Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 100 (2016) 78 – 84

Procedia Computer Science

Conference on ENTERprise Information Systems / International Conference on Project MANagement / Conference on Health and Social Care Information Systems and Technologies, CENTERIS / ProjMAN / HCist 2016, October 5-7, 2016

A Depth Camera-based Human Activity Recognition via Deep Learning Recurrent Neural Network for Health and Social Care Services

S. U. Park[a], J. H. Park[a], M. A. Al-masni[a], M. A. Al-antari[a], Md. Z. Uddin[b], T. -S. Kim[a]*

[a]*Department of Biomedical Engineering, Kyung Hee University, Yongin, Republic of Korea*
[b]*Department of Computer Education, Sungkyunkwan University, Seoul, Republic of Korea*

Abstract

Human activity recognition (HAR) has become an active research topic in the fields of health and social care, since this technology offers automatic monitoring and understanding of activities of patients or residents. Depth camera-based HAR recognizes human activities using features from depth human silhouettes via conventional classifiers such as Hidden Markov Model (HMM), Conditional Random Fields etc. In this paper, we propose a new HAR system via Recurrent Neural Network (RNN) which is one of deep learning algorithms. We utilize joint angles from multiple body joints changing in time which are represented a spatiotemporal feature matrix (i.e., multiple body joint angles in time). With these derived features, we train and test our RNN for HAR. In order to evaluate our system, we have compared the performance of our RNN-based HAR against the conventional HMM- and Deep Belief Network (DBN)-based HAR using a database of Microsoft Research Cambridge-12 (MSRC-12). Our test results show that the proposed RNN-based HAR is able to recognize twelve human activities reliably and outperforms the HMM- and DBN-based HAR. We have achieved the average recognition accuracy of 99.55% for the activities. The results are 7.06% more accurate than that of the HMM-based HAR and 2.01% more accurate than that of the DBN-based HAR.

Peer-review under responsibility of the organizing committee of CENTERIS 2016

* Corresponding author. Tel.: +82-31-201-3731; fax: +82-31-201-3731;
E-mail address: tskim@khu.ac.kr

Peer-review under responsibility of the organizing committee of CENTERIS 2016
doi:10.1016/j.procs.2016.09.126

Keywords: Human Activity Recognition, Depth Imaging Sensor, Deep Learning, Recurrent Neural Network, Health Care

1. Introduction

Human activity recognition (HAR) is a technique to recognize various human activities via external sensors such as inertial or video sensors. In recent years, HAR has evoked significant interest among researchers in the areas of health care, social care, and life care services[1], since it allows automatic monitoring and understanding of activities of patients or residents in smart environments such as smart hospitals and smart homes. For instance, at smart home, a HAR system can automatically recognize residents' activities and create daily, monthly, and yearly activity logs. These life logs can provide residents' habitual patterns which medical doctors evaluate for further health care suggestions. Especially for elderly people, a HAR system can recognize their falls or unusual activity patterns and alert or inform their caregivers.

The basic methodology of activity recognition involves activity feature extraction, modeling, and recognition techniques. Video-based HAR is a challenging task as it has to consider whole body movement of a human and does not follow rigid syntax like hand gestures or sign languages. Hence, a complete representation of a full human body is essential to characterize human movements properly in this regard. Though many researchers have been exploring video-based HAR systems due to their practical applications, accurate recognition of human activities still remains as a major challenge.

Generally, video-based HAR can be divided into two categories according to motion features: namely, marker-based and vision-based. The former is based on a wearable optical marker-based motion capture (MoCap) system that is widely used as it offers an advantage of accurately capturing complex human motions. However, it has a disadvantage that the optical sensors must be attached to the body and requires multiple camera settings. The latter is based on depth video cameras and it is marker-free[2]. This approach is getting more attention these days due to the absence of tracking wearable markers, hence making the HAR system easy to be deployed in daily applications.

As for the recognition techniques, until now HMMs have been widely used in many HAR systems, as HMMs are capable of temporal pattern decoding[3-5]. However, recently recognition via deep learning is getting considerable attentions due to its power to learn deep structures of patterns[6-12]. Basically, deep learning refers to neural networks that exploit layers of non-linear data processing for feature classification. These layers are hierarchically organized and process the outputs of the previous layers. Deep learning techniques have outperformed many traditional methods in computer vision[8-12]. Deep learning techniques are very promising to address the requirements of HAR in two ways. First, performance can be significantly improved over conventional recognition techniques. Second, deep learning approaches have the potential to uncover features that are tied to the dynamics of human motion (i.e., from simple motion encoding in lower layers to more complex motion dynamics in upper layers of the network). This may be useful to scaling up HAR to activities that are more complex.

Recently, there has been a HAR work via Deep Belief Network (DBN)[6] which is one of Deep Neural Networks (DNNs) proposed by Hilton in 2004[7]. DBN uses Restricted Boltzmann Machines (RBMs) in learning and it avoids local minimum problem with less training time. However, Recurrent Neural Networks (RNNs) is a better choice than DBN, since it could offer more discriminative power over DBN as time sequential information can be encoded or learned through RNNs. Although HMM can handle time sequential information, now researchers prefer RNN over HMM for its improved discriminant capability.

In this paper, we present a RNN-based HAR system. We have performed HAR with the features of body joint angles. The performance of RNN for HAR has been compared to other conventional recognizers such as HMM and DBN.

2. Materials and Methods

In this section, we introduce our RNN-based HAR system. Our HAR system proceeds to the following steps. First, we create an input feature matrix of joint angles computed from the MSRC-12 activity dataset[13]. Second, we train RNN with the training feature matrix. Third, we evaluate the trained RNN with test data sets by recognizing

twelve human activities. The recognition performance is compared to the results from the conventional HMM- and DBN-based HAR systems.

2.1. MSRC-12 Gesture Dataset and Features

We have evaluated the HAR systems using the MSRC-12 dataset. This dataset consists of sequences of human activities containing 12 activities: namely lift arms, duck, push right, goggles, wind it up, shoot, bow, throw, had enough, change weapon, beat both and kick respectively (denoted as A_1 through A_{12}) from 30 human subjects using a depth camera. In total, there are 6,244 activities. The dataset includes the human skeletal joint positions at each frame.

To avoid the dependency on body joint locations for HAR, from 14 key body parts we extracted 28 joint angle features (i.e., two joint angles from each part). The 14 key body parts are spine, neck, right lower arm, right upper arm, right shoulder, left lower arm, left upper arm, left shoulder, right hip, right upper leg, right lower leg, left hip, left upper leg, and left lower leg respectively. The joint angles were calculated in a spherical coordinate so that they are invariant to position and scale. We defined the number of frame for one activity is 50. Finally, we created the input feature matrix with 28 joint features from 50 frames, making the size of each input feature matrix (28×50). Each row of the feature matrix represents a change in joint angle in time.

2.2. HMM-based HAR

On the feature matrix, we perform Principal Component Analysis (PCA) to reduce a dimension of the feature vector from 28 to 17, which includes more than 99% of information of the frame. Then each of the reduced feature vectors of (1×17) is clustered to be one of 64 codes via Linde-Buzo-Gray algorithm[14]. Then a set of 50 frames is represented in a (50×1) sequence of codes. Lastly, HMMs are trained with the sequences of codes via Baum-Welch algorithm. Details of our settings for HMMs are available in[5]. After training HMM, we have applied it for HAR.

2.3. DBN-based HAR

For DBN-based HAR, we use a vector of (1x1400) from 28 joint features from 50 frames. Training DBN requires two steps: pre-training and fine-tuning. Pre-training is a process of determining the appropriate initial weights to avoid local minimum solution in the network. This step initializes Restricted Boltzmann Machines[15]. The weights of RBMs are adjusted in a fine-tuning step through backpropagation. After training DBN, we have applied the system for HAR. More details can be found in the work[6].

2.4. RNN-based HAR

In our feature data, human activities are represented as time sequential changes in multiple joint angles. For this reason, a recognizer capable of encoding time sequential information is needed. Therefore, we have adopted Recurrent Neural Networks (RNNs) which is one of the sequential Deep learning approaches[16]. RNNs have recurrent connections between hidden units in their structure which connect past information to current information. That is a role of the memory in neural networks. In general, basic RNNs have a vanishing gradient problem, so having a limitation of processing long term information. This is known as the problem of Long-Term Dependencies. As a solution to this vanishing gradient problem, Long Short-Term Memory (LSTM) was proposed by Sepp Hochreiter and Jurgen Schmidhuber[17].

As shown in Fig. 1, our designed RNN consists of 50 LSTMs with 90 hidden units. In our RNN model, the number of LSTMs reflects the length of the activity video frames.

Each LSTM block contains a cell state and three gates including input gate, forget gate, and output gate as indicated in Fig. 1. The input gate i_t determines which values are updated as expressed in Equation (1). The forget gate f_t determines what information is thrown away as expressed in Equation (2). The long-term memory is stored in a cell state vector of c_t as expressed in Equation (3)[16, 18]. The output gate o_t determines what is going to be an

output as expressed in Equation (4) and the hidden state h_t is a multiply of the output gate o_t and the non-linear transformed cell state c_t as expressed in Equation (5).

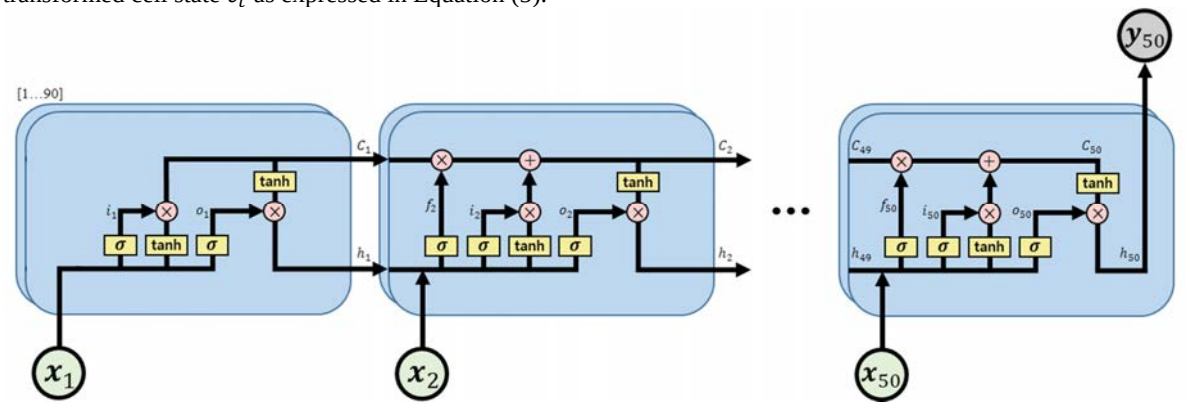

Figure 1 Our designed RNN architecture with fifty LSTMs

$$i_t = \sigma(W_{xi}x_t + W_{hi}h_{t-1} + b_i) \quad (1)$$

$$f_t = \sigma\left(W_{xf}x_t + W_{hf}h_{t-1} + b_f\right) \quad (2)$$

$$c_t = f_t c_{t-1} + i_t tanh(W_{xc}x_t + W_{hc}h_{t-1} + b_c) \quad (3)$$

$$o_t = \sigma(W_{xo}x_t + W_{ho}h_{t-1} + b_o) \quad (4)$$

$$h_t = o_t tanh(c_t) \quad (5)$$

where x_t is an input sequence. W denotes weight matrices. b denotes bias vectors. σ is the logistic sigmoid function. i, f, o are the input gate, forget gate, output gate respectively. h is the hidden state vector[16].

$$y = softmax(W_y h_{50} + b_y) \quad (6)$$

The output vector of y comes from the hidden state vector of h_{50} at the last time step of 50 which is multiplied by the weight matrix and added a bias as expressed in Equation (6). We use the softmax function as the network output activation function[16].

Our RNN is trained with the training feature data via an extended backpropagation algorithm called Backpropagation Through Time (BPTT)[19-21]. To optimize our RNN, we used Adam optimizer which an algorithm for the first-order gradient-based optimization of stochastic objective functions[22].

Fig. 2 shows the processes of our implemented RNN-based HAR system.

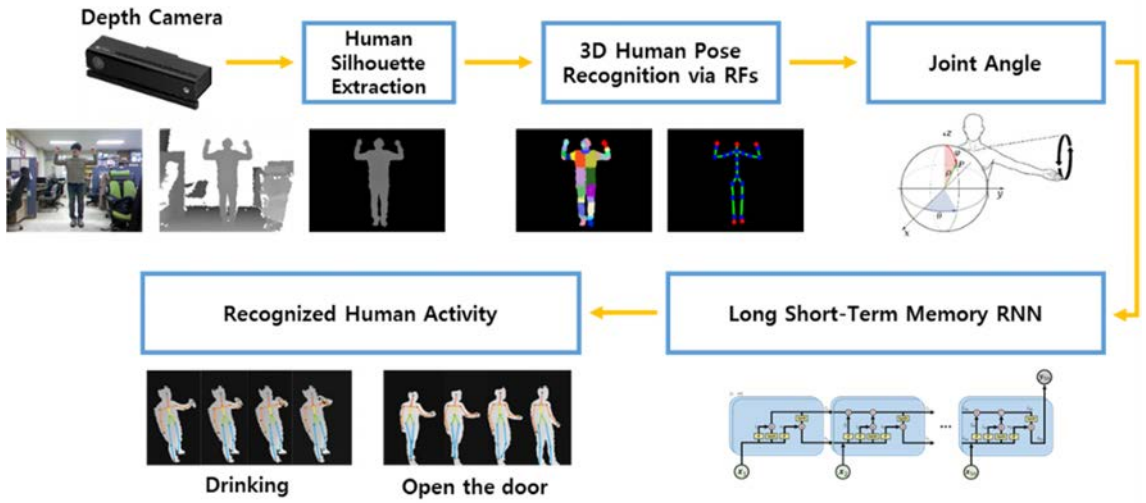

Figure 2 Our RNN-based HAR system

3. Results

To evaluate our HAR system, we have compared the accuracy of HAR based on HMM, DBN and RNN respectively using the same MSRC-12 dataset. Table 1 shows the accuracies of the proposed HAR system via RNN and the comparison results of using HMM and DBN.

The results by HMM show the lowest performance compared to the other two cases. The results show that the average accuracy of 92.49% with standard deviation of 4.48. The results by DBN show the higher recognition rate than those of HMM. The accuracy of DBN-based HAR is 97.54% with standard deviation of 1.92 [6]. The results by RNN show the best recognition accuracy of 99.55% with standard deviation for 0.11. The RNN-based HAR is 7.06% more accurate than those of HMM-based HAR are and 2.01% more accurate than those of DBN-based HAR. Table 2 shows a confusion matrix for 12 activities using RNN.

Table 1. Comparison of accuracies of HAR based on HMM, DBN and RNN.

Activities	Recognition Accuracy (%)		
	HMM	DBN	RNN
A_1	87.5	94.3	100
A_2	98.8	98.8	99.40
A_3	84.2	98.8	100
A_4	91.8	96.9	99.23
A_5	86.1	98.7	99.38
A_6	97.6	100	99.40
A_7	92.9	98.8	100
A_8	95.1	93.8	100
A_9	93.9	98.0	98.49
A_{10}	94.8	96.1	100
A_{11}	92.4	98.7	99.06
A_{12}	94.9	97.4	100
Mean (SD)	92.49 (4.48)	97.54 (1.92)	99.55 (0.11)

Table 2. The confusion matrix of HAR with the proposed RNN system

Activities	A_1	A_2	A_3	A_4	A_5	A_6	A_7	A_8	A_9	A_{10}	A_{11}	A_{12}
A_1	**100**											
A_2		**99.40**							**0.60**			
A_3			**100**									
A_4				**99.23**								**0.77**
A_5			**0.63**		**99.38**							
A_6						**99.40**			**0.60**			
A_7							**100**					
A_8								**100**				
A_9				**1.01**					**98.49**	**0.50**		
A_{10}										**100**		
A_{11}				**0.63**						**0.31**	**99.06**	
A_{12}												**100**

4. Conclusions

In this paper, we present the work of RNN-based HAR. Our proposed RNN-based HAR achieves much higher accuracy than the conventional HMM- or DMB-based HAR. We believe this improvement is due to the time sequential encoding of activity features (i.e., temporal changes in joint angle features). We expect that our proposed HAR system using RNN should be a better choice of HAR in the areas of health care or social care services by automatically monitoring and recognizing the activities of patients or residents in smart environments.

Acknowledgements

This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MEST) (2014R1A2A2A09052449).

References

1. Nam S.B., Park S.U., Park J.H., Uddin MZ, Kim T.-S. Accurate 3D human pose recognition via fusion of depth and motion sensors. *International Journal of Future Computer and Communication.* 2015;4.5:336-340.
2. Jalal A, Uddin MZ, Kim T.-S. Depth video-based human activity recognition system using translation and scaling invariant features for life logging at smart home. *Consumer Electronics, IEEE Transactions on.* 2012;58.3:863-871.
3. Iengo S, Rossi S, Staffa M, Finzi A. Continuous gesture recognition for flexible human-robot interaction. *Robotics and Automation (ICRA), 2014 IEEE International Conference on.* IEEE; 2014:4863-4868.
4. Piyathilaka L, Kodagoda S. Gaussian mixture based HMM for human daily activity recognition using 3D skeleton features. *Industrial Electronics and Applications (ICIEA), 2013 8th IEEE Conference on.* IEEE; 2013:567-572.
5. Jalal A, Sarif N, Kim JT, Kim T.-S. Human activity recognition via recognized body parts of human depth silhouettes for residents monitoring services at smart home. *Indoor and Built Environment.* 2013;22.1:271-279.
6. Nam S.B., Park S.U., Park J.H., Kim T.-S. A single depth sensor based human activity recognition via deep belief network. *Proceedings of the 4th World Conference on Applied Sciences, Engineering and Technology 24-26 October 2015.* Japan: Kumamoto University; 2015:015-019.
7. Hinton GE, Osindero S, Teh Y. A fast learning algorithm for deep belief nets. *Neural Comput.* 2006;18.7:1527-1554.
8. Krizhevsky A, Sutskever I, Hinton GE. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems.* 2012:1097-1105.
9. LeCun Y, Bengio Y. Convolutional networks for images, speech, and time series. *The handbook of brain theory and neural networks.* 1995;3361.10:1995.
10. Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556.* 2014.
11. Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, Erhan D, Vanhoucke V, Rabinovich A. Going deeper with convolutions. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.* 2015:1-9.

12. Duffner S, Berlemont S, Lefebvre G, Garcia C. 3D gesture classification with convolutional neural networks. *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on.* IEEE; 2014:5432-5436.
13. Fothergill S, Mentis H, Kohli P, Nowozin S. Instructing people for training gestural interactive systems. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems.* ACM; 2012:1737-1746.
14. Linde Y, Buzo A, Gray RM. An algorithm for vector quantizer design. *Communications, IEEE Transactions on.* 1980;28.1:84-95.
15. Minsky M, Papert S. Perceptron: An introduction to computational geometry. *The MIT Press, Cambridge, expanded edition.* 1969;19.88:2.
16. Graves A, Mohamed A, Hinton G. Speech recognition with deep recurrent neural networks. *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on.* IEEE; 2013:6645-6649.
17. Hochreiter S, Schmidhuber J. Long short-term memory. *Neural Comput.* 1997;9.8:1735-1780.
18. Zaremba W, Sutskever I, Vinyals O. Recurrent neural network regularization. *arXiv preprint arXiv:1409.2329.* 2014.
19. Gers FA, Schraudolph NN, Schmidhuber J. Learning precise timing with LSTM recurrent networks. *The Journal of Machine Learning Research.* 2003;3:115-143.
20. Sak H, Senior AW, Beaufays F. Long short-term memory recurrent neural network architectures for large scale acoustic modeling. *INTERSPEECH.* 2014:338-342.
21. Williams RJ, Peng J. An efficient gradient-based algorithm for on-line training of recurrent network trajectories. *Neural Comput.* 1990;2.4:490-501.
22. Kingma D, Ba J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980.* 2014.
23. Christopher Olah. Understanding LSTM Networks. Available at: http://colah.github.io/posts/2015-08-Understanding-LSTMs/. Accessed 4/13/2016, 2016.