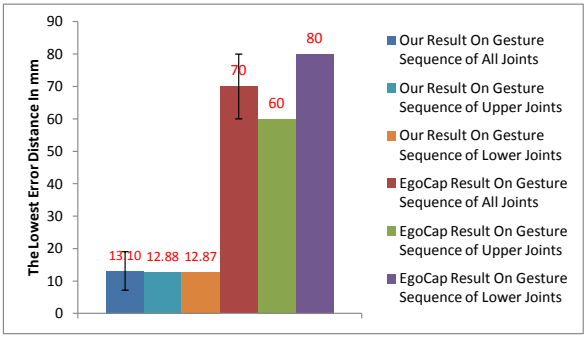

Figure 2. The comparison between our prediction error distance and the prediction error distance of EgoCap on the *"Gesture"* sequence.

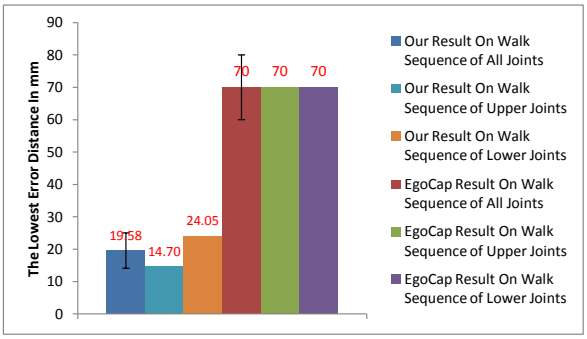

Figure 3. The comparison between our prediction error distance and the prediction error distance of EgoCap on the *"Walk"* sequence.

by a great margin. In the future, we consider to capture and collect a bigger dataset for fisheye fullbody tracking for benefiting other relevant research. Meanwhile, We are going to adopt this our method in designing different interactive AR/VR applictions.

ACKNOWLEDGMENTS

This research was supported in part by the Ministry of Science and Technology of Taiwan under MOST107-2634-F-002-007 and MOST106-2221-E-002-211-MY2.

REFERENCES

1. Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long Short-Term Memory. *Neural Comput.* 9, 8 (Nov. 1997), 1735–1780.
2. H. Rhodin, C. Richardt, D. Casas, E. Insafutdinov, M. Shafiei, H.-P. Seidel, B. Schiele, and C. Theobalt. 2016. EgoCap: Egocentric Marker-less Motion Capture with Two Fisheye Cameras. *ACM Trans. Graph.* 35, 6, Article 162 (Nov. 2016), 11 pages.
3. J. Shotton, A. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman, and A. Blake. 2011. Real-time Human Pose Recognition in Parts from Single Depth Images. In *IEEE CVPR*. 1297–1304.
4. C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. 2016. Rethinking the Inception Architecture for Computer Vision. In *IEEE CVPR*. 2818–2826.