Multi-channel convolutional neural networks for human action recognition using RGB and depth images

Ch.Raghava Prasad

Department of Electronics and Communication Engineering, Koneru Lakshmaiah Education Foundation (KLEF), Deemed to be University, Vaddeswaram, Green fields, Guntur, Andhra Pradesh, India

chrp@kluniversity.in,

DOI : 10.48047/IJFANS/V11/ISS8/350

Abstract

One of the most researched fields over the past decade has been Human Action Recognition (HAR). This is because there is no single model that can adequately reflect the wide variety of difficulties presented by the mechanics of the human body. As a result, many techniques, applicable to a wide range of data, have been created. The difficulty in accurately representing data is linked to the recognition difficulty. This issue was gradually fixed by increasing the dimensionality of the data, which led to inexpensive 3D RGB-D data in place of 2D RGB films. The purpose of this paper was to develop a 4-stream convolutional neural network (CNN) for use with RGB and depth data. Each of these 4 CNNs is fed this multi modal data as spatial and motion data. Optical flow maps are used to display the RGB and Depth motion data. The network was created to be completely re-programmable.

1.Introduction

When it comes to motion, human action is unparalleled in its adaptability. Human body dynamics are very malleable and diverse, making each individual a special case even within groups with comparable characteristics. Classifying human actions in the digital realm becomes more complicated by the increasing diversity within classes. The term "Human Action Recognition" (HAR) [1] refers to a class of algorithms used to organize information gathered from sensors in various locations that document human actions. Anything done by a human body has a purpose in this world, whether it's communicating emotion or accomplishing a goal. Surprisingly, HAR has several important uses, including smart surveillance [2, 3], person re-identification [4, 5], gait analysis [5, 6], sports action investigation [6, 7], fall detection [8, 9], and crowd sensing [10, 11]. Multi-stream convolutional neural networks (CNNs) can adapt to new data variances and boost recognition accuracy. Individual streams' SoftMax scores are combined to form a global probability class distribution. The deep architectures that have been created are fully trainable. It is well-known that increasing the training sample sizes leads to better performance over a wider spectrum of human actions. The optimum method for carrying out the aforementioned procedure is known as data augmentation. Typically, data is transformed by scaling, rotation, and other methods to perform augmen tation. As a result, the networks can learn to adapt to new types of data by analyzing more examples from each class. The better deep learning models perform, the more data should be used to train them. However, HAR discovers the answer in color, depth, and skeleton data from a low-priced multi-camera array called Microsoft Kinect. While recording, three distinct data sets, or modalities, are produced. The use of Kinect data to train Deep Learning Models for HAR purposes has increased recently. Prior to our study, most multi-stream CNNs had between two and three streams, each of which contained a separate color channel, depth map, and flow map of the same color. These models have proven to be capable of providing respectable accuracy for the aforementioned file types.

2.Analysis of global spatial features

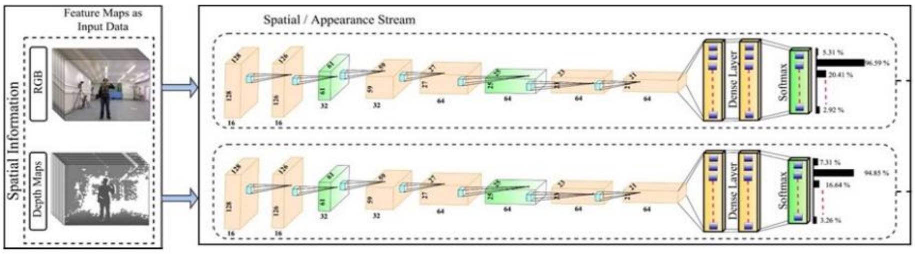

Fig. 1: Spatial Stream CNN model

As shown in Figure 1, the network is composed of the following: six convolutional layers, three maximum pooling layers, two dense layers, and one SoftMax layer. There are 12 segments in C-RGB and 12 in CDPT. Convolutional layers utilize a fixed activation function represented by rectified linear units (ReLu) and a dense activation function represented by tanh. This is because the activation function of the SoftMax layer is sigmoid.

3. Extracting characteristics from motion

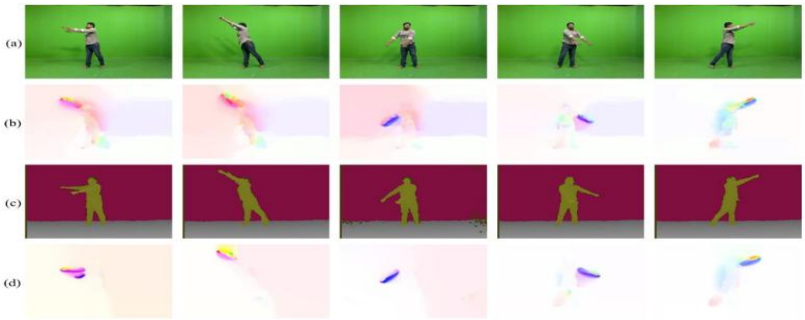

Fig. 2: Optical Flow Maps [142] calculated from RGB and Depth video sequences for action 'Diagonal Stretching'. (a) RGB video sequences. (b) Optical flow maps of (a). (c) Depth sequences and (d) Depth optical flow maps.

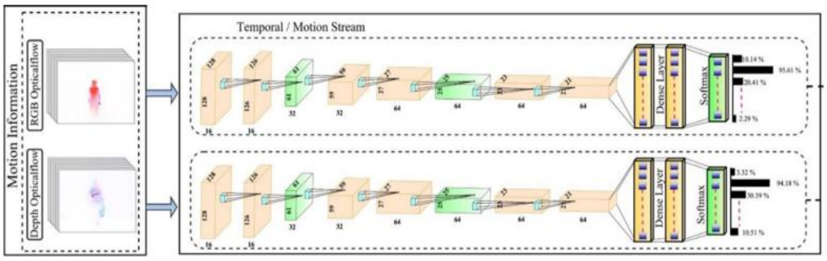

Fig. 3: Motion Stream CNN Model.

In Figure 2, we can see the optical flow maps of an RGB and depth action sequence and how they're distributed across a video frame. High motion content is depicted in the flow maps by dark colors, whereas low motion is shown by bright colors. Each each pixel in the motion frame has a value between 0 and 255; 255 indicates no motion while [0,254] indicates variations in the motion. Colors (from white to light blue to orange to dark blue to purple to green and red) are assigned to each of these groups of motion features in the order specified. Figure 3 depicts the two independent CNN pipelines that were used to learn the RGB and depth motion maps.

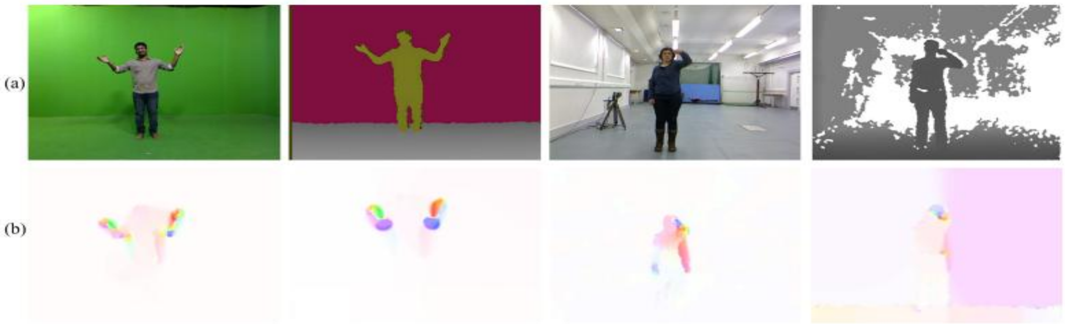

Fig. 4: Motion maps on two different datasets.

Figure 4 displays the computed motion maps for RGB and Depth sequences from multiple action datasets. The impacts of depth motion maps are examined in detail in the work's outcomes section. In the following part, we use these four inputs to build together the full architecture for action recognition. Conclusions drawn from this analysis motivated the development of the four-stream deep learning model shown in Figure 5.

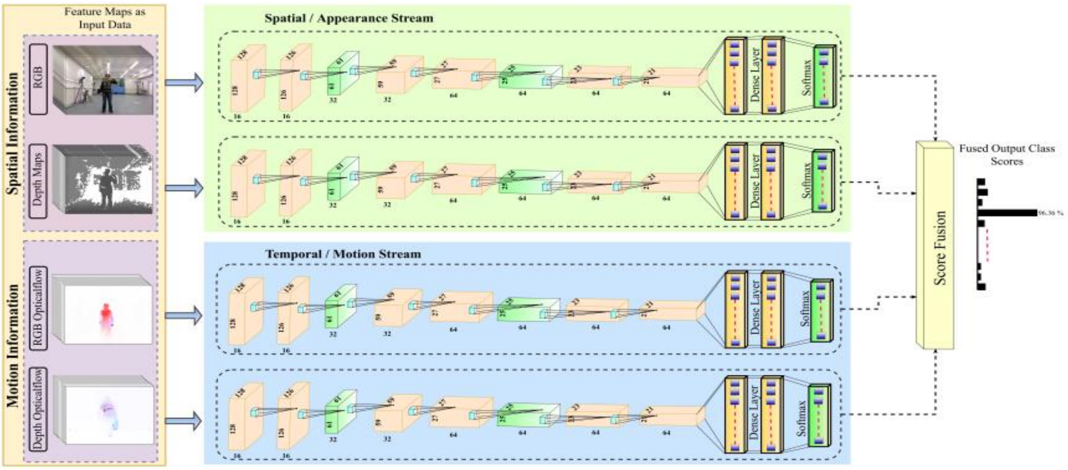

Fig. 5: Action Recognition Framework of the poposed model: The 4 – Stream CNN

4. Result analysis

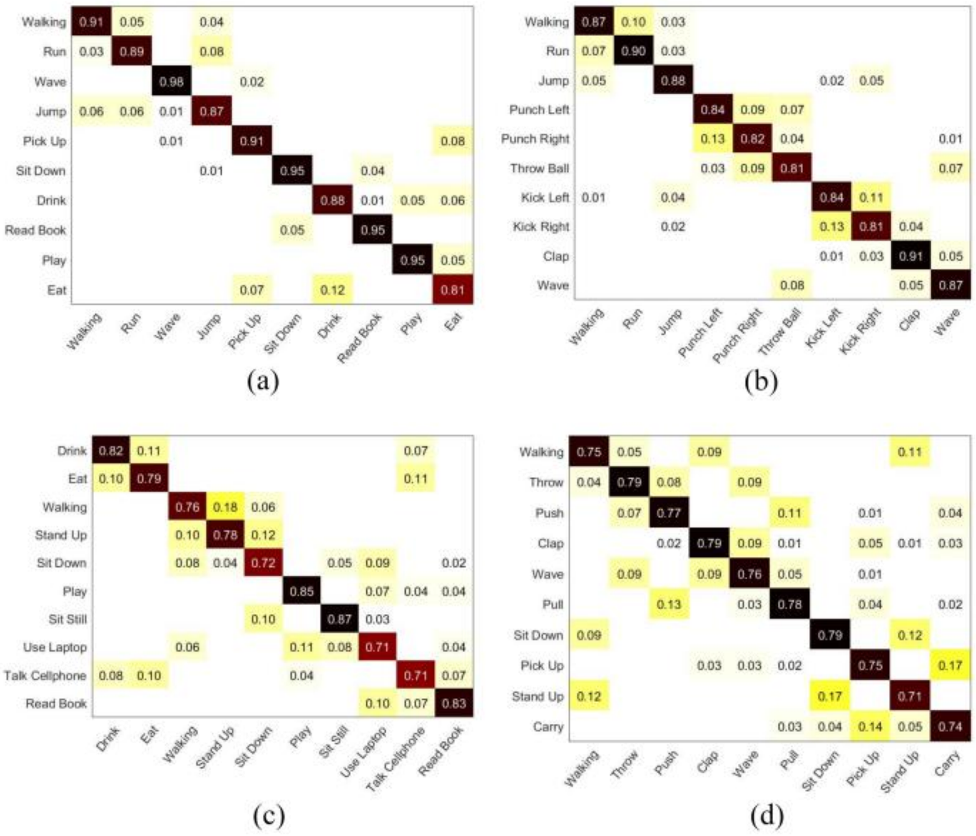

Fig. 6: Confusion Matrices for 10 classes on: (a) BVRC3DAction, (b) NTU RGB-D, (c) MSRDailyActivity3D and (d) UT Kinect.

Inference is performed exclusively using chaotic action data extracted from BVRC3DAction. Ten samples from each dataset were used to generate a confusion matrix, which is depicted in Figure 6. Confusion matrices for all inputs are constructed and illustrated in figure 7. In Figure 8, we show the layered feature maps that we created from our noiseless dataset. The insight into the depth of the convolutional layers is provided through the visualization of the feature maps in Figure 8.

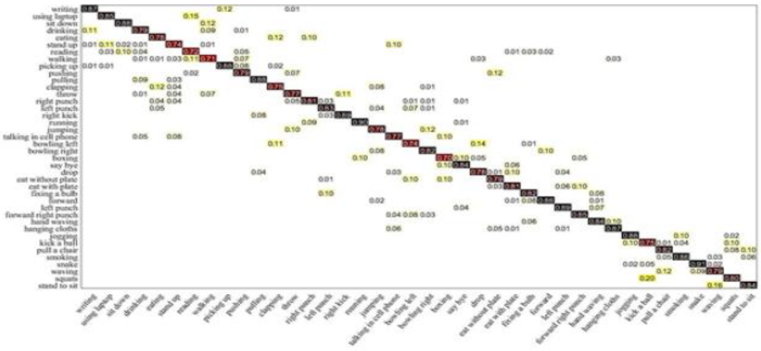

Fig. 7: Shows the Confusion matrix on the entire BVRCAction3D dataset

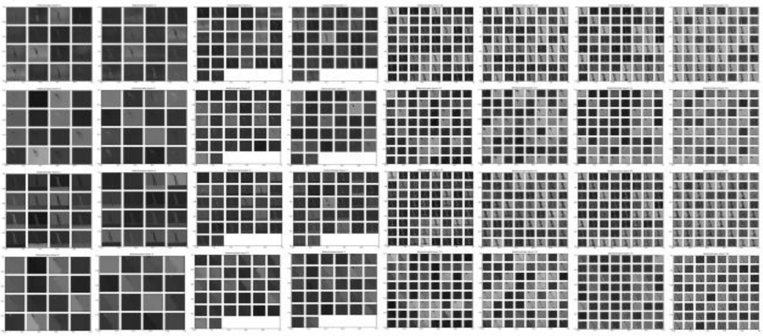

Fig. 8: Intermediate features obtained in each of the 8 CNN layers in 4 streams for BVRC3DAction dataset. The rows in the image are streams and the columns are convolutional operation outputs.

5. Conclusions

To feed a series of multi stream CNN models, past models have focused on either spatial RGB or motion RGB, and depth at varying scales. To classify actions, the combined outputs from these several sources are employed. In contrast, the proposed model is a four-stream convolutional neural network (CNN) with two streams devoted to spatial red-green-blue (RGB) and depth representations and two more streams for apparent motion. Action can be represented more deeply and robustly than with only RGB motion extraction thanks to the depth motion stream's ability to provide subtle small variations in action sequences. In both RGB and depth movies, optical flow is calculated between frames to create a color-coded representation of movement, known as a "motion map." The 4-stream deep learning model has proven effective at representing spatial and motion information to boost the network's overall efficiency.

References

1. R. Poppe, “A survey on vision-based human action recognition,” Image and vision computing, vol. 28, no. 6, pp. 976–990, 2010.

2. S. Shinde, A. Kothari, and V. Gupta, “Yolo based human action recognition and localization,” Procedia computer science, vol. 133, pp. 831–838, 2018.

3. L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian, “Scalable person re identification: A benchmark,” in Proceedings of the IEEE international conference on computer vision, 2015, pp. 1116–1124.

4. C.-C. Yu, H.-Y. Cheng, C.-H. Cheng, and K.-C. Fan, “Efficient human action and gait analysis using multiresolution motion energy histogram,” EURASIP journal on advances in signal processing, vol. 2010, pp. 1–13, 2010.

5. N. A. Rahmad, M. A. As’Ari, N. F. Ghazali, N. Shahar, and N. A. J. Sufri, “A survey of video based action recognition in sports,” Indonesian Journal of Electrical Engineering and Computer Science, vol. 11, no. 3, pp. 987–993, 2018.

6. S. Miao, G. Chen, X. Ning, Y. Zi, K. Ren, Z. Bing, and A. Knoll, “Neuromorphic vision datasets for pedestrian detection, action recognition, and fall detection,” Frontiers in neurorobotics, vol. 13, p. 38, 2019.

7. A. Tharwat, H. Mahdi, M. Elhoseny, and A. E. Hassanien, “Recognizing human activity in mobile crowdsensing environment using optimized k-nn algorithm,” Ex□ pert Systems with Applications, vol. 107, pp. 32–44, 2018.

8. B. Jiang, X. Yin, and H. Song, “Single-stream long-term optical flow convolution network for action recognition of lameness dairy cow,” Computers and Electronics in Agriculture, vol. 175, p. 105536, 2020.

9. B. Allaert, I. M. Bilasco, and C. Djeraba, “Consistent optical flow maps for full and micro facial expression recognition,” in VISAPP, vol. 5. SCITEPRESS-Science and Technology Publications, 2017, pp. 235–242.

10. K. Simonyan and A. Zisserman, “Two-stream convolutional networks for action recognition in videos,” arXiv preprint arXiv:1406.2199, 2014.

11. C. Feichtenhofer, A. Pinz, and A. Zisserman, “Convolutional two-stream network fusion for video action recognition,” in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 1933–1941.