https://doi.org/10.21289/KSIC.2021.24.6.693

Human Activity Recognition with LSTM Using the Egocentric Coordinate System Key Points

Sheilla Wesonga[1], Jang-Sik Park[2*]

⟨Abstract⟩

As technology advances, there is increasing need for research in different fields where this technology is applied. On of the most researched topic in computer vision is Human activity recognition (HAR), which has widely been implemented in various fields which include healthcare, video surveillance and education. We therefore present in this paper a human activity recognition system based on scale and rotation while employing the Kinect depth sensors to obtain the human skeleton joints. In contrast to previous approaches that use joint angles, in this paper we propose that each limb has an angle with the X, Y, Z axes which we employ as feature vectors. The use of the joint angles makes our system scale invariant. We further calculate the body relative direction in the egocentric coordinates in order to provide the rotation invariance. For the system parameters, we employ 8 limbs with their corresponding angles each having the X, Y, Z axes from the coordinate system as feature vectors. The extracted features are finally trained and tested with the Long short term memory (LSTM) Network which gives us an average accuracy of 98.3%.

Keywords : Human Activity Recognition (HAR), Kinect Depth Sensor, Long Short Term Memory (LSTM)

[1] Dept. of Electronic Eng., Kyungsung University, Student
E-mail: sheilla2@ks.ac.kr

[2*] Corresponding Author, Dept. of Electronic Eng., Kyungsung University, Professor
E-mail: jsipark@ks.ac.kr

1. Introduction

One of the most researched topic in computer vision is Human activity recognition (HAR), and it is widely implemented in various fields which include but not limited to healthcare[1], video surveillance[2] and education[3]. HAR tackles the problem of activity detection for both simple and complex actions performed in the real world environment in the presence of many variables. The most used HAR method is the implementation of high precision sensors[4] for inferring captured action observations. in this paper, we employed the Kinect depth sensors[5] to obtain human skeleton joints known as key points. For each limb on the human skeleton, we obtain the X, Y, Z axes from the coordinate system and these angles are used as the feature vectors.

Recurrent neural networks (RNNs) are famous for modeling long sequence data as a result of their memory state which displays temporal dynamic behavior. Despite their success, RNNs still face a problem of vanishing gradient and for this reason, we implement the Long short term memory (LSTM) network for activity recognition. LSTM is a recurrent neural network capable of handling long sequence data and can model wide-range temporal dependencies, and most importantly, LSTM avoids the problem of vanishing gradient faced by the RNNs[6]. We implement the LSTM to recognize 5 human activities namely; walking, punching, clapping, lifting and standing. We obtain a 98.3% accuracy from the confusion matrix used to evaluate the model.

The following chapters describe the egocentric coordinate system with which we extract joint points from limbs of the human skeleton. We further elaborate the implementation of the LSTM network for the HAR task and results from the experiment, and conclude with what we hope to accomplish in the future.

2. Methodology

2.1 Egocentric Coordinate System

Let $\hat{u}$ be a unit vector denoted as:

$$\hat{u} = u\hat{u}$$

and

$$\hat{u} = u_x\hat{i} + u_y\hat{j} + u_z\hat{k}$$

where

$\hat{i} = (1,0,0)$ is the unit vector on the x axis, $\hat{j} = (0,1,0)$ is the unit vector on the y axis and $\hat{k} = (0,0,1)$ is the unit vector on the z axis.

For two known vectors $\hat{u}$ and $\hat{v}$ the third vector orthogonal to both vectors is found by calculating their cross product as:

$$|\vec{u}\times\vec{v}| = uv\sin\theta$$

where

θ is the angle between $\vec{u}$ and $\vec{v}$.

Normalizing the axes (X, Y, Z) also normalizes the limb vectors shown in Fig. 1. which makes it easy to get the angles between the three axes X, Y, Z separately and each limb vector. Finally, we get the angle between the two vectors in a range of $[0, \pi]$ by applying the arccosine to the dot product.

For us to construct the egocentric coordinate system that is relative to the direction of the human body, we assume that: the x axis is a normalized vector between the left and right hips, and also, the y axis is the vertical unit vector of the general coordinate system. The depth axis z is constructed by calculating the cross product of the known vectors on the x and y axes.

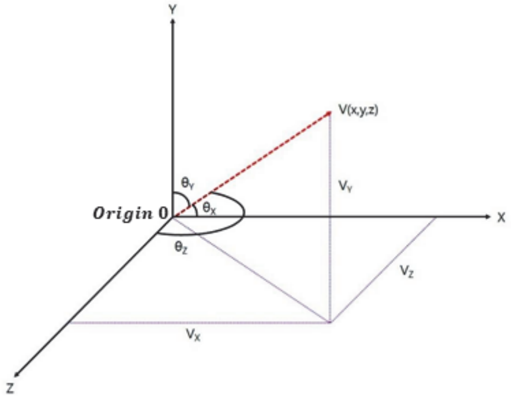


Fig. 1 Limb vector and X, Y, Z axes angles in the egocentric system

2.2 RNNs and LSTM

RNNs can be described as a class of artificial neural networks where connections between nodes form a directed graph along a temporal sequence, and this allows it to exhibit temporal dynamic behavior[7]. Using loops, RNNs can keep information across time with their design also enabling them to process variable length sequences of input data. However, as the sequence grows longer, it becomes harder for RNN to recognize the interrelationship between the components of the same sequence which are far apart from each other.

The challenge facing the RNNs architecture is the problem of short-term memory because the gradients tend to explode or vanish known ass the "vanish gradient problem"[8,9]. To solve this problem, LSTM is a mechanism that is popularly implemented as it was specifically designed to answer the issue of long-term dependencies remembering information for a long time as its default state[10,11]. The LSTM network built on the basis of the cell state which runs across the network in a straight line with minimal linear interaction, hence allowing easy information flow.

As seen in Fig. 2, while a regular RNN node contains only simple calculation (e.g. tanh operation) which consists of a single neural network layer, the LSTM cell carries 4 neural network layers. These additional neural network layers and linear operators together

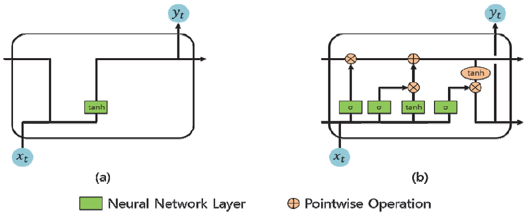


Fig. 2 simple RNN node (a) and an LSTM cell (b)

form structures called gates (i.e.; input gate, forget gate, output gate)[12]. Therefore, the LSTM the network is able to handle longer sequences of sentences better compared to the standard RNN network.

The illustration shows the RNN node that contains a single neural network layer, in contrast, the LSTM cell carries 4 layers interacting uniquely. The LSTM is elevated by recurrent gates known as forget gates.

3. Implementation and Results

The limb vectors extracted from the egocentric coordinate system are used as input to the LSTM network. In Table 1. we summarize the parameters we selected for the training and testing process.

The input data includes ground truth for five different activities (standing, walking, punching, lifting and clapping) shown in Fig. 3. The data was split into 70:30 for training and testing respectively.

The accuracy graph is shown in Fig. 4. while For evaluating the LSTM model for the recognised activities, we used the confusion matrix shown in Fig. 5. where we obtained

Table 1. Specification and libraries used in this paper

Category	Hyper-Parameters	Value
Data Preprocessing	Time Step Window Size Batch Size Epochs	10 100 64 75
Network Architecture	Hidden Layers Number of Neurons	2 30
Training	Activation Function Bias Weight Initialization	Softmax constant=1.0
Learning	Optimizer Learning Rate Loss Rate	Adam 0.0025 0.0015

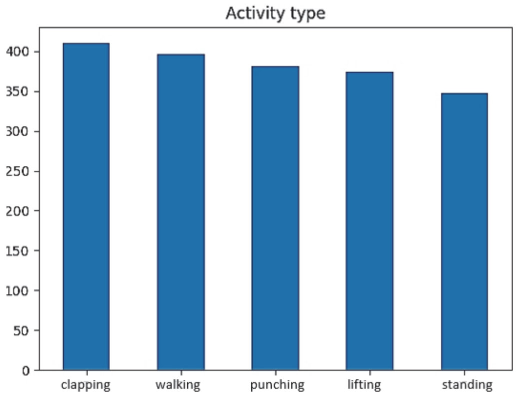


Fig. 3 HAR data set distribution

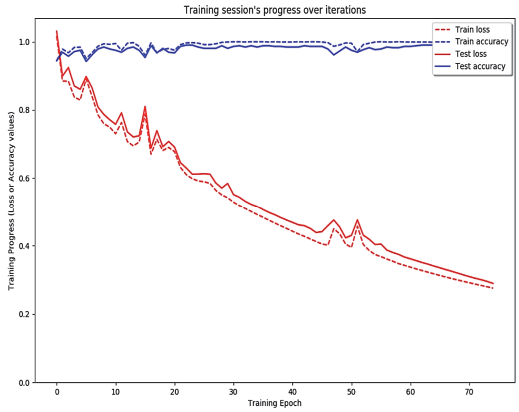


Fig. 4 Training (blue) and testing (red) accuracy and loss

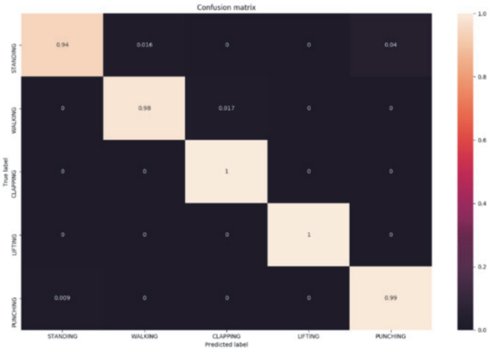
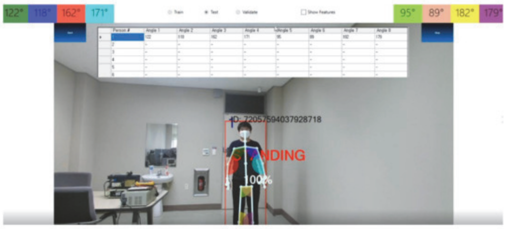


Fig. 5 Confusion matrix.



○ Train ● Test ○ Validate □ Show Features

	Person #	Angle 1	Angle 2	Angle 3	Angle 4	Angle 5	Angle 6	Angle 7	Angle 8
▸		122	118	162	171	95	89	182	179
	2	-	-	-	-	-	-	-	-
	3	-	-	-	-	-	-	-	-
	4	-	-	-	-	-	-	-	-
	5	-	-	-	-	-	-	-	-
	6	-	-	-	-	-	-	-	-

Fig. 6 Inference on the activity labeled standing

an average accuracy of 98.3% which we believe can be improved upon with more focus put on the activity recognition network.

An example of inference performed using the kinect camera is shown below in Fig. 6.

4. Conclusion

Human activity recognition has diverse applications and in this paper we implemented an egocentric coordinate system for extracting X, Y, Z axes for every limb in the human skeleton represented as feature vectors, using the Kinect depth sensor. The use of the joint angles made our system scale invariant. We further calculated the body relative direction in the egocentric coordinates in order to provide the rotation invariance. For the system parameters, we employed 8 limbs with their corresponding angles each having the X, Y, Z axes from the coordinate system as feature vectors. The features were implemented in the Long Short Term Memory network, a recurrent neural network capable of handling long-term sequence data, for the task of activity recognition. We obtained a 98.3 accuracy from the LSTM evaluation for five different activities (walking, punching, lifting, standing, clapping).

In future, we plan on testing and implementing the system in real time scenarios using the CCTV surveillance cameras.

Acknowledgement

This research was supported by the Kyungsung University Research Grants in 2021 and Brain Busan 21+ 2021.

References

[1] M. Tentori and J. Favela, “Activity-aware computing for healthcare”, Pervasive Computing IEEE, vol. 7, pp. 51-57, 2008.

[2] K. K. Htike, O. O. Khalifa, H. A. Mohd Ramli and M. A. M. Abushariah, “Human activity recognition for video surveillance using sequences

of postures," The Third International Conference on e-Technologies and Networks for Development (ICeND2014), pp. 79-82, 2014.

[3] A. Jalal and M. A. Zeb, Security Enhancement for E-learning portal, Int. J. Comput. Sci. Netw. Security 8 (2008), no. 3, 41– 45.

[4] J. Wang, Y. Chen, S. Hao, X. Peng, & L. Hu, "Deep learning for sensor-based activity recognition: a survey. Pattern Recogn". Lett. 119, 3–11 (2019).

[5] L.F. Yeung, Z. Yang, K.C.C. Cheng, D. Du. and R.K.Y. Tong, Effects of camera viewing angles on tracking kinematic gait patterns using Azure Kinect, Kinect v2 and Orbbec Astra Pro v2. Gait & Posture, 87, pp. 19-26. 2021.

[6] Y. Zhao," Deep Residual Bidir-LSTM for Human Activity Recognition Using Wearable Sensors," 1-5, 2018.

[7] J. Guo, K. Tian, K. Ye and C. -Z. Xu, "MA-LSTM: A Multi-Attention Based LSTM for Complex Pattern Extraction," 2020 25th International Conference on Pattern Recognition (ICPR), pp. 3605-3611, 2021.

[8] J. F. Kolen; S. C. Kremer, "Gradient Flow in Recurrent Nets: The Difficulty of Learning LongTerm Dependencies," in A Field Guide to Dynamical Recurrent Networks, IEEE, pp. 237-243, 2001.

[9] Y. Bengio, P. Simard and P. Frasconi, "Learning long-term dependencies with gradient descent is difficult," in IEEE Transactions on Neural Networks, vol. 5, no. 2, pp. 157-166, March 1994.

[10] H. Sepp & S. Jürgen. Long Short-term Memory. Neural computation. 9. 1997.

[11] Y. Zhao, Deep Residual Bidir-LSTM for Human Activity Recognition Using Wearable Sensors. Math. Probl. Eng. 2018, 2018.

[12] FJ, Ordóñez, D. Roggen Deep Convolutional and LSTM Recurrent Neural Networks for Multimodal Wearable Activity Recognition. Sensors. 2016

[13] J. Wang, Y. Chen, S. Hao, X. Peng, L. Hu, Deep learning for sensor-based activity recognition: A survey, Pattern Recognition Letters, Volume 119, Pages 3-1 2019.

[14] S. Qiu, H. Zhao, N. Jiang, Z. Wang, L. Liu, Y. An, H. Zhao, X. Miao, R. Liu, G. Fortino, Multi-sensor information fusion based on machine learning for real applications in human activity recognition: State-of-the-art and research challenges, Information Fusion, Volume 80.

(Manuscript received November 08, 2021;
revised December 03, 2021; accepted December 03, 2021)