A Single Depth Sensor Based Human Activity Recognition via Convolutional Neural Network

J.H. Park, S.U. Park, Md. Zia Uddin, M.A. Al-antari, M.A. Al-masni, and T.-S. Kim

Abstract

Human activity recognition (HAR) has become an active research topic in the various fields. Depth sensor-based HAR recognizes human activities using features from depth human silhouettes via classifiers such as Hidden Markov Model (HMM), Conditional Random Fields Model etc. In this paper, we propose a new HAR system via Convolutional Neural Network (CNN), one of deep learning algorithms. We extract joint angles from multiple body joints changing in time and create a spatiotemporal feature matrix (i.e., multiple body joint angles in time). With these derived features, we train and test our CNN for HAR. In order to evaluate our system, we have compared the performance of our CNN-based HAR against the 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 CNN-based HAR is able to recognize twelve human activities reliably and it outperforms the HMM- and DBN-based systems. We have achieved the average recognition accuracy of 98.59% for the activities. The results are 6.1% more accurate than that of the HMM-based HAR and 1.05% more accurate than that of the DBN-based HAR.

Keywords

Human activity recognition • Depth imaging sensor • Deep learning • Convolutional neural network

1 Introduction

Human activity recognition (HAR) is to recognize various human activities via external sensors such as acceleration or video sensors. In recent years, HAR from video has evoked significant interest among researchers in the areas of computer vision, e-healthcare, lifecare, human computer interface, etc. [1]. In fact, human activity recognition exhibits practical applications such as human computer interaction, automated surveillance, and human healthcare. For instance, in a smart environment, an automatic human activity recognition system can recognize residents' activities and can create daily, monthly, and yearly activity logs. These life logs can provide residents' habitual patterns, which medical doctors evaluate for further healthcare suggestions. 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 and does not follow rigid syntax like hand gestures or sign languages. Hence, a

J.H. Park · S.U. Park · M.A. Al-antari · M.A. Al-masni · T.-S. Kim (✉)
Department of Biomedical Engineering, Kyung Hee University, Kyunghee University Global Campus, Seocheon-dong, Giheung-gu, Yongin-si, Gyeonggi-do, Republic of Korea
e-mail: tskim@khu.ac.kr

J.H. Park
e-mail: psmt2655@khu.ac.kr

Md.Zia Uddin
Department of Computer Education, Sungkyunkwan University, Seoul, Republic of Korea

T. Vo Van et al. (eds.), *6th International Conference on the Development of Biomedical Engineering in Vietnam (BME6)*, IFMBE Proceedings 63, https://doi.org/10.1007/978-981-10-4361-1_92

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 an optic wearable marker-based motion capture (MoCap) system that is widely used as it offers an advantage of accurately capturing complex human motions. However, it has the disadvantage that the optical sensors must be attached to the body and requires multiple camera settings. The latter is based on RGB or depth image sensors and it is marker-free. 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 [2–4]. Recently, deep learning is getting considerable attentions due to its power to learn deep structures of patterns [5–11]. Basically, deep learning refers to neural networks that exploit layers of non-linear data processing for feature classification which is hierarchically organized where each layer processing the outputs of the previous layer. Deep learning techniques have outperformed many traditional methods in computer vision [7–11]. Deep learning techniques are very promising to address the requirements of HAR in two ways. First, performance can be significantly improved over existing 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, in [5], the authors recognized activities via Deep Belief Network (DBN) which is one of Deep Neural Networks (DNNs) proposed by Hilton in 2004 [6]. DBN uses Restricted Boltzmann Machines (RBMs) in learning and it avoids local minimum problem with less training time. In addition to DBN, Convolutional Neural Networks (CNNs) are also attracting many researchers due to their more discriminative power over DBN. CNN is a type of DNN consisting of feature extractions and several convolutional stacks to create a progressive hierarchy of more abstract features. One of key attributes of CNN is to conduct different processing units such as convolution, pooling, sigmoid/hyperbolic tangent squashing, rectifier, and normalization [7]. Such various processing units yield an effective representation the features: this deep architecture allows multiple layers of the processing units to be stacked so that this deep learning model can characterize the salience of features in different scales. Moreover, the features extracted via CNN exhibit more discriminative power as it can learn under the supervision of output labels. Yann Lecun and Yoshua Bengio introduced the concept of CNN in 1995 [8]. Later on, various structures of CNN were proposed including AlexNet [7], VGGNet [9], and GoogleNet [10].

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

2 Materials and Methods

In this section, we introduce our CNN-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 [12]. Second, we train CNN with the training feature matrix. Third, we evaluate the trained CNN using test data sets 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 594 sequences and 719,359 frames (approx. 6 h. 40 min.) collected from 30 people performing 12 activities. In total, there are 6244 activities. Twelve different activities are denoted as G1–G12, indicating the following activities: lift arms, duck, push right, goggles, wind it up, shoot, bow, throw, had enough, change weapon, beat both, and kick respectively.

For HAR, from 14 key body parts, we derive 28 joint angle features (i.e., two joint angles from each part). The 14 key body parts include 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. Finally, we create 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 [13]. 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 [4]. After training HMM, we have applied it for HAR.

2.3 DBN-Based HAR

For DBN-based HAR, we use a vector of (1 × 1400) 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 weight to avoid local minimum solution in network. This step initializes Restricted Boltzmann Machines [14]. 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 [5].

2.4 CNN-Based HAR

Convolutional Neural Network is a kind of multilayer perceptron that is designed to use minimum preprocesses [10]. In general, the structure of CNN consists of multiple layers including convolution layer, pooling layer, and fully connected layer. Compared to other deep learning structure, CNN shows a good performance in the video and audio sector. CNN has the advantage of using a small number of the bias values and weight values than other deep learning approaches.

Figure 1 shows the structure of our proposed CNN for HAR. This structure consists of seven layers, including three convolution layers, three pooling layers, and one fully connected layer. The convolution layers and pooling layers pass input feature matrix 3 times repeatedly. The output is classified into 12 activities through a single fully connected layer.

The first layer is the convolution layer called c^1. In this layer, the input matrix is convolved with a 1 × 3 convolution kernel and the matrix of (28 × 48) is generated. Note the convolution kernel is 1D, since each body joint is independent of each other and only temporal convolution is performed in each row. Equation (1) represents the convolution layer.

$$c_k^{(l+1)}(m,n) = ReLU\left(\sum_{(g=1)}^{u} x(m, n-g+(u+1)/2)w_k^l(g) + b_k^l\right) \tag{1}$$

where $c_k^{l+1}(m,n)$ indicates (m, n) coordinates of the l + 1-th layer and k-th convolution map. w_k^l represents the l-th layer and k-th convolution kernel. b_k^l represents the l-th layer and k-th bias value. x represents the map of previous layer and u represents the size of kernel. ReLU is one of the active functions, which receives the weight sum value of the previous layer and then passes to the next layer. CNN usually uses ReLU which is expressed as Max(0,x) to the active function [11].

The second layer is the pooling layer called s^1. This layer converts the result of the c^1 layer to a matrix of (28 × 24) using a (1 × 2) max pooling. Equation (2) represents the pooling layer. We select the maximum value in the 1 × 2 window of previous map.

$$p_k^{l+1}(m,n) = max_{1 \le g \le u} x(m, (n-1) * u + g) \tag{2}$$

Then repeat the convolution layer and pooling layer two more times. In the c^2 and s^2 layers, the result of s^1 is

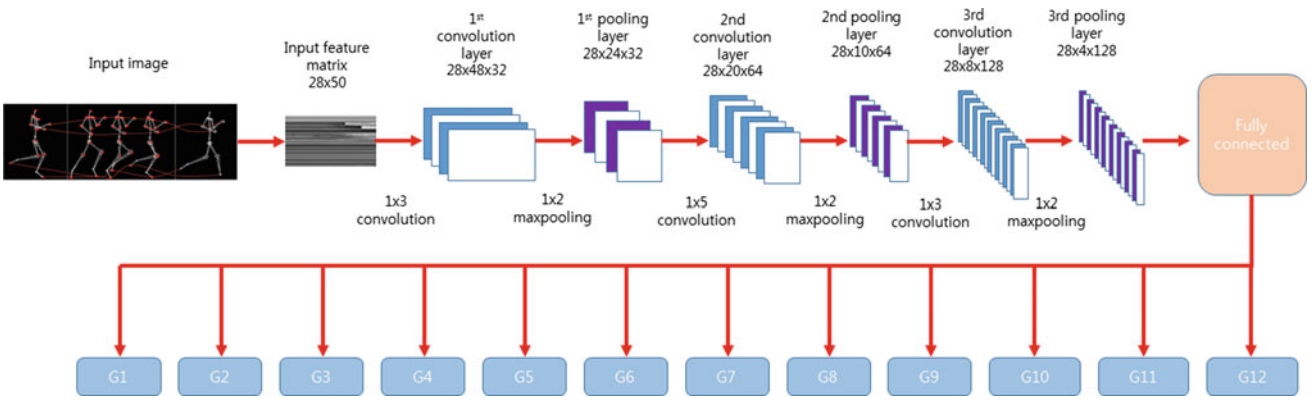

Fig. 1 Structure of our proposed CNN for HAR consisting of seven layers, including three convolution layers, three pooling layers, and one fully connected layer

convolved with a 1×5 convolution kernel and convolution map is reduced in its size, so we generate a matrix of (28×10). In the c^3 and s^3 layers, we generate a matrix of (28×4) using a 1×3 convolution kernel and 1×2 max-pooling.

Finally, we generate a fully connected layer using the result of s^3. Equation (3) represents the pooling layer. w_{ij}^l represents the i-th node of the l-th layer to the j-th node of the $l + 1$-th layer weight value.

$$f_j^{(l+1)} = ReLU\left(\sum_i x_i^l w_{ij}^l + b_j^l\right) \tag{3}$$

Now we need to train the CNN. First, we initialize the weight and bias values with random numbers. Second, we calculate error using the ground truth and the output from the initialized CNN. Third, we update the weight and bias values of all the layer of our CNN structure via backpropagation [15]. These processes are repeated until the error is smaller than the maximum error tolerance or exceed the maximum iteration.

3 Result

To evaluate our HAR system, we compared the accuracy of HAR based on HMM, DBN, and CNN using the same MSRC-12 dataset. Table 1 shows the comparison results in terms of recognition accuracy. 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.

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

Activities	HMM (%)	DBN (%)	CNN (%)
G1	87.5	94.3	97.7
G2	98.8	98.8	100
G3	84.2	98.8	97.6
G4	91.8	96.9	94.9
G5	86.1	98.7	100
G6	97.6	100	100
G7	92.9	98.8	98.8
G8	95.1	93.8	100
G9	93.9	98.0	98.0
G10	94.8	96.1	96.1
G11	92.4	98.7	100
G12	94.9	97.4	100
Mean (STD)	92.49 (4.48)	97.54 (1.92)	98.59 (0.69)

Table 2 The confusion matrix of HAR with the proposed CNN-based HAR system

G1	97.7			1.14							1.14	
G2		100										
G3			97.6		2.44							
G4				94.9					2.04		3.06	
G5					100							
G6						100						
G7			1.19				98.8					
G8								100				
G9									98.0		2.02	
G10			1.30							96.1	2.60	
G11											100	
G12												100
	G1	G2	G3	G4	G5	G6	G7	G8	G9	G10	G11	G12

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. The results by CNN show the best recognition accuracy of 98.59% with standard deviation for 0.69. The CNN-based HAR is 6.1% more accurate than those of HMM-based HAR are and 1.05% more accurate than those of DBN-based HAR. Table 2 shows a confusion matrix for 12 activities using CNN.

4 Conclusions

In this paper, we present a work of CNN-based HAR. Our CNN-based HAR results show that CNN outperforms HMM and DBN. The proposed CNN-based HAR system can be adopted as a smart system in smart hospitals for better healthcare of patients and in smart homes for better elderly care of residents.

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

References

1. Nam SB, Park SU, Park JH, Uddin MZ, Kim T-S (2015) Accurate 3D human pose recognition via fusion of depth and motion sensors. Int Conf Comput Commun Devices 4(5):336–340
2. Iengo S, Rossi S, Staffa M, Finzi A (2014) Continuous gesture recognition for flexible human-robot interaction. IEEE Trans ICRA 2014 4863–4868
3. Piyathilaka L, Kodagoda S (2013) Gaussian mixture based HMM for human daily activity recognition using 3d skeleton features. IEEE 8th Conf ICIEA 2013 567–572
4. Jalal A, Sarif N, Kim JT, Kim T-S (2013) Human activity recognition via recognized body parts of human depth silhouettes for residents monitoring services at smart home. Indoor Built Environ 22:271–279
5. Nam SB, Park SU, Park JH, Kim T-S (2015) A single depth sensor based human activity recognition via deep belief network. World Conf Appl Sci 2015 015–019
6. Hinton GE, Osindero S, Tes Y (2006) A fast learning algorithm for deep belief nets. Neural Computation 1527–1554
7. Krizhevsky A, Sutskever I, Hinton GE (2012) ImageNet classification with deep convolutional neural networks. NIPS 2012 1097–1105
8. LeCun Y, Bengio Y (1995) Convolutional networks for images, speech, and time-series. Handb Brain Theory Neural Networks 3361(10):1995
9. Simonyan K, Zisserman A (2014) Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556
10. Szegedy C, Liu W, Jia Y, Sermanet P, Reed S et al (2015) Going deeper with convolutions. In: Proceedings of the IEEE conference on computer vision and pattern recognition 2015, pp 1–9
11. Duffner S, Berlemont S, Lefebvre G, Garcia C (2014) 3D gesture classification with convolutional neural networks. In: ICASSP 2014, pp 5432–5436
12. Fothergill S, Mentis H, Kohli P, Nowozin S (2012) Instructing people for training gestural interactive systems. ACM Conference on Human Factors in Computing Systems 2012, pp. 1737–1746
13. Linde Y, Buzo A, Gray RM (1980) An algorithm for vector quantizer design. IEEE Trans Commun 702–710
14. Minsky M, Papert S (1969) Perceptrons. An introduction to computational geometry. M.I.T Press, Cambridge, vol 165, No. 3895, pp 780–782
15. Hecht-Nielsen R (1989) Theory of the backpropagation neural network, In: IJCNN 1989, pp 593–605