Leveraging CNN and Transfer Learning for Vision-based Human Activity Recognition

Samundra Deep
Department of Computing
Macquarie University
Sydney, Australia
Email: samundra.deep@hdr.mq.edu.au

Xi Zheng
Department of Computing
Macquarie University
Sydney, Australia
Email: james.zheng@mq.edu.au

***Abstract*—With the advent of the Internet of Things (IoT), there have been significant advancements in the area of human activity recognition (HAR) in recent years. HAR is applicable to wider application such as elderly care, anomalous behaviour detection and surveillance system. Several machine learning algorithms have been employed to predict the activities performed by the human in an environment. However, traditional machine learning approaches have been outperformed by feature engineering methods which can select an optimal set of features. On the contrary, it is known that deep learning models such as Convolutional Neural Networks (CNN) can extract features and reduce the computational cost automatically. In this paper, we use CNN model to predict human activities from Wiezmann Dataset. Specifically, we employ *transfer learning* to get deep image features and trained machine learning classifiers. Our experimental results showed the accuracy of 96.95% using VGG-16. Our experimental results also confirmed the high performance of VGG-16 as compared to rest of the applied CNN models.**

***Index Terms*—Activity recognition, deep learning, convolutional neural network.**

I. Introduction

Human activity recognition (HAR) is an active research area because of its applications in elderly care, automated homes and surveillance system. Several studies has been done on human activity recognition in the past. Some of the existing work are either wearable based [1] or non-wearable based [2] [3]. Wearable based HAR system make use of wearable sensors that are attached on the human body. Wearable based HAR system are intrusive in nature. Non-wearable based HAR system do not require any sensors to attach on the human or to carry any device for activity recognition. Non-wearable based approach can be further categorised into sensor based [2] and vision-based HAR systems [3]. Sensor based technology use RF signals from sensors, such as RFID, PIR sensors and Wi-Fi signals to detect human activities. Vision based technology use videos, image frames from depth cameras or IR cameras to classify human activities. Sensor based HAR system are non-intrusive in nature but may not provide high accuracy. Therefore, vision-based human activity recognition system has gained significant interest in the present time. Recognising human activities from the streaming video is challenging.

Video-based human activity recognition can be categorised as marker-based and vision-based according to motion features [4]. Marker-based method make use of optic wearable marker-based motion capture (MoCap) framework. It can accurately capture complex human motions but this approach has some disadvantages. It require the optical sensors to be attached on the human and also demand the need of multiple camera settings. Whereas, the vision based method make use of RGB or depth image. It does not require the user to carry any devices or to attach any sensors on the human. Therefore, this methodology is getting more consideration nowadays, consequently making the HAR framework simple and easy to be deployed in many applications.

Most of the vision-based HAR systems proposed in the literature used traditional machine learning algorithms for activity recognition. However, traditional machine learning methods have been outperformed by deep learning methods in recent time [5]. The most common type of deep learning method is Convolutional Neural Network (CNN). CNN are largely applied in areas related to computer vision. It consists series of convolution layers through which images are passed for processing. In this paper, we use CNN to recognise human activities from Wiezmann Dataset. We first extracted the frames for each activities from the videos. Specifically, we use *transfer learning* to get deep image features and trained machine learning classifiers. We applied 3 different CNN models to classify activities and compared our results with the existing works on the same dataset. In summary, the main contributions of our work are as follows:

1) We applied three different CNN models to classify human recognition activities and we showed the accuracy of 96.95% using VGG-16.

2) We used *transfer learning* to leverage the knowledge gained from large-scale dataset such as ImageNet [6] to the human activity recognition dataset.

The rest of the paper is as follows: Section II provide an overview of the related work in video-based HAR systems. We provide an overview of transfer learning in section III. Section IV outline the research methodology, sources of data, research approach and discuss the experimental results. Conclusion and future work are drawn in Section V.

978-1-7281-3673-8/19/$31.00 ©2019 IEEE

II. Related Work

There have been a lot of research on vision-based human activity recognition in recent years. Most of the studied methods have depend on handcrafted feature extraction from the videos/images and employed traditional classifiers for activity recognition. The traditional approaches often achieved optimum results and exhibited high performances. However, traditional methods are not feasible to deploy in real life because handcrafted features are highly dependent on data and are not robust to the environment change.

Hidden Markov Model (HMMs) methods have been largely used as the recognition techniques in the past because of its capability of temporal pattern decoding [7]. However, researchers are more interested in using deep learning techniques because of its ability to automatically extract the features and learn deep pattern structures [5] [7]. Deep learning methods have clearly ruled out traditional classification methods in the domain of computer vision [5] [8]. Deep learning techniques have been largely employed recently in the domain of computer vision and have achieved tremendous results. Therefore, video-based human activity recognition using deep learning models have gained a lot of interest in recent years [5].

Zhu et al. [4] proposed an action classification method by adding a mixed-norm regularization function to a deep LSTM network. One of the most popular deep learning methods in frames/image processing is Convolutional Neural Network (CNN). There have been several works that utilized 2D-CNNs that take advantages of spatial correlation between the video frames and combine the outputs employing different strategies [9] Many have also used additional input such as optical flow to 2D-CNN to get temporal correlations information [10]. Subsequently, 3D-CNNs [11] were introduced that demonstrated exceptional results in the classification of videos and frames.

Wang et al. [12] applied CNN to RGB and depth frames to automatically extract the features. The obtained features were passed through a fully connected neural network and achieved an improved accuracy. Ji et al. [13] proposed a 3D CNN model which performs 3D convolutions and extract spatial and temporal features by capturing the motion information for activity recognition. Simonyan et al. [8] introduced ConvNet, a two-stream convolution layer architecture that could achieve good results despite of limited training data. Khaire et al. [14] proposed a model that train convnets from RGB-D dataset and combined the softmax scores from depth, motion and skeleton images at the classification level to identify the activities. Karpathy et al. [15] proposed the extension of CNN architecture in the first convolutional layers over a 4D video chunk. Similarly, Tran et al. [16] used a deep 3D CNN architecture (quiet similar to VGGnet [17]) that utilise spatiotemporal convolutions and pooling in all layers to improve the accuracy of the model.

In comparison, we are more interested to explore how *transfer learning* can be leveraged with CNN models on benchmark dataset to improve classification accuracy.

III. Transfer Learning

Transfer learning [18] is a method of transferring *knowledge* that a model has learned from earlier extensive training to the current model. The deep network models can be trained with significantly less data with transfer learning. It has been used to reduce training time and improve accuracy of the model. In this work, we use *transfer learning* to leverage the knowledge gained from large-scale dataset such as ImageNet. We first extract the frames for each activities from the videos. We use *transfer learning* to get deep image features and trained machine learning classifiers. For all CNN models, pre-trained weights on ImageNet are used as starting point for transfer learning. ImageNet [6] is a dataset containing 20000 categories of activities. The knowledge is transferred from pre-trained weights on ImageNet to Weizmann dataset, since set of activities recognised in this work fall within the domain of ImageNet. The features are extracted from the penultimate layer of CNNs. The basic idea of transfer learning is as shown in Figure 1.

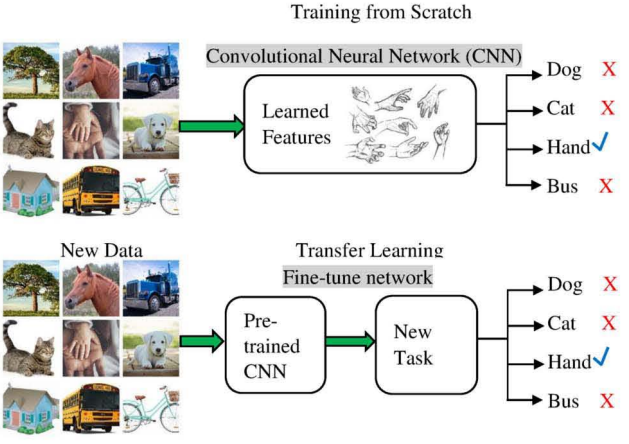

Fig. 1. Schematic diagram to demonstrate transfer learning.

The main approaches in transfer learning are: (1). to preserve the original pre-trained neural model of large-scale dataset and update weights of the trained model on the target dataset, and (2). use the pre-trained neural model for feature extraction and representation followed by a generic classifier such as Support Vector Machine or Logistic Regression.

IV. Implementation

A. Dataset

In order to evaluate the effectiveness of the models, we do experiments on benchmark activity recognition, namely **Weizmann dataset**. It consists of 90 low-resolution video sequences showing 9 different people performing 10 activities namely, *bend*, *jack* (or jumping-jack), *jump* (or jump-forward-on-two-legs), *pjump* (or jump-in-place-on-two-legs), *run*, *side* (or gallopsideways), *skip*, *walk*, *wave1* (waveone-hand), and *wave2* (wave-two-hands).We used nine actions (not including *pjump* jump-in-place-on-two-legs) for our experiment. We first convert all videos into individual frames based on their

activity. Table I shows the total number of frames per activity based on the extracting frames for all 9 people. The entire dataset is divided into *Training* (70%), *Validation*(10%), and *Testing* (20%).

TABLE I
DATASET STATISTICS IN TERMS OF NUMBER OF FRAMES PER ACTIVITY

Activity	Number of Frames
Bend	639
Jack	729
Jump	538
Run	346
Side	444
Skip	378
Walk	566
Wave1	653
Wave2	624
Total	4917

B. Discussion and Results

In order to classify activities, we experiment with 3 different Convolutional Neural Networks (CNN) for activity recognition, namely *VGG-16*, *VGG-19* and Google's *InceptionNet-v3*. We used *transfer learning* to leverage the knowledge gained from large-scale dataset such as ImageNet. Transfer learning technique transfer knowledge from pre-trained model to train a new domain in neural network. We performed experiment on Weizmann dataset using the knowledge learned from pre-trained weights on ImageNet. The features are extracted from the penultimate layers of CNNs. We applied transfer learning on VGG-16 CNN model and achieved accuracy of 96.95% . For VGG-16, image of dimensions 224×224 is given as an input and features from *fc1* layer are extracted which gives 4096-dimensional vector for each image.

We also applied transfer learning to other CNN models such as VGG-19 and Google's InceptionNet-v3 to examine the performance of the different CNN models. VGG-19 and Google's InceptionNet-v3 achieved 96.54% and 95.63% respectively. Experimental results showed that VGG-16 performs better than the rest of the CNN models after all the models have been applied transfer learning. Table II reports accuracy score, precision, recall, and f1-score of the applied CNN models. The confusion matrix of 3 different CNN models are as shown in Figure 2, 3 and 4.

We compared the results achieved from some of the other approaches that do not employed transfer learning on Weizmann dataset with our approach. The experiment results showed that applying transfer learning to the same dataset achieved better recognition scores. The recognition accuracy is improved by $1 - 6\%$ by applying transfer learning. The comparison of results utilising transfer learning on VGG-16 model and other approaches are presented in Table III. The comparison with state-of-the-art approaches is done to explore how effective is transfer learning when leveraged with CNN models for improving recognition scores.

TABLE II
RESULTS ON ACTIVITY RECOGNITION BASED ON DIFFERENT CNN MODELS IN TERMS OF ACCURACY SCORE, PRECISION, RECALL, AND F1-SCORE

Model	Accuracy (in %)	Precision (in %)	Recall (in %)	F1-score (in %)
VGG-16	96.95	97.00	97.00	97.00
VGG-19	96.54	97.00	97.00	96.00
Inception-v3	95.63	96.00	96.00	96.00

TABLE III
PERFORMANCE COMPARISON USING WEIZMANN DATASET

Model	Accuracy (in %)
VGG-16	96.95
Cai et al. [19]	95.70
Kumar et al. [20]	95.69
Feng et al. [21]	94.10
Han et al. [22]	90.00

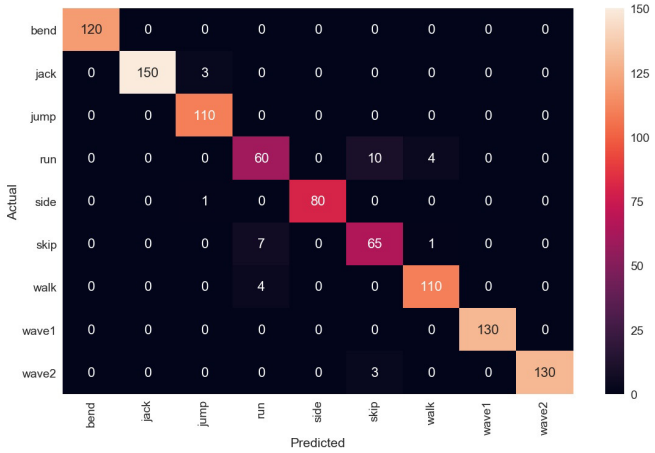

Fig. 2. Confusion Matrix for recognising 9 activities on Weizmann Dataset using VGG-16 Convolutional Neural Network

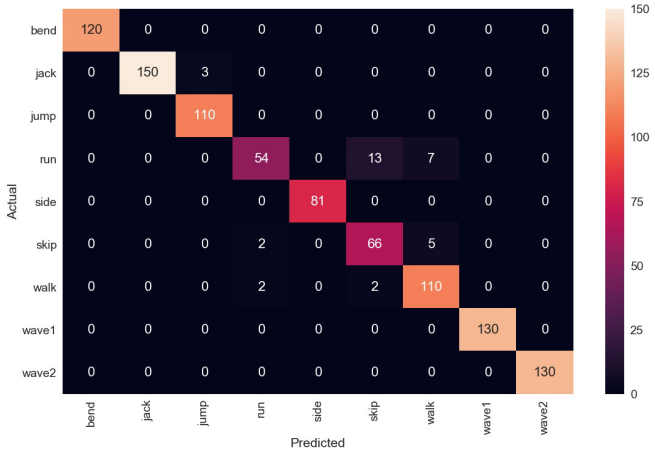

Fig. 3. Confusion Matrix for recognising 9 activities on Weizmann Dataset using VGG-19 Convolutional Neural Network

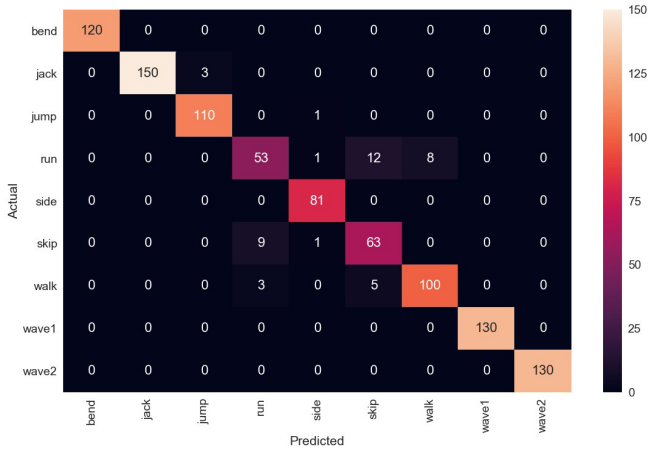

Fig. 4. Confusion Matrix for recognising 9 activities on Weizmann Dataset using Inception-v3 Convolutional Neural Network

Figure 2, 3 and 4 show the confusion matrix of 3 different Convolutional Neural Networks (CNN) after applied transfer learning, which were used to classify frames of different activities using *VGG-16*, *VGG-19* and Google's *InceptionNet-v3* respectively. It is evident from Figure 2, 3 and 4 that VGG-16 has mis-classification in predicting *run* activity as *skip*, VGG-19 has mis-classification in predicting *run* activity as *skip*, and *skip* as *walk*, and Google's InceptionNet-v3 has mis-classification in predicting *run* activity as *skip* which are very similar in terms of their visual perception. Employing transfer learning on the CNN models have increased the accuracy of the activity recognition. However, transfer learning technique used in our work with the knowledge transferred from pre-trained weight on Imagenet may be compromised, since Imagenet contains images of several different categories.

V. Conclusion

We used CNN models to predict the human activities from Wiezmann Dataset. We experimented with 3 different Convolutional Neural Networks (CNN) for activity recognition. We have employed *transfer learning* to get the deep image features and trained machine learning classifiers. Our experimental results showed the accuracy of 96.95% using VGG-16 with the implementation of transfer learning. Our experimental results showed that VGG-16 outperformed other CNN models in terms of feature extraction. Our experimental results with transfer learning technique also showed high performance of VGG-16 as compared to state-of-the-art methods.

In future, we aim to extend this study by developing the context-aware recognition system to classify human activities. Also, we will extend our work to recognise complex human activities such as cooking, reading books, and watching TV.

References

[1] B. Bhandari, J. Lu, X. Zheng, S. Rajasegarar, and C. Karmakar, "Non-invasive sensor based automated smoking activity detection," in *Proceedings of Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, 2017, pp. 845–848.

[2] L. Yao, Q. Z. Sheng, X. Li, T. Gu, M. Tan, X. Wang, S. Wang, and W. Ruan, "Compressive representation for device-free activity recognition with passive rfid signal strength," *IEEE Transactions on Mobile Computing*, vol. 17, no. 2, pp. 293–306, 2018.

[3] I. Lillo, J. C. Niebles, and A. Soto, "Sparse composition of body poses and atomic actions for human activity recognition in rgb-d videos," *Image and Vision Computing*, vol. 59, pp. 63–75, 2017.

[4] W. Zhu, C. Lan, J. Xing, W. Zeng, Y. Li, L. Shen, and X. Xie, "Co-occurrence feature learning for skeleton based action recognition using regularized deep lstm networks," in *Thirtieth AAAI Conference on Artificial Intelligence*, 2016.

[5] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1–9.

[6] J. Deng, W. Dong, R. Socher, L. Li, Kai Li, and Li Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, June 2009, pp. 248–255.

[7] A. Jalal, N. Sarif, J. T. Kim, and T.-S. Kim, "Human activity recognition via recognized body parts of human depth silhouettes for residents monitoring services at smart home," *Indoor and built environment*, vol. 22, no. 1, pp. 271–279, 2013.

[8] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," in *Advances in neural information processing systems*, 2014, pp. 568–576.

[9] G. Gkioxari, R. Girshick, and J. Malik, "Contextual action recognition with r* cnn," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1080–1088.

[10] L. Wang, Y. Xiong, Z. Wang, and Y. Qiao, "Towards good practices for very deep two-stream convnets," *arXiv preprint arXiv:1507.02159*, 2015.

[11] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Deep end2end voxel2voxel prediction," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2016, pp. 17–24.

[12] P. Wang, W. Li, J. Wan, P. Ogunbona, and X. Liu, "Cooperative training of deep aggregation networks for rgb-d action recognition," in *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[13] S. Ji, W. Xu, M. Yang, and K. Yu, "3d convolutional neural networks for human action recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 1, pp. 221–231, 2013.

[14] P. Khaire, P. Kumar, and J. Imran, "Combining cnn streams of rgb-d and skeletal data for human activity recognition," *Pattern Recognition Letters*, vol. 115, pp. 107–116, 2018.

[15] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei, "Large-scale video classification with convolutional neural networks," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2014, pp. 1725–1732.

[16] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3d convolutional networks," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 4489–4497.

[17] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[18] Z. Wharton, E. Thomas, B. Debnath, and A. Behera, "A vision-based transfer learning approach for recognizing behavioral symptoms in people with dementia," in *2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*. IEEE, 2018, pp. 1–6.

[19] J. Cai, X. Tang, and R. Zhong, "Silhouettes based human action recognition by procrustes analysis and fisher vector encoding," in *International Conference on Image and Video Processing, and Artificial Intelligence*, vol. 10836. International Society for Optics and Photonics, 2018, p. 1083612.

[20] S. S. Kumar and M. John, "Human activity recognition using optical flow based feature set," in *Proceedings of IEEE International Carnahan conference on security technology (ICCST)*. IEEE, 2016, pp. 1–5.

[21] W. Feng, H. Tian, and Y. Xiao, "Research on temporal structure for action recognition," in *Chinese Conference on Biometric Recognition*. Springer, 2017, pp. 625–632.

[22] P. Y. Han, K. E. Yee, and O. S. Yin, "Localized temporal representation in human action recognition," in *Proceedings of International Conference on Network, Communication and Computing*. ACM, 2018, pp. 261–266.