Real-time human motion forecasting using a RGB camera

Conference Paper · November 2018

DOI: 10.1145/3281505.3281598

CITATIONS
2

READS
474

2 authors:

Erwin Wu
Tokyo Institute of Technology
10 PUBLICATIONS **6** CITATIONS
SEE PROFILE

Hideki Koike
Tokyo Institute of Technology
193 PUBLICATIONS **2,395** CITATIONS
SEE PROFILE

Real-time Human Motion Forecasting using a RGB Camera

Erwin Wu
School of Computing, Tokyo Institute of Technology
wu.e.aa@m.titech.ac.jp

Hideki Koike
School of Computing, Tokyo Institute of Technology
koike@acm.org

ABSTRACT

We propose a real-time human motion forecasting system which visualize the future pose in virtual reality using a RGB camera. Our system consists of three parts: 2D pose estimation from RGB frames using a residual neural network, 2D pose forecasting using a recurrent neural network, and 3D recovery from the predicted 2D pose using a residual linear network. To improve the prediction learning quantity of temporal feature, we propose a special method using lattice optical flow for the joints movement estimation. After fitting the skeleton, a predicted 3d model of target human will be built 0.5s in advance in a 30-fps video.

CCS CONCEPTS

• **Computing methodologies** → **Motion capture**; • **Human-centered computing** → *Mixed / augmented reality*; Information visualization;

KEYWORDS

Motion forecasting, Real-time pose prediction, Deep neural network

ACM Reference Format:
Erwin Wu and Hideki Koike. 2018. Real-time Human Motion Forecasting using a RGB Camera. In *VRST 2018: 24th ACM Symposium on Virtual Reality Software and Technology (VRST '18), November 28-December 1, 2018, Tokyo, Japan.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3281505.3281598

1 INTRODUCTION

Motion capture system plays an important role in Human-Computer Interactions, especially in sports estimations. However, most recent precise 3D motion capture systems require players to wear specific suits or sensors, or use special depth cameras, which might not be applicable to all types of sports.

Therefore, we propose a real-time 3D human pose forecasting system using deep neural network to predict human's future pose from normal RGB frames and visualize the prediction result on VR head-mounted display. Our system uses a customized residual network to obtain 2D human joints from an input image and uses recurrent networks to learn the temporal features of the human motion. To get a precise motion prediction, we develop a novel lattice optical flow algorithm to calculate the joint movement with less computation. Finally, we import the recovery network for the 3D construction developed by Martinez et al. [3] to reconstruct the predicted 2D joint positions to a 3D skeleton. Because these operations are done in parallel with low computational cost, a future 3D pose of our subject can be generated in real time and visualized in virtual reality.

VRST '18, November 28-December 1, 2018, Tokyo, Japan

ACM ISBN 978-1-4503-6086-9/18/11.
https://doi.org/10.1145/3281505.3281598

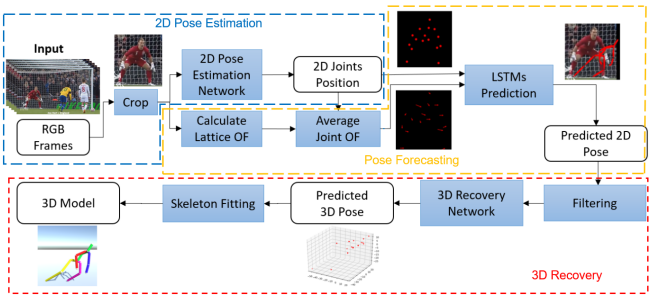

Figure 1: System overview

2 RELATED WORKS

Chao et al. [1] proposed the 3D Pose Forecasting Network (3D-PFNet) as the first study on forecasting human dynamics from single RGB images. Their method of forecasting 2D skeletal poses and converting them into 3D space was shown to have quantitative and qualitative results, with average joint position errors is of approximately 87.6mm. Nevertheless, 3D-PFNet is a offline networks requiring a large amount of computation, and is therefore difficult to use in the contenxt of sports which require high-speed feedback. Horiuchi et al. [2] forecast human body motions 0.5s (15 frames in 30 fps video) in advance using a five-layered neural network with motion data input taken by a Microsoft Kinect V2 camara, the maximum difference in the prediction was 7.9cm which was acceptable for their experiment. However, Kinect is a depth camera using IR sensors, as previously mentioned: therefore, it is not suitable to use in an outdoor environment or a large area, which means it is not applicable to most of the sports.

3 IMPLEMENTATION

As shown in Figure 1, our system can be considered as having three main components: the detection of 2D joints , the forecasting of the joints positions, and the 3D recovery for visualization in VR.

3.1 2D pose estimation

We refer to the VNect pose estimation model of Dushyant et al. [4], which uses a customized ResNet50 to allow the convolutional layer to regress the 2D joint data and is trained on an annotated 3D human pose dataset. Because we do not need the 3D joints for the prediction, we adjusted the network to directly predict the 2D joints instead of generating location maps for the depth estimation. The former part of our network is exactly the same as the ResNet50 stride 2 till level 4, with 224×224-size cropped RGB image as input, after a pooling layer, the result of the convolution is fully connected to the linear regression to obtain the 2D joint positions. In our approach, we only defined 17 key joints of the human body not

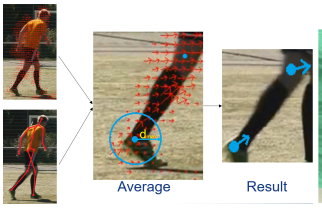

Figure 2: Keypoint lattice optical flow method

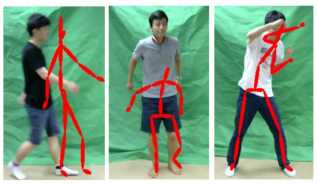

Figure 3: Prediction result

including the foot tip and hand tip; therefore, the output dimension of the pose estimation is 17×2. The output joint positions from five continuous frames (the present frame and the former four frames) are then passed to the LSTM layer for prediction together with the calculated optical flow data mentioned in the next paragraph.

3.2 Pose Forecasting

To obtain full temporal features of human movement, we used the optical flow and joint positions data to do a regression on the LSTMs. This type of two-stream networks are widely used and obtain high accuracy in understanding human actions. However, computing the dense optical flows and pass them into an LSTM network would cause a large inference time even if the image's resolution were low, in that case, it is difficult to estimate poses in real time. Therefore, we develop a new type of sparse optical flow called keypoint lattice-optical flow, which creates several lattice points around a keypoint (in this case, the human joint) and only calculates the optical flows of these lattice points. We decide the distance d of each neighbor point to have a sum of $W/d \times H/d$ (W and H are the width and height of cropped image, respectively) lattice points, then we calculate the Lucas Kanade optical flow vector of each point, and define the vector of the corresponding point (x, y) to be $LK(x, y)$. As a result, the computation was reduced by at least d^2 times.

As shown in Figure 2, we can obtain the average optical flow Avg_j representing the movement of specific joints by averaging all the optical flow vector near a joint. While the radius r_d of the averaging circle need to be tuned for different applications to obtain higher accuracies; however, smaller d and larger r_d will lead to heavier computations. In most of our implementations, with cropped images with a size of 224×224 as input, we used $d = 8$ and $r_d = 24$, which means that, at most, 28 lattice optical flow vectors are averaged for one joint. In our demonstration, we test the effect of predicting 0.5s future of 3 kinds of motion: walking, squat and boxing. By connecting all the predicted 2D joints positions, we can get the 2D pose forecasting results shown in Figure 3.

3.3 3D Pose Recovery

We fine-tuned the model of Martinez et al. [3], who developed an effective network for 3D pose recovery, by setting the input and output dimensions to 17 joints and using our dataset to re-train the model. After the forecasted 2D joint positions were output from the LSTM network, we used a simple threshold filter, which filters the joints that are far from the center of body more than 70% of height. The filtered data are then passed to the recovery network for the 3D construction, with only two linear layers and two residual blocks, which means that there are only six linear layers in total.

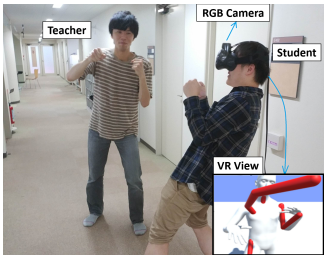

Figure 4: Virtual boxing prediction

4 APPLICATION

Safe Martial Training: martial arts such as boxing, karate, or taekwondo are often accompanied by injuries. Professional practitioners may predict an opponent's attack from their current pose or movements, based on years of experience. By visualizing the prediction motion on VR head mounted display, we can create a virtual representation model of two users and allow them to fight in close quarters in a virtual environment. With the help of the forecasting, an amateur can also "predict" a professional player's attack and, because there is no real contact between the two users, fewer injuries will occur than in normal training.

Real-time Remote Coach: since the motions are captured by a RGB camera, it is not necessary to have the opponent person stand in front of the user. So it is possible to do remote coach of sports or athletes using a simple video chat, while the user can observe the 3D motion clearly through the VR HMD. The 0.5s prediction can fix the problem of network latency and make the training more like real-time training.

5 DISCUSSION & CONCLUSIONS

In this paper, we presented a real-time 3D human pose forecasting system with a novel optical flow method for two stream deep neural network. To improve the learning quantity of the motion's temporal features, we also designed a special lattice optical flow for the joint optical flow vector estimation. Although our method does show some effects on improving the computation speed, we have not done any evaluation experiment about the inference time by yet, which will be conduct in our future work as well as some accuracy tests of the forecasted poses. Regardless of the above-mentioned problems, our system is capable of 3D pose estimation and forecasting in virtual reality, which could be applied to real-time visualizations, sports estimations, or trainings.

ACKNOWLEDGMENTS

This work was supported by JST CREST under Grant No.: JPMJCR17A3.

REFERENCES

[1] Yu-Wei Chao, Jimei Yang, Brian Price, Scott Cohen, and Jia Deng. 2017. Forecasting human dynamics from static images. In *IEEE CVPR*.
[2] Yuuki Horiuchi, Yasutoshi Makino, and Hiroyuki Shinoda. 2017. Computational Foresight: Forecasting Human Body Motion in Real-time for Reducing Delays in Interactive System. In *Proceedings of the 2017 ACM International Conference on Interactive Surfaces and Spaces*. ACM, 312–317.
[3] Julieta Martinez, Rayat Hossain, Javier Romero, and James J. Little. 2017. A simple yet effective baseline for 3d human pose estimation. In *ICCV*.
[4] Dushyant Mehta, Srinath Sridhar, Oleksandr Sotnychenko, Helge Rhodin, Mohammad Shafiei, Hans-Peter Seidel, Weipeng Xu, Dan Casas, and Christian Theobalt. 2017. VNect: Real-time 3D Human Pose Estimation with a Single RGB Camera. *ACM Transactions on Graphics* 36, 4, 14.