MIPOSE: A Micro-Intelligent Platform for Dynamic Human Pose Recognition

Zhishuai Han*
University of Science & Technology, Beijing
Beijing, China
zsani.han@gmail.com

Xiaokun Wang*
University of Science & Technology, Beijing
Beijing, China
wangxiaokun@ustb.edu.cn

Xiaojuan Ban†
University of Science & Technology, Beijing
Beijing, China
banxj@ustb.edu.cn

Jianyu Wu
University of Science & Technology, Beijing
Beijing, China
ustbjerome@163.com

*Both authors contributed equally to this research.

†Corresponding author.

ABSTRACT

Giving computers the ability to learn from demonstrations is important for users to perform complex tasks. In this paper, we present an intelligent self-learning interface for dynamic human pose recognition. We capture 20 samples for an unknown pose to train a stable generative adversarial networks (GAN) system which aims to conduct data enhancement, then we adopt a threshold isolation method to distinguish relatively similar poses. A few minutes of learning time is sufficient to train a GAN system to successfully generate qualified pose samples. Our platform provides a feasible scheme for micro-intelligent interface, which can benefit to human-robot interaction greatly.

Asian HCI Symposium'19, May 4–9, 2019, Glasgow, Scotland UK

ACM ISBN 978-1-4503-6679-3/19/05.
https://doi.org/10.1145/3309700.3338440

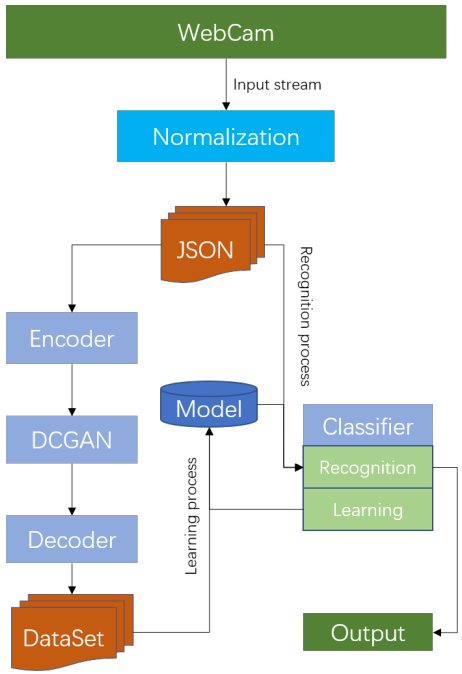

Figure 1: End-to-end architecture of MIPOSE. Training and recognition process don't require manual intervention.

CCS CONCEPTS

• **Human-centered computing** → **Gestural input**; **User interface management systems**; *HCI theory, concepts and models*; *Activity centered design.*

KEYWORDS

Pose Recognition, Micro-Intelligent interface, DCGAN, Human-Robot Interaction

ACM Reference Format:
Zhishuai Han, Xiaokun Wang, Xiaojuan Ban, and Jianyu Wu. 2019. MIPOSE: A Micro-Intelligent Platform for Dynamic Human Pose Recognition. In *Asian HCI Symposium'19 (Asian HCI Symposium'19), May 4–9, 2019, Glasgow, Scotland Uk*. ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3309700.3338440

INTRODUCTION

A major interior of most existing human-computer interfaces is that they can only perform predefined tasks. Traditionally they obtain some priori knowledge to "remember" established rules and regulations, however, this kind of strategy lacks self-learning ability, and they can't expand to new recognition patterns quickly. Artificial intelligence especially deep learning (DL) has shown fantastic generalization performance on pose estimation and gesture classification [5], but DL based methods are highly dependent on sufficient sample data. Collecting enough samples and manually labelling will consume plenty of time, and technically speaking, training is still modelling some patterns and telling machines to "remember" them, but not to "learn". We think an intelligent machine should be capable of learning from real-time demonstrations or several samples, and can distinguish new patterns and off-the-shelf ones [4]. Hence we design a micro-intelligent platform for dynamic human pose recognition, it only needs 20 samples as basic knowledge to learn target human pose. Our designed platform is based on end-to-end architecture, and it can finish recognizing an unknown human pose within 10 minutes. The learning and recognition process without human intervention will greatly improve the human-computer interaction experience.

In this extended abstract, we first present the technical details that power our critical micro-intelligent human pose recognition platform. Next, we show how to create a new dynamic human pose and make the platform recognize it. Besides, we provides effect contrast between using our proposed agent and just raw sample data. Finally, we conclude with a brief discussion of future work.

RELATED WORK

GAN has developed a lot since it was proposed, such as CVAE-GAN [1] and StarGAN [3]. Among all the derivative methods of GAN, deep convolutional generative adversarial networks (DCGAN) is one of the most widely used and stable methods [6]. It adopts two CNN modules as generative network

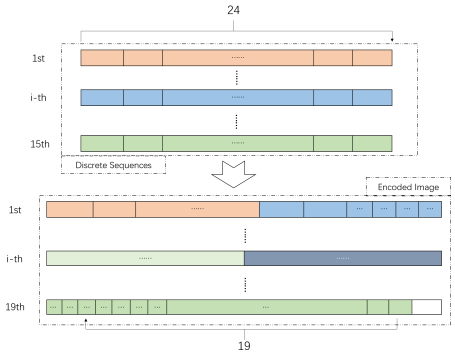

Figure 2: Regulation of encoding discrete sequences into images.

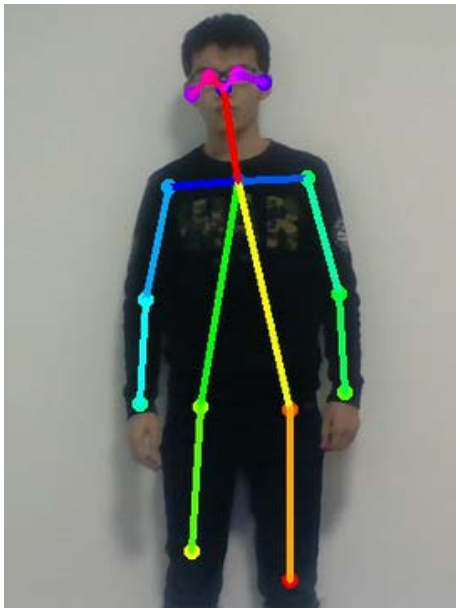

Figure 3: Body keypoints display using OpenPose. The center keypoint in the head represents the head keypoint.

and adversarial network. It uses a set of random noise as input to generate data approximating the original samples, and stops training when the network can't tell whether the generated data is real or not. In this document, we adopt DCGAN as intelligent agent to do data enhancement.

TECHNICAL DETAILS

The offline real-time dynamic human pose recognition platform is built using OpenPose [2], the most popular open-source human pose recognition system based on state-of-art methods. Figure 1 demonstrates the architecture of our proposed platform. We use OpenPose to detect human keypoints. When a new pose starts, the platform saves real-time keypoints to a JSON file. Normalization module processes the JSON data to make them conform to a specific length. Encoder module is used to encode discrete pose sequences into images. DCGAN module uses encoded images to generate more synthetic samples. Decoder module decodes generated samples back to sequences of original format. Classifier module uses a sliding window to match webcam input stream with stored samples and output recognition results. Our platform runs on GTX 1050 only using webcam, written in Python for the development.

MODULES

OpenPose: OpenPose can provide human body keypoints from still images and videos. In order to accelerate recognition speed of webcam input, we adopt a slight model which can run 10±2 FPS on our laptop.

Normalization: Since that it's impossible for all samples of a pose having the same length, we design the normalization module to modify random-length sequences to 15 frames. It means that each human pose contains 15 frames of body keypoints and should be performed within about 2 seconds. If an acquired pose sequence has more than 15 frames, we use every 15 frames in this sequence as one sample. Otherwise, this acquired sequence is used as a sample.

Encoder: DCGAN can only process images, so we encode discrete body keypoints sequence into images. The encoding regulation is shown as Figure 2. The platform uses 12 keypoints to configure and recognize target pose(see Figure 3, "head" is one keypoint), and each keypoint has two values, each pose contains 15 frames. Therefore one pose sequence can be encoded into a 19×19 image.

DCGAN: DCGAN is the most time-consuming process in the platform. We have experimented that the network can reach a stable state after 300 echos of training. We generate 1k samples for a new pose. On our laptop (Ubuntu, 18.04 LTS), it will take around 5 minutes to collect such number of samples.

Decoder: Decoder is exactly reversible process of encoder. Its main function is to transform generated images back to keypoint sequences.

	Test Conditions						
Pose	*DCGAN*	*Raw*	*Raw&Noise*	*Pose*	*DCGAN*	*Raw*	*Raw&Noise*
Swipe Up	99.5%	95.25%	97.0%	Zoom in	100.0%	98.5%	99.0%
Swipe Down	98.0%	94.50%	96.75%	Zoom out	99.5%	96.0%	99.0%
Swipe Left	99.0%	96.75%	98.50%	Rotate Left	99.25%	92.5%	93.75%
Swipe Right	99.75%	90.0%	97.50%	Rotate Right	99.25%	97.0%	98.0%

Table 1: Recognition results comparison of DCGAN agent, raw samples, and raw samples with noise. We can see that our agent has better robustness that the others.

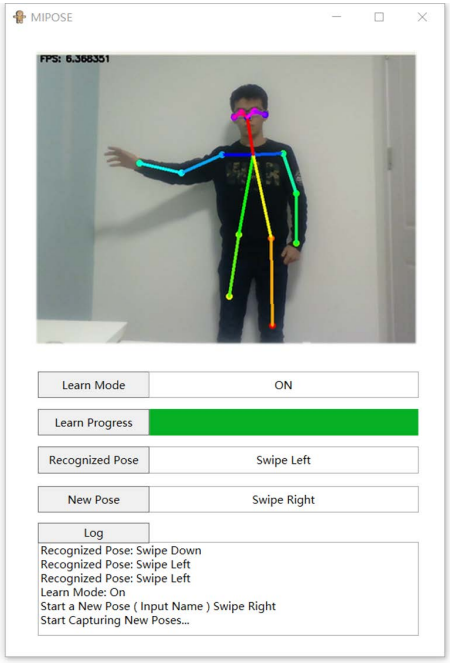

Figure 4: An Instant Process of MIPOSE Platform Running.

Classifier: Classifier has two major functions. One is in the process of recognition. We store 20 frames of data before the current frame, then set a sliding window traversing from -19(the farthest frame) to -14, each step output a undetermined pose, totally 6 poses. Finally we choose the pose with the highest frequency as the target pose. The other function is in the process of learning. When the platform is going to record a new pose, and this pose is pretty similar with an existing pose, our platform gives up recording new pose and tells the users that a similar pose is already in use. Classifier adopts a predefined threshold to perform similarity measurement.

EXPERIMENT

Setup

Figure 4 shows the main interface of our platform. We use left hand as the switch to start and turn on/off learning mode. Left hand going up/down means "mode on/off", while going right means "start learning". When the platform learns a new pose, there should be 3 seconds of stationary time before the start and after the end of the pose, to ensure that the normalization module can segment target keypoint sequences easily.

The main contribution of our MIPOSE platform is DCGAN based intelligent agent. We experiment three strategies to validate the superiority and practicability of our proposed architecture, which are (1) DCGAN agent, (2) 20 raw samples, (3) 20 raw samples with another 20 samples with random noise. We collect 400 recognition results each for 8 poses from 8 people, shown in Table 1.

Result

The comparison result shows in Table 1. We can see that our micro-intelligent agent has better performance that the others.

CONCLUSION

We introduce a micro-intelligent platform named MIPOSE for dynamic human pose recognition, an end-to-end interface which is not to "remember" predefined patterns, but has the ability to "learn" pose patterns. Our platform can greatly reduce the time for designing rules for recognition. We believe the intelligent agent and designed architecture could benefit HCI practitioners who are interested in creating more human-like robots.

FUTURE WORK AND IMPACT

Our platform realizes end-to-end pose recognition, and it could learn a new pose within a few of minutes. The next step will be optimizing DCGAN module to reduce the training time, to further improve the efficiency of learning.

Our research combines GAN, one of the most potential methods in deep learning, with human pose recognition. GAN is one of the hottest research topics in deep learning and it will make continuous improvement in the following years. Our designed architecture integrated with GAN can be more intelligent, also it can be extended to other areas of HCI like intelligent robot. The future direction of HCI will be automation and intellectualization, our main contribution is to design a core intelligent agent, bringing reinforcement learning to traditional HCI. This method provides a new feasible scheme for researchers those will study intelligent HCI and create more human-like robots in the future.

ACKNOWLEDGEMENTS

This work was supported by The National Key Research and Development Program of China (Grant No. 2016YFB1001404) and National Natural Science Foundation of China (61873299, 61702036, 61572075). We would like to thank Ryosuke Takada, Kaori Ikematsu and Sotetsu Koyamada and for their valuable advice on manuscript revision. We would also like to thank Börje Karlsson from Microsoft Research Asia for his guidance in designing the platform.

REFERENCES

[1] Jianmin Bao, Dong Chen, Fang Wen, Houqiang Li, and Gang Hua. 2017. CVAE-GAN: Fine-Grained Image Generation Through Asymmetric Training. In *The IEEE International Conference on Computer Vision (ICCV)*.

[2] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. 2017. Realtime Multi-Person 2D Pose Estimation Using Part Affinity Fields. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[3] Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. 2018. StarGAN: Unified Generative Adversarial Networks for Multi-Domain Image-to-Image Translation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[4] Nishanth Koganti, Abdul Rahman H. A. G., Yusuke Iwasawa, Kotaro Nakayama, and Yutaka Matsuo. 2018. Virtual Reality As a User-friendly Interface for Learning from Demonstrations. In *Extended Abstracts of the 2018 CHI Conference on Human*

Factors in Computing Systems (CHI EA '18). ACM, New York, NY, USA, Article D310, 4 pages. https://doi.org/10.1145/3170427.3186500

[5] Behnam Neyshabur, Srinadh Bhojanapalli, David Mcallester, and Nati Srebro. 2017. Exploring Generalization in Deep Learning. In *Advances in Neural Information Processing Systems 30*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.). Curran Associates, Inc., 5947–5956.

[6] Alec Radford, Luke Metz, and Soumith Chintala. 2015. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434* (2015).

AUTHOR BIOS

Zhishuai Han

Master student at University of Science & Techonology, Beijing. His research topics include human-computer interaction and deep learning. He gained plenty of engineering and research experience in Microsoft Research Asia, and published several international papers included in EI.

Xiaokun Wang

Postdoctoral researcher and lecturer at University of Science & Technology, Beijing. His research interests include computer graphics, virtual reality and human-computer interaction.

Xiaojuan Ban

Professor at University of Science & Technology, Beijing. She has published over 200 international conference papers and journals, including "Expert Systems with Applications", "Journal of Visualization", "Mathematical Problems in Engineering", and so on. More than 100 papers are included in EI and SCI.

Jianyu Wu

Bachelor student of University of Science & Technology, Beijing. His research topics include human-computer interaction and computer vision, he participated in the experiments of several international papers indexed in EI.