Human activity recognition using deep learning approaches

Bradley J. Pillay [1] [0000-0002-0369-4680], Anban Pillay[1] [0000-0001-7160-6972], and Edgar Jembere[1] [0000-0003-1776-1925]

[1] School of Mathematics, Statistics and Computer Science, University of KwaZulu-Natal, Westville Campus, Private Bag X54001, Durban 4000, South Africa
215031687@stu.ukzn.ac.za

Abstract. Human activity recognition using video data has been an active research area in computer vision for many years. This research presents an architecture that employs deep learning techniques to effectively solve human activity recognition of single individuals using visual information from videos. The architecture adopts an Octave Convolutional neural network as a feature extractor and a weighting strategy that regresses the importance of each video segment. This architecture was trained and evaluated on the KTH human activity dataset. The results are promising and validates the approach taken.

Keywords: human activity recognition (HAR), octave convolution neural network, temporal segment network, deep adaptive temporal pooling

1 Introduction

Human activity recognition is the process of identifying the actions and goals of individuals from a series of observations of activities performed within a given environment. This recognition process can be applied to many areas that have the goal of monitoring human actions such as surveillance systems for detecting illegal activities or sporting arenas for detecting foul play. This recognition task is solvable using either video information, sensor data, or both.

The application of deep learning has shown tremendous performance improvements over traditional techniques. The use of deep learning has an advantage through its automatic feature extraction abilities [1]. In this paper, a deep learning architecture that can successfully classify human activities is presented. The performance of the architecture on classifying activities of the KTH human activity dataset is also given.

2 Proposed Architecture Design

The proposed architecture utilizes a temporal segment network (TSN) as the base architecture. A video V is divided into *N* segments of equal lengths. A frame from the middle of each video segment serves as input for the spatial stream. The motion representations of each segment were also extracted using the TVL1 optical flow algo-

Copyright © 2019 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0)

CEUR-WS.org/Vol-2540/FAIR2019_paper_56.pdf

rithm on a stack of 10 consecutive frames. An OctResNet50 model pre-trained on the CIFAR-100 dataset was used as the frame-level feature extractor. The OctResNet50 model is a ResNet50 model that utilizes the octave convolution [4]. The temporal features are stacked together and parsed to a deep adaptive temporal pooling (DATP) module to generate the importance of each temporal segment [2]. The two-component Gaussian Mixture Model (GMM) was chosen as the weights-generator employed by the DATP module. The generated weights were assigned to the temporal segments and a late fusion strategy was adopted for fusing both the spatial and temporal streams. The fused vector was then parsed to a feedforward neural network that outputs the predicted class label for the entire video. All trainable parameters in this architecture was trained through backpropagation.

3 Results

Table 1 displays the confusion matrix obtained by the proposed architecture on the test set made up of 121 videos from the KTH dataset [5]. The accuracy rate obtained by the proposed architecture is 63.6%. The state-of-the-art results of 95.33% classification accuracy were achieved in [3].

Table 1. Confusion Matrix obtained by the proposed architecture on the KTH dataset

Actual Class Labels	Predicted Class Labels					
	Boxing	Handclapping	Handwaving	Jogging	Running	Walking
Boxing	**18**	0	1	0	1	0
Handclapping	1	**18**	1	0	1	0
Handwaving	1	2	**17**	0	0	0
Jogging	1	1	0	**5**	10	4
Running	0	1	0	1	**14**	4
Walking	0	0	0	5	10	**5**

4 Conclusion

Most of the boxing, handclapping and handwaving videos were classified correctly. The jogging and walking actions were misclassified as running. This misclassification may have occurred because the motion representations of the 3 actions were similar. The results obtained are not as good as the state-of-the-art results, but they are promising and validates the approach taken.

References

1. Khurana, R. and Kushwaha, A.K.S., 2019. Delving Deeper with Dual-Stream CNN for Activity Recognition. In Recent Trends in Communication, Computing, and Electronics (pp. 333-342). Springer, Singapore.
2. Song, S., Cheung, N.M., Chandrasekhar, V. and Mandal, B., 2018. Deep Adaptive Temporal Pooling for Activity Recognition. arXiv preprint arXiv:1808.07272.

3. Arunnehru, J., Chamundeeswari, G. and Bharathi, S.P., 2018. Human action recognition using 3D convolutional neural networks with 3D motion cuboids in surveillance videos. Procedia computer science, 133, pp.471-477.
4. Chen, Y., Fang, H., Xu, B., Yan, Z., Kalantidis, Y., Rohrbach, M., Yan, S. and Feng, J., 2019. Drop an Octave: Reducing Spatial Redundancy in Convolutional Neural Networks with Octave Convolution. arXiv preprint arXiv:1904.05049.
5. Nada.kth.se. (2019). Recognition of human actions. [online] Available at: http://www.nada.kth.se/cvap/actions/ [Accessed 14 Aug. 2019]