ISASE 2020

Time Series Human Motion Prediction Using RGB Camera and OpenPose

Andi Prademon YUNUS*, Nobu C. SHIRAI*, Kento MORITA*, and Tetsushi WAKABAYASHI*

** Mie University, 1577 Kurima-machiya, Tsu, 514 Japan*
andi@hi.info.mie-u.ac.jp

Abstract: The report projects that by 2050, the population aged 60 and above will reach 2.1 billion. This aging society is more likely to suffer from locomotive syndrome. In order to reduce the spread of locomotive syndrome, it is best to increase awareness before the citizens become elderly. We propose the system to predict human motion as the first step to realize the locomotive syndrome estimation. Previous researches were using the Kinect camera which has a depth sensor that the camera used to detect the pose of a human body. However, in this research we are using an RGB camera as a reliable alternative. We set a goal to predict 1 second ahead of the motion which includes simple motions such as hand gesture and walking movement. We used OpenPose to extract the features of a human body pose including 14 points. YOLOv3 is used to crop the main feature in the frames before OpenPose process the frame. Distance and direction which are calculated from the features by comparing two consecutive frames as the input of Recurrent Neural Network Long Short-term Memory (RNN-LSTM) model and Kalman Filter. Mostly, Kalman Filter show better accuracy then RNN-LSTM and based on the human motions, motion such as hand gesture and moving to the right side are easier than more complex motion like hand gesture and moving to the left side. We confirmed the validity of RGB-camera based method in the simple human motion case from the result.

Keywords: *Human motion prediction, RNN-LSTM, Kalman Filter, OpenPose, YOLOv3, Deep learning*

1. INTRODUCTION

The aging society becomes a large issue in many countries, governments and hospitals considering the application of counter-measurement. The human body motion evaluation attracts attention in the medical field because human motion in some activities (walking, running, going upstairs, etc.) are correlated with their health. For example, Sanwa Newtec Co., Ltd. has developed a system to numerically evaluate the locomotive syndrome using Kinect. As well as Hiroki Tamura, et. al.[1], with their research on locomotive syndrome estimation based on walking motion using Kinect. The cost and hardware size will become an obstacle when introducing the system to a small clinic or a patient's house. In general, the level of the locomotive syndrome is evaluated by doctors in the hospital. Therefore, a system to evaluate the human motion using low cost and widely available device is required.

Some researches develop their systems with data from the RGB-D camera since it has depth parameter for human pose estimation [2, 3]. RGB-D camera such as Kinect camera can precisely estimate precisely human body parts. However, in this research we start with using the RGB camera as the other option that we can rely on. A research has been performed for human motion prediction with RGB camera [4]. They focused on human motion forecasting of sports activity especially for safe martial arts such as boxing, karate or taekwondo. As a result, they obtained 0.5 second of human motion prediction by forecasting 15 frame steps in a 30fps video. Nonetheless, this paper does not show the accuracy of the prediction.

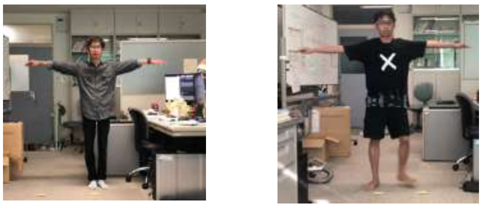

Figure 1. Our dataset samples

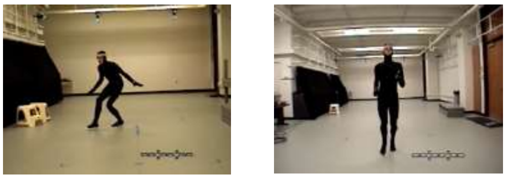

Figure 2. CMU dataset samples

This paper proposes the method to estimate human body motion using video images acquired by an RGB camera. The proposed method estimates one second future motion of human by using the time series estimation method based on the current and the past body motion estimated by OpenPose. On the other hand, human motion as the object of this research is difficult to predict due to the countless motion in human behaviors, as well as the differences of the individual behaviors. For these reasons, we decided to cover the scopes of the human motion for the simple first step to memorize the human motion. Since the RGB camera is more available than other type of camera, we propose the prediction based on the RGB camera data.

2. PRELIMINARIES

2.1 Dataset

This study uses the dataset from the COCO dataset keypoints for the human body pose which contains 18 points consist of human body (e.g., nose, neck, and shoulders)[7,13]. We develop a dataset that contains simple motion such as hand gesture and moving aside. Our dataset consists with 30 fps (frame per second) and frame dimension of 960 × 540 pixels as shown in Fig. 1. As well as our dataset, as a comparison with another complex motion, we use the CMU dataset which contains walking motion. CMU dataset consists 2605 videos with 30 fps and frame dimension of 352 × 240 pixels as shown in Fig. 2.

2.2 Kalman Filter

Kalman Filter is an efficient recursive filter that estimates the internal state of a linear dynamic system from a series of noisy measurements. Kalman Filter has also been used in some applications such as short-term forecasting and the analysis of life lengths from dose-response experiments [12]. Kalman Filter has two steps. The first step is predicting which it makes a first guess about what we think is true (an estimate) and how certain we are that is true (uncertainty). Next, Kalman Filter makes a new guess by using a weighted average. More certain numbers are more important in this weighted average. After doing these two steps, we use the new guess to start these steps again.

2.3 Recurrent Neural Networks (RNN)

RNN is a class of neural network where connections between the computational units form a directed graph along a temporal sequence. Unlike feed-forward networks, RNN can use their internal memory to process arbitrary sequence of inputs. Each of the computing unit in an RNN has a time varying real valued activation and modifiable weight. RNNs are created by applying the same set of weights recursively over a graph-like structure [10]. The learned model in RNN has the same input size, since it has terms of the transition from one state to the other state.

2.4 Long Short-Term Memory (LSTM)

LSTM is an extended version of RNN which has the extended memory to memorize not only the short term of the sequence data, but further long term of the sequence data. LSTM networks were discovered by Hochreiter and Schmidhuber in 1997[8]. LSTM works even given long delays between significant events and can handle signals that mix low and high frequency components.

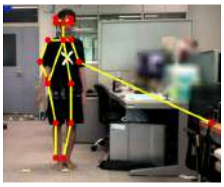

Figure 3: Example of human body parts estimation error.

Figure 4. Frame cropping using YOLOv3.

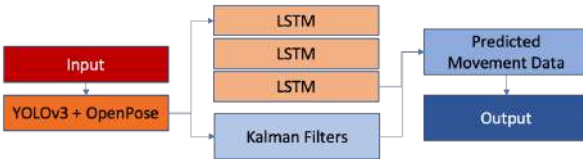

Figure 5. Proposed network model

3. PROPOSED METHOD

We set goal to predict one second ahead of the motion, and we prepared 30 fps videos. Nodes are defined by human body parts which cover head, neck, shoulders, elbows, wrists, hip, knees, and ankles. Before we proceed to obtain the prediction, we need to convert the coordinate data into movement data which contains distance and direction. These movement data obtained from the change of the coordinate from the frame F_i to the frame F_{i+30} with Euclidean formula. After that, we proceed the processing method that includes RNN-LSTM and Kalman Filter for the prediction.

3.1 Feature Extraction

OpenCV implements OpenPose to provide the human body pose estimation in RGB still image that we can use in this research. OpenPose obtains the x and y coordinate value of 18 body points (nose, neck, left/right shoulder, left/right elbow, left/right wrist, left/right hip, left/right knee, and left/right ankle, left/right eye, left/right ear). However, body points given by the OpenPose are not always as reliable as expected (*e.g.* The left wrist point is not estimated correctly in Fig. 3). The proposed method solves this large estimation error by restricting the image region to apply the OpenPose.

The proposed method performs the OpenPose estimation in the human body region estimated using YOLOv3. YOLOv3 is a single neural network that directly predicts bounding boxes and its probability for a class from an input image [13]. As shown in Fig. 4, the proposed method crops the human body region from the original input image frame. The proposed method applies the OpenPose to the cropped image frames, therefore the proposed method can obtain the coordinates of human body parts without large estimation error.

The obtained raw *x* and *y* coordinate values are not suitable to the motion estimation using our estimation model because their value range depends on the image size. Equation 1 and 2 convert the obtained coordinate value at *i*-th frame x_i and y_i to the movement data expression that consists of distance d_i and direction θ_i.

$$d_i = \sqrt{\left(x_i - x_{i-fs}\right)^2 + \left(y_i - y_{i-fs}\right)^2} \qquad (1)$$

$$\theta_i = \arcsin\left(\frac{y_i - y_{i-fs}}{d_i}\right) \qquad (2)$$

where *fs* is the frame step which has a constant value of 30 since the proposed method estimate the one second ahead motion and the input video has 30 fps property.

3.2 Pose Prediction Using Kalman Filter

In Kalman Filter, we have two important function to predict the sequence data. At *i*-th frame, the proposed method predicts the distance d_i and direction θ_i by using Eq. 3 and Eq. 4, respectively.

$$d_i = d_{i-1} + \left(\frac{(\sigma_i \times d_{i-1}) + (\sigma_c \times \delta_i)}{\sigma_i \times \sigma_c}\right) \qquad (3)$$

$$\theta_i = \theta_{i-1} + \left(\frac{(\sigma_i \times \theta_{i-1}) + (\sigma_c \times \delta_i)}{\sigma_i \times \sigma_c}\right) \qquad (4)$$

where σ_i is the initial cost at first and an updated value of the d_i, σ_c represents the constant value of noise, δ_i represents the measurement data which produced by OpenPose.

3.3 Pose Prediction using RNN-LSTM

The Kalman Filter based pose estimation may fail when the human moves suddenly. This study proposes the RNN-LSTM based human motion estimation method.

In this research, the input will be 14 nodes of human body parts and we use three stacked hidden layers for the learning model in RNN-LSTM. Fig. 5 shows the illustration of proposed network model. The last output will be 14 nodes as well as the input. Some other related researches used three stacked layer RNN-LSTM as well [2][3][4]. We used the Mean Squared Error (MSE) that is defined by the following equation to estimate the loss value in the training of RNN-LSTM.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(\hat{x} - x_i)^2 \qquad (5)$$

4. EXPERIMENTAL RESULTS

We performed experiments on the CMU dataset and our dataset. The prediction data are approached from the prediction process, but we are still not sure about the accuracy of these predictions. We set the ground truth prediction for the *i*-th frame F_i to the coordinate data in *i*+30-th frame F_{i+30}. In the experiment, the following equation proposed in the related research [1] calculates the Euclidean distance between two nodes from different frames.

$$E = \sqrt{\left(x_{i+30} - x_p\right)^2 + \left(y_{i+30} - y_p\right)^2} \qquad (6)$$

where *i* is the number of the frame, x_i and y_i represent *x* and *y* coordinate value at *i*-th frame. By using the predicted movement data at *i*-th frame d_i and θ_i, the coordinate value in *i+30*-th frame (x_p, y_p) is calculated by:

$$x_p = x_i + d_i \qquad (7)$$

$$y_p = y_i + d_i \qquad (8)$$

where x_p is the *x* coordinate of the prediction, d_i is the value of distance movement of prediction result, y_p is the *y* coordinate of the prediction.

Figure 6a and 6b show estimation results on our dataset, where the red points are the current position and the blue points are the prediction position. As well as the prediction results on CMU dataset in Fig. 7a and 7b. Table 1 shows the evaluation distances for each node defines the frequency of the value lower than 1.8% of the diagonal frame pixels away from the ground truth in percentage.

Generally, Kalman Filter prediction results show the better result than RNN-LSTM, where neck and right shoulder on our dataset have been obtained 99% of the prediction data are reliable performed by Kalman Filter. Although, RNN-LSTM performed better than Kalman Filter on some nodes like right knee, right ankle, left knee, and left ankle for our dataset. While the rest of the results are varied, elbow and wrist show that RNN-LSTM has a difficulty on predicting the data since elbow and wrist nodes are the human body parts that move more than other nodes in the video. As well as the result of the experiment on CMU dataset, mostly Kalman Filter shows better performance than RNN-LSTM with 88% as the highest frequency of the reliable prediction result that is left knee node. Whereas RNN-LSTM only shows the better result than Kalman Filter on left knee and left ankle with 82%

and 83%, the rest of the prediction result performed by RNN-LSTM have less reliable results. In general, the prediction result on our dataset is more reliable than on CMU dataset.

4. CONCLUSIONS AND DISCUSSIONS

Based on the result of this experiment, we have proposed the human movement prediction with RNN-LSTM and Kalman Filter based on RGB camera for the motion prediction of one second ahead. We used samples of video that cover hand gesture, sideways moving, and walking. The result showed most of the predictions are close to the correct position that is a prediction for one second of human movement. We confirmed the validity of the RGB based method with the unstable data in the simple human motion case from the result, and we conclude that this is an important step to realize the prediction of more complex human motion. For the future works, we need to see the prediction result with combination of RNN-LSTM and Kalman Filter.

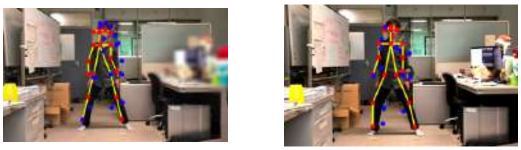
(a) RNN-LSTM (b) Kalman filter

Figure 6. Pose estimation results on our dataset.

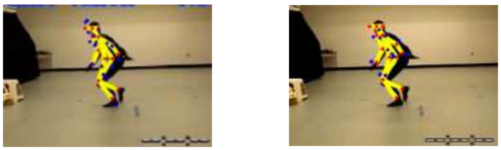
(a) RNN-LSTM (b) Kalman filter

Figure 7. Pose estimation results on CMU dataset.

Table 1: Percentage of evaluation distance lower than 1.8% of the diagonal frame pixels.

Nodes	Our Dataset		CMU Dataset	
	RNN-LSTM	Kalman Filter	RNN-LSTM	Kalman Filter
Head	84	97	57	77
Neck	88	99	46	83
Right Shoulder	93	99	15	79
Right Elbow	39	91	40	72
Right Wrist	29	98	36	67
Left Shoulder	88	98	31	72
Left Elbow	50	94	50	72
Left Wrist	30	93	40	69
Right Hip	88	95	58	84
Right Knee	95	90	63	83
Right Ankle	94	81	59	79
Left Hip	90	98	72	88
Left Knee	95	89	82	81
Left Ankle	93	83	83	74

REFERENCES

[1]. Hiroki Tamura, Kurumi Tsuruta, Etsuo Chosa; Method for estimating locomotive syndrome based on walking motion, Japanese Journal of Clinical Biomechanics, 40, pp.223-229, 2019.

[2]. M. Julieta, B. J. Michael, and R. Javier; On Human Motion Prediction Using Recurrent Neural Networks, arXiv:1705.02445v1, 2017

[3]. Tang Yongyi, Ma Lin, Liu Wei, and Zheng Wei Shi; Long-Term Human Motion Prediction by Modeling Motion Context and Enhancing Motion Dynamics, In:935-941.10.24963/ijcai.2018/130, 2018

[4]. Wu, Erwin, Koike, and Hideki; Real-time human motion forecasting using a RGB camera, 1-2. 10.1145/3281505.3281598, 2018

[5]. C. Yujiao,Z. Weiye, L. Changliu, and T. Masayoshi; Human Motion Prediction using Adaptable Neural Networks, arXiv:1810.00781, 2018

[6]. R. Akita, A. Yoshihara, T. Matsubara and K. Uehara; Deep learning for stock prediction using numerical and textual information, IEEE/ACIS 15th International Conference on Computer and Information Science (ICIS) pp. 1-6. doi: 10.1109/ICIS.2016.7550882, 2016

[7]. Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh; OpenPose: Real-time Multi-Person 2D Pose Estimation using Part Affinity Fields, arXiv:1812.08008[cs.CV]. 2018

[8]. Hochreiter, Sepp Schmidhuber, and Jrgen; Long Short-term Memory. Neural computation, 9.1735-80. 10.1162/neco.1997.9.8.1735, 1997

[9]. Alex Graves; Generating Sequences With Recurrent Neural Networks, arXiv:1308.0850v5 [cs.NE], 2014

[10]. Che Zhengping, Purushotham Sanjay, Cho Kyunghyun, Sontag David, and Liu Yan; Recurrent Neural Networks for Multivariate Time Series with Missing Values, Scientific Reports 8:6085 DOI:10.1038/s41598-018-24271-9, 2018

[11]. S. Selvin, R. Vinayakumar, E. A. Gopalakrishnan, V. K. Menon and K. P. Soman; Stock price prediction using LSTM, RNN and CNN-sliding window model, International Conference on Advances in Computing, Communications and Informatics (ICACCI), Udupi, pp. 1643-1647. doi: 10.1109/ICACCI.2017.8126078, 2017

[12]. Richard J. Meinhold and Nozer D. Singpurwalla; Understanding the Kalman Filter, The American Statistician, 37:2, 123-127, DOI:10.1080/00031305.1983.10482723, 1983

[13]. Asif Sattar; Human detection and distance estimation with monocular camera using YOLOv3 neural network, University of Tartu, Faculty of Science and Technology,Institute of Technology, Master Thesis (30 ECTS), 2019