Journal of Electrical Engineering, Vol. 75, No. 1, 2024, pp. 8-13

sciendo

PAPERS

Human activity recognition with fine-tuned CNN-LSTM

Erdal Genc[1], Mustafa Eren Yildirim[2,3], Yucel Batu Salman[4]

Human activity recognition (HAR) by deep learning is a challenging and interesting topic. Although there are robust models, there is also a bunch of parameters and variables, which affect the performance such as the number of layers, pooling type. This study presents a new deep learning architecture that is obtained by fine-tuning of the conventional CNN-LSTM model, namely, CNN (+3)-LSTM. Three changes are made to the conventional model to increase the accuracy. Firstly, kernel size is set to 1×1 to extract more information. Secondly, three convolutional layers are added to the model. Lastly, average pooling is used instead of max-pooling. Performance analysis of the proposed model is conducted on the KTH dataset and implemented on Keras. In addition to the overall accuracy of the proposed model, the contribution of each change is observed individually. Results show that adding layers made the highest contribution followed by kernel size and pooling, respectively. The proposed model is compared with state-of-art and outperformed some of the recent studies with a 94.1% recognition rate.

Category: smart and intelligent computing

Keywords: human activity recognition, convolutional neural network, KTH dataset

1 Introduction

Human activity recognition (HAR) has been one of the most studied fields in today’s era. Several techniques have been applied for video surveillance systems, fall detection analysis, and virtual reality (VR) games. Most of the HAR methods need infrastructural support even if computer vision-based methods are decent to recognize human activity. For instance, the installation of video cameras is required in monitoring applications. One of the efficient approaches for the recognition process is to analyze the data collected by the inertial measurement unit (IMU) sensors which are settled on a person’s body [1, 2]. Moreover, IMU can be installed on a smartphone so that the action motions of a person can be analyzed.

The capture of certain patterns during an activity is significant for understanding that activity. Discovering activity patterns can sometimes relate to finding unknown patterns that come directly from a sensor without describing a predefined model or a presumption. The majority of the existing models for HAR have a high complexity. This leads to high computation time. On the other hand, the models with low complexity cannot supply high accuracy for recognition. Thus, this paper aims to provide high accuracy with low computation time.

This paper proposes a convolutional neural network (CNN) based on deep learning architecture for HAR purposes. We did three modifications to the conventional CNN architecture. First, we added three convolutional layers to the model. Second, kernel size is set to 1×1 for the feature extraction stage. At last, we used average pooling rather than max-pooling. The extracted features are fed into the LSTM module for action recognition. The analysis and results of the proposed model are given in further sections.

2 Related work

In computer vision, vision-based HAR is a major portion of human monitoring applications [3]. There are a vast amount of HAR based publications in the literature [4], including gesture-based interactive games and physical therapy [5], etc.

Instead of using a keyboard and mouse, human gesture movements ideally can be utilized in human-computer based intelligent systems. While hand movements can conduct presentations flow [6] in terms of previous or following pages, employees can also gain experience and learn new methods about manufacturing steps [7].

By focusing on HAR and pose estimation, a system can employ fall detection, as it is a vital issue for disabled and elderly people. An intelligent system that works with only a blinking eye is an example of it [8]. On the other hand, trying to understand human mental capabilities is another study field. It can be done by

[1] Equinor UK Ltd, London, United Kingdom
[2] Department of Electronics and Communications Engineering, American University of Malta
[3] Department of Electrical and Electronics Engineering, Bahçeşehir University, Turkey
[4] Department of Software Engineering, Bahcesehir University, Turkey
erdal.genc09@gmail.com, eren.yildirim@aum.edu.mt, mustafaeren.yildirim@bau.edu.tr, batu.salman@bau.edu.tr

https://doi.org/10.2478/jee-2024-0002, Print (till 2015) ISSN 1335-3632, On-line ISSN 1339-309X

observing the relations and movements of people in a real-world environment. In [9], the authors introduced a study examining young children's process of learning and the development of cognition.

In Balloon Game and Microsoft Xbox [10], the body movements of players can be utilized to conduct a game. Assistance systems of the smart drive are also feasible by checking a driver's body [11]. According to the driver's behavior and posture analysis, awareness of a driver is observed [12]. Besides, the interaction between the driver's head and finger can be compared to check whether the driver is distracted or not [13].

Intelligent systems can also be used in these fields such as parking lots and transportation in public [14]. Video annotation has become significant since there are tons of videos that are recorded, shared, uploaded, and downloaded these days. Therefore, annotating a video of a football game or a broadcast for outdoor sports can be very useful [15].

3 Proposed CNN (+3)-LSTM model

3.1 Convolutional neural networks

Deep learning helps us to understand and identify an unknown image without feature extraction manually. It extracts features and classifies images by self-learning from many input images. CNN is a widely used deep learning scheme for feature extraction.

CNN consists of one or several different convolutional layers. Fully connected layers are sometimes attached to these convolutional layers. The main idea of CNN is to analyze the input that is most likely to be a 2D image. The algorithm also has a loss function in the last layer to calculate the loss and increase the performance.

Visual mechanisms that are used by animals are also used by CNN`s. Hubel and Wiesel specified two cell types in the vision area of the brain in their paper [16]. Simple cells, straight edges maximize the output according to relative positions within their receptive field. In complex cells, there are larger receptive fields and edge positioning does not affect the output.

A CNN consists of a variety of convolutional and subsampling layers, fully connected layers added at the end. The inspiration for using activation functions is coming from biological neurons. It consists of multiple layers such as convolutional, pooling, fully connected layer, and weights of each operation. This method is highly in use in today's deep learning era specifically for human activity recognition.

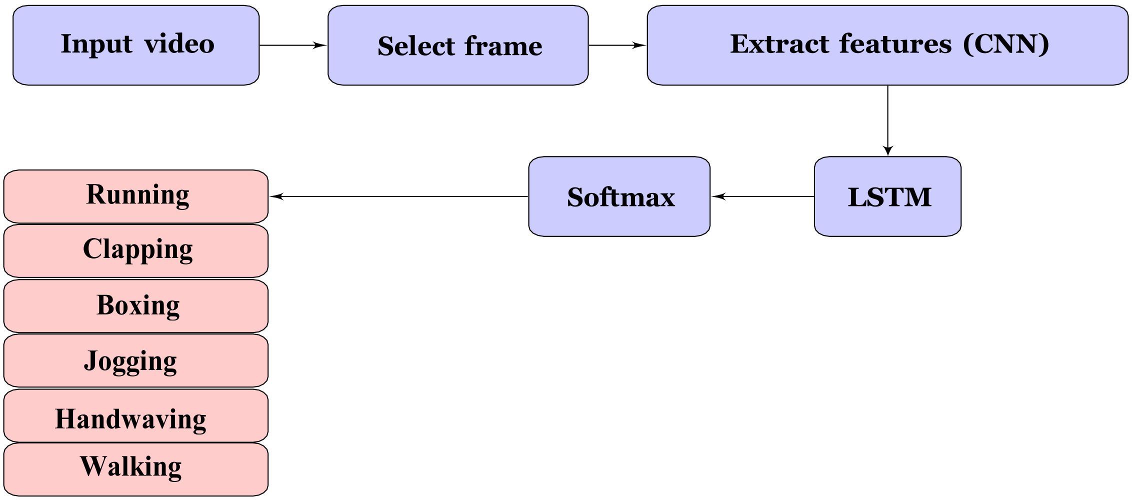

Fig. 1. The flow of the proposed model

3.2 Flow of the proposed model

In this work, a camera-based system is proposed. Features from the dataset are extracted by CNN and output data are transferred to LSTM. Figure 1 gives the flowchart of the proposed model. Captured frames of the dataset with certain activity are given into the proposed CNN model for feature extraction. Output features are given to the LSTM module for classification.

In this study, the proposed method has three contributions: the change in kernel size, convolutional layer, and lastly the pooling type.

3.3 Change in kernel size

In this study, the proposed model is based on the conventional LeNet model [16] for feature extraction. LeNet is one of the first models of convolutional neural networks in literature. It uses 3×3 convolutional kernels. Although 1×1 convolution is considered a feature pooling, or dimension reduction layer, it acts as a coordinate-dependent transformation in the filter space [17]. Although this operation is linear, it is followed by a non-linearity function like ReLU. The transformation then is learned through stochastic optimization (such as stochastic gradient descent (SGD)).

In the proposed method, these kernel sizes are set to 1×1 for all layers. The resolution of the images in the used dataset is 160×120 that is smaller than the default sizes in other deep learning methods. Thus, using the 1×1 kernel treats each pixel as a feature and extracts complex features from the low-resolution image.

3.4 Change in convolutional layers

Adding convolutional layers to a network may increase the overall accuracy. However, if overdone, it may cause vanishing/exploding issues [18], preventing the network from converging to the global minima. Normalization techniques such as batch normalization enable networks to start converging while enabling the addition of more convolutional layers [19].

As deep neural networks are on the verge of convergence, a degradation problem occurs; accuracy saturates, and drops rapidly afterward. This issue is not due to overfitting; moreover, adding more layers to a candidate model causes higher training error [20, 21]. Therefore, balancing the number of convolutional layers and adding more layers is a delicate task.

The degrading training accuracy points to the fact that not all deep learning models are simply optimum. For example, in comparing a shallow architecture and its deeper counterpart, there is a method of construction of the deeper model. Added layers to the model are identity mapping, and other layers are copied from the earlier shallow model. This method points out that a deeper model should produce a lower training error than its shallow counterpart should. However, current solutions are unable to find the optimal answer, or at least in a feasible period.

The conventional model uses only two convolutional layers in the CNN part. To abstract more features from input images, three convolutional layers are added to the architecture. Increasing the number of convolutional layers, just enough to prevent overfitting allows the model to extract more features that in turn can be used as more detailed input for the classification layer.

We explain the method and reason for choosing the layer number in the results section.

3.5 Change in pooling operation

Max-pooling is used to extract important features like edges, whereas the average pooling allows the extraction of more smooth features. Even though both pooling operations are used for the same reason, average pooling takes all the information in its receptive field, sends it down to the next layer which means all pixel values in an image are considered. The average pooling operation is done by taking the average of the values in its receptive field, which differs from max-pooling where only the highest value pixel is used, and the rest is discarded.

4 Experimental results and discussion

4.1 KTH dataset

This video database was published by authors, in 2004 [20]. The same authors also proposed an algorithm for this dataset. The dataset contains 6 human actions in four different locations. There are four different scenarios: outdoors, outdoors with scale variation, outdoors with different clothes, and indoors. It includes 2,391 sequences. A static camera with 25fps records these sequences. We apply downsampling to the dataset and decreased the spatial resolution to 160×120. In short, 600 videos correspond to 25 subjects, 6 actions, and 4 scenarios respectively. In this study, images are extracted from videos and processed by using Python. For all methods, the dataset is divided into 75% is for training, 20% for test and 5% for the validation process.

The extracted images are used as input for benchmark studies and the proposed methods. The training procedure uses 16 randomly selected individuals, out of 25, while the rest are used in the testing phase following the setup in [21].

4.2 Implementation details

The proposed model consists of five convolutional layers and an LSTM block. It uses 1×1 kernel size as well as average pooling. Specifically, batch size and hidden layers are 32. It also uses a softmax classifier. We coded our model in Keras, which uses the Tensorflow backend. The number of epochs is 30 and the learning rate is 0.001. Training roughly took 1 hour 55 minutes. Accuracy and loss fluctuations during training and testing are shown in Fig. 2.

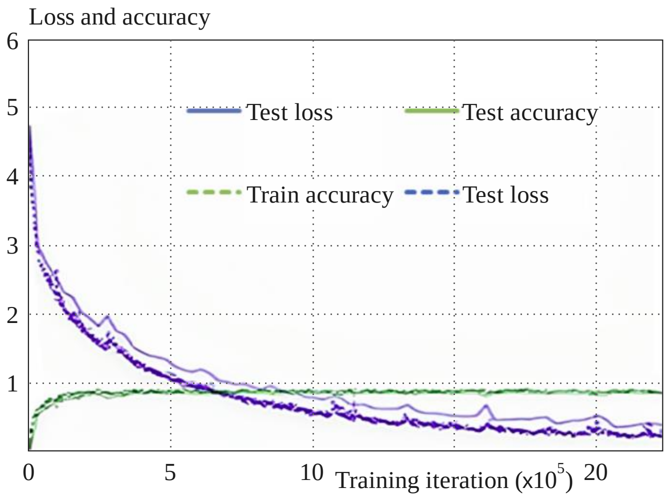

Fig. 2. Training graph of proposed CNN model

4.3 Performance of proposed model

The confusion matrix can be seen in Fig. 3. It perfectly classifies walking among the other similar activities. Besides, it achieved over 90% accuracy for similar activities including running, jogging, and waving-clapping.

	Box	Clap	Wave	Jog	Run	Walk
Box	0.95	0.05	0	0	0	0
Clap	0.07	0.93	0	0	0	0
Wave	0.04	0.04	0.92	0	0	0
Jog	0	0	0	0.91	0.07	0.02
Run	0	0	0	0.06	0.93	0.01
Walk	0	0	0	0	0	1.0

Fig. 3. Confusion matrix of proposed CNN model

The reason why three convolutional layers are added is explained in Fig. 4. In the simulation environment, analysis is done by altering the number of convolution operations at two, five, seven, and nine respectively. In Fig. 4, the highest accuracy rate is obtained with a total of five layers, whereas the conventional CNN model with two layers could achieve a 91.6% rate. When we increased the number of layers, the accuracy started to degrade.

As they all are explained in the previous parts, there are three contributions to this study. These are setting kernel size to 1×1, using average pooling, and adding three more convolutional layers.

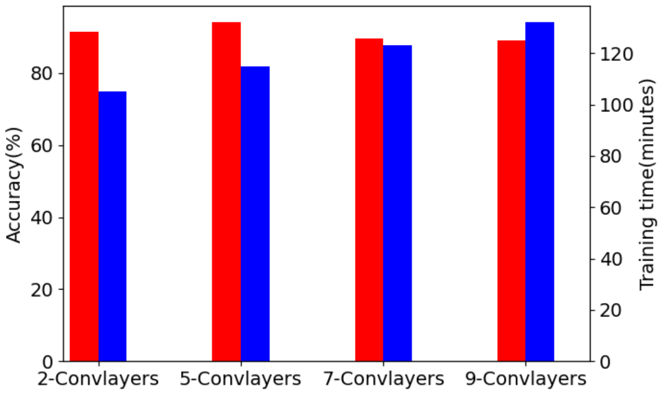

Fig. 4. Recognition accuracy (red) and training time (blue) with different convolutional layers

The other two modifications are unchanged to analyze the impact of each modification. For example, while using a 1×1 kernel, the number of convolutional layers remained unchanged (two convolutional layers) and max-pooling was used instead of average pooling. Others are also analyzed in the same way above. The impacts of each modification are in Table 1.

Table 1. Effect of each modification on the recognition rate

Modification	Accuracy (%)
None	91.6
Kernel size	91.9
Average pooling	91.7
Convolutional layers	92.7
All modifications together	94.1

The accuracy rate of conventional CNN-LSTM is 91.6%. According to Table 1, adding three convolutional layers has the most significant effect on performance. Moreover, average pooling increases the performance by 0.1, whereas kernel size has an impact of 0.3.

Each modification has a small improvement on accuracy when applied separately. On the other hand, when all modifications are applied together, they trigger each other and lead to a much higher accuracy rate.

4.4 Comparison with state-of-art

A comparison is made with studies in the literature on the same dataset to show the performance of the proposed model. The results are given in Table 2.

According to the results, the proposed model showed 94.1% accuracy rate. Although our model could not outperform all the studies in the literature, it showed a performance that is in the state-of-art level.

If we recall from Table 1, addition of three convolutional layers gave 92.7% accuracy, which already outperformed some of the state-of-art models. Moreover, the least effective contributions of this study, which are kernel size and average pooling, also outperformed some recent studies.

Table 2. Comparison with state-of-the-art methodologies on the KTH dataset

Methodology	Accuracy (%)
SMPT [22]	93.40
3D-CNN [23]	94.90
CNN (+3)-LSTM	**94.10**
CNN+LSTM	91.60
Online DL [24]	91.99
MHI [25]	86.70
Deep Representations [26]	98.15
SIFT+BoW+SVM [27]	97.89
Point Detector [28]	90.02
Gaussian mixture [29]	91.97
Large-scaling [30]	91.30
SVD+SVM [31]	96.51
DBN [32]	94.83

5 Conclusion

This paper presents a fine-tuned CNN model for human activity recognition. There are three modifications to increase the performance of recognition. These are adding three more convolution operations, changing the pooling layer, and altering the kernel sizes. The proposed CNN model is a more stable and robust method in human activity recognition.

Nevertheless, there are some limitations of the proposed model. The first one is the used dataset. The videos in the used dataset do not contain any noise such as occlusion, huge illumination variations nor multiple people in the same scene. Thus, use of the proposed model in real world mediums can supply lower accuracy rates. The second one is the computation time. Although the accuracy improved, there is a slight increase in the training time also. This must be decreased by a faster hardware.

For future work, a more complex dataset can be used. In addition, training time can be decreased by altering the algorithm to a simpler structure. Moreover, computational cost can be minimized to accomplish a smoother model where a fewer number of parameters are needed.

References

[1] P. Casale, O. Pujol, and P. Radeva, "Human activity recognition from accelerometer data using a wearable device," in *Proceedings of Iberian Conference on Pattern Recognition and Image Analysis*, Spain, 2011, pp. 289-296.

[2] N. C. Krishnan, D. Colbry, C. Juillard, and S. Panchanathan, "Real-time human activity recognition using tri-axial accelerometers," in *Proceedings of Sensors, Signals and Information Processing Workshop*, Sedona, 2008.

[3] A. H. Moeslund and V. Kruger, “A survey of advances in vision-based human motion capture and analysis”, *Computer Vision and Image Understanding*, vol. 104, no. 2-3, pp. 90-126, Dec. 2006.

[4] M. B. Holte, "Vision-Based 2D and 3D Human Activity Recognition," Ph.D. dissertation, Aalborg University, Aalborg, Denmark, 2012.

[5] Kinect Physical Therapy, http://x-tech.am/kinect-physical-therapy.

[6] H. Lee and J. H. Kim, "An hmm-based threshold model approach for gesture recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 21, no. 10, pp. 961-973, Oct. 1999.

[7] A. B. Postawa, M. Kleinsorge, J. Krueger, and G. Seliger, "Automated image based recognition of manual work steps in the remanufacturing of alternators," in *Proceedings of the Conference on Sustainable Manufacturing*, Berlin, 2011, pp. 209-214.

[8] A. A. Alonso, R. D Rosa, L. D. Val, M. I. Jimenez, and S. Franco, "A robot controlled by blinking for ambient assisted living," in *Proceedings of the Int. Work-Conference on Artificial Neural Networks: Part II: Distributed Computing, Artificial Intelligence, Bioinformatics, Soft Computing, and Ambient Assisted Living*, Salamanca, 2009, pp. 839-842.

[9] Y. Chen, L. B. Smith, S. Hongwei, A. F. Pereira, and T. Smith, “Active information selection: Visual attention through the hands,” *IEEE Transactions on Autonomous Mental Development*, vol. 1, no. 2, pp. 141-151, Sep. 2009.

[10] C. Tran and M. M. Trivedi, "Introducing XMOB: Extremity Movement Observation Framework for Upper Body Pose Tracking in 3D," in *Proceedings of IEEE Int. Symposium on Multimedia*, San Diego, 2009, pp. 446-447.

[11] M. M. Trivedi and S. Y. Cheng, “Holistic sensing and active displays for intelligent driver support systems,” vol. 40, no. 5, *IEEE Computer Magazine*, pp. 60-68, May 2007.

[12] E. M. Chutorian and M. M. Trivedi, "Head pose estimation and augmented reality tracking: An integrated system and evaluation for monitoring driver awareness," *IEEE Transactions on Intelligent Transportation Systems*, vol. 11, no. 2, June 2010.

[13] C. Tran and M. M. Trivedi, “Driver assistance for 'Keeping hands on the wheel and eyes on the road’,” in *Proceedings of the IEEE Int. Conf. on Vehicular Electronics and Safety*, Pune, 2009, pp. 97-101.

[14] S. Park and M. M. Trivedi, "Understanding Human Interactions with Track and Body Synergies (TBS) Captured from Multiple Views," *Computer Vision and Image Understanding*, vol. 111, no. 1, pp. 2-20, July 2008.

[15] J. Assfalg, M. Bertini, C. Colombo, A. D. Bimbo, and W. Nunziati, "Semantic annotation of soccer videos: automatic highlights identification," *Computer Vision and Image Understanding*, vol. 92, no. 2-3, pp. 285-305, Nov-Dec. 2003.

[16] D. H. Hubel and T. N. Wiesel, "Receptive Fields and Functional Architecture of Monkey Striate Cortex,"*The Journal of Physiology*, vol. 195, no. 1, pp. 215-243, March 1968.

[17] M. Lin, Q. Chen, and S. Yan, "Network in network," arXiv:1312.4400, 2013.

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

[19] S. Ioffe and C. Szegedy, "Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift," in *Proceedings of Int. Conf. on Machine Learning*, Lille, 2015, pp.448-456.

[20] C. Schuldt, I. Laptev, and B. Caputo, "Recognizing human actions: a local SVM approach," in *Proceedings of Int. Conf. on Pattern Recognition*, 2004, pp. 32-36.

[21] H. Jhuang, T. Serre, L. Wolf, T. Poggio, "A biologically inspired system for action recognition," in Proceedings of *Int. Conf. on Computer Vision*, 2007, Rio de Janeiro.

[22] Z. Lin, Z. Jiang and L. S. Davis, "Recognizing actions by shape-motion prototype trees," in Proceedings of *Int. Conf. on Computer Vision Workshops*, Kyoto, 2009, pp. 444-451.

[23] J. Arunnehru, G. Chamundeeswari and S. P. Bharathi, "Human Action Recognition using 3D Convolutional Neural Networks with 3D Motion Cuboids in Surveillance Videos," *Procedia Computer Science*, vol.133, pp. 471-477, 2018.

[24] K. Charalampous and A. Gasteratos, "On-line deep learning method for action recognition," *Pattern Analysis & Applications*, vol. 19, no. 2, pp. 337-354, May 2016.

[25] M. A. R. Ahad, M. N. Islam, and I. Jahan, "Action recognition based on binary patterns of action-history and histogram of oriented gradient," *Journal on Multimodal User Interfaces*, vol. 10, pp. 335-344, Dec. 2016.

[26] A. B. Sargano, X. Wang, P. Angelov and Z. Habib, "Human action recognition using transfer learning with deep representations," *in Proceedings of Int. Joint Conf. on Neural Networks*, Anchorage, 2017, pp. 463-469.

[27] M. M. Moussa, E. Hamayed, M. B. Fayek, H. A. El Nemr, "An enhanced method for human action recognition," *Journal of Advanced Research*, vol. 6, no. 2, pp. 163-169, Mar. 2015.

[28] D. Zahraa, A. Amel, "Human action recognition using interest point detector with KTH dataset," *International Journal of Civil Engineering and Technology,* vol.4, no.10, pp.333-34, 2019.

[29] F. Najar, S. Bourouis, N. Bouguila, S. Belghith, "Unsupervised learning of finite full covariance multivariate generalized Gaussian mixture models for human activity recognition," *Multimed Tools Appl* 78, 18669-18691 (2019).

[30] P. Antonik, N. Marsal, D. Brunner, D. Rontani, "Human action recognition with a large-scale brain-inspired photonic computer," *Nat Mach Intell* 1, 530–537 (2019).

[31] Yoon, Byung Woo, et al., "Human activity recognition using inter-joint feature fusion with SVD," *ICIC Express Letters, Part B: Applications*, vol.12, no.3, pp. 215-221, 2021.

[32] A. Mehrez and A. Douik, "Human Action Recognition in Video Sequences Using Deep Belief Networks," *Traitement du Signal,* vo. 37, no. 1, pp. 37-44, 2020.

Received 8 November 2023