Sports Classification in Sequential Frames Using CNN and RNN

Mohammad Ashraf Russo, Alexander Filonenko, Kang-Hyun Jo

Graduate School of Electrical Engineering,
University of Ulsan, Ulsan, Republic of Korea

ashrafrusso@islab.ulsan.ac.kr, alexander@islab.ulsan.ac.kr, acejo@ulsan.ac.kr

***Abstract*—Automatic sports classification is basic but important task for archiving digital contents in broadcasting companies as well as for general video scene understanding. In this paper, deep learning approach combining convolutional and recurrent neural networks is applied to classify five different classes of sports. Features extracted from CNN are combined temporal analysis using RNN. Multiple experiments results show that the effect of different frame sequence correctly classified test data up to 96.66%.**

Keywords—Sports Classification, Deep learning, CNN, RNN

I. INTRODUCTION

Sports takes a large portion of TV broadcasting and an enormous amount of sports video is being captured every day. Automatic sports classification is important to index each sport according to their category for further processing such as post-game analysis, the formation of tactics for coaches. Also, it helps broadcasting companies to manage their archives easily, and find videos without relying on tedious manual work. This type of classification falls under the group of scene context analysis; thus, it can contribute to further develop a high-level visual understanding of machines.

Researchers applied many approaches to analyzing scene context and classifying visual information according to that over the years, recently deep learning based models have become increasingly popular for complex tasks like these. To recognize any sports activity humans mostly consider a set of actions, sometimes surrounding environment are also considered. Our goal is to build a system which can classify football, cricket, tennis, basketball, ice-hockey 5 different classes of sports based on visual information only. The proposed method approaches this problem based on human intuition which is - the system needs to consider the visual information not only in the spatial domain but also in the time domain. For, learning features or visual cues convolutional neural networks (CNN), and to determine the relation between frames in the time domain recurrent neural networks (RNN) were used.

Section II briefly summarizes the related works, Section III gives an explanation about the networks, Section IV shows the experimental results and Section V concludes the paper.

II. RELATED WORKS

Automatic classification of sports has seen varied approaches in past years. [1] used two separate methods of neural net and texture code cue method and also combined them to compare the results between all three approaches, the neural net cue method was best performing. In [2], thermal imaging technique was used, they produce heatmaps and then project them into low dimensional space using PCA and Fischer's Linear Discriminant. Their overall result was promising for 5 classes. [3] combines audio and video features to classify sports, authors combine MFCC features from audio and visual motion features and could obtain very good results. [4] aims to classify sports among other classes from continuous tv broadcasts, they apply CNN fc7 + PCA 200 to extract features and SVM as the classifier. Obtained results are claimed to be good. The Main focus of [5] was to classify sports type for mobile videos. Authors consider fusion of video, audio and sensor, 3 kinds of data and develop multiclass-SVM for their work. Spatial and temporal analysis was also considered. They give an extensive experimental analysis of the different combinations of fusion approaches. [6] tries to recognize events and classify them in sports videos, Hidden Markov Models (HMM) approach was used for their work. Only computational time performance was given, no mention about the accuracy of their method.

III. PROPOSED METHODOLOGY

For the proposed sports classification method, a combination of convolutional and recurrent network is used, which is shown in Fig.1. This network was inspired from [7] where authors use video sequence to determine smoke's existence in the videos.

A sequence of RGB color frames is given as input to the network. Each frame is fed as input to a separate convolutional layer. All of the convolutional layers share their weights. A rectified linear unit or ReLU is used as activation function and placed at the output of CNN layers. The convolutional part (CONV in Fig. 1) is shown in detailed in Fig. 2 and described in Table 1.

The convolutional part has four layers; the first layer consists of 32 features map with 3x3 receptive field and is followed by max pooling with stride 2. The next three layers are 3 different kinds of dilated convolutions layers which we can describe as the context module. To solve the sports classification problem, the network is taught relations between human action sequences and their surrounding environmental context. Dilated convolutions let us see a wider area which is effective for this task. Also, in experiments, it was seen that adding context module significantly increased the accuracy.

Dilated convolutions were zero-padded. Next in the network is a fully connected layer of 64 neurons which is used to feed the results of the convolutional part to the recurrent part. 0.2 dropout was used to prevent overfitting and get better results.

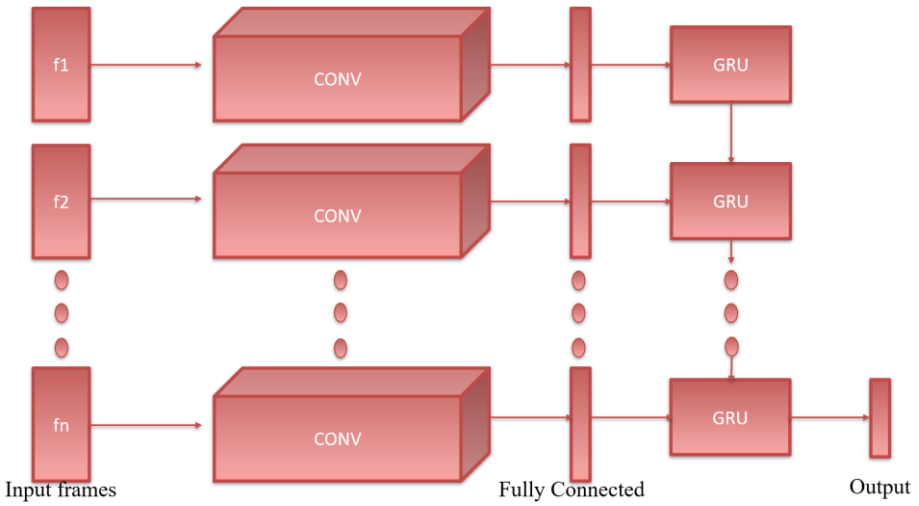

Fig. 1. A Proposed Combined network

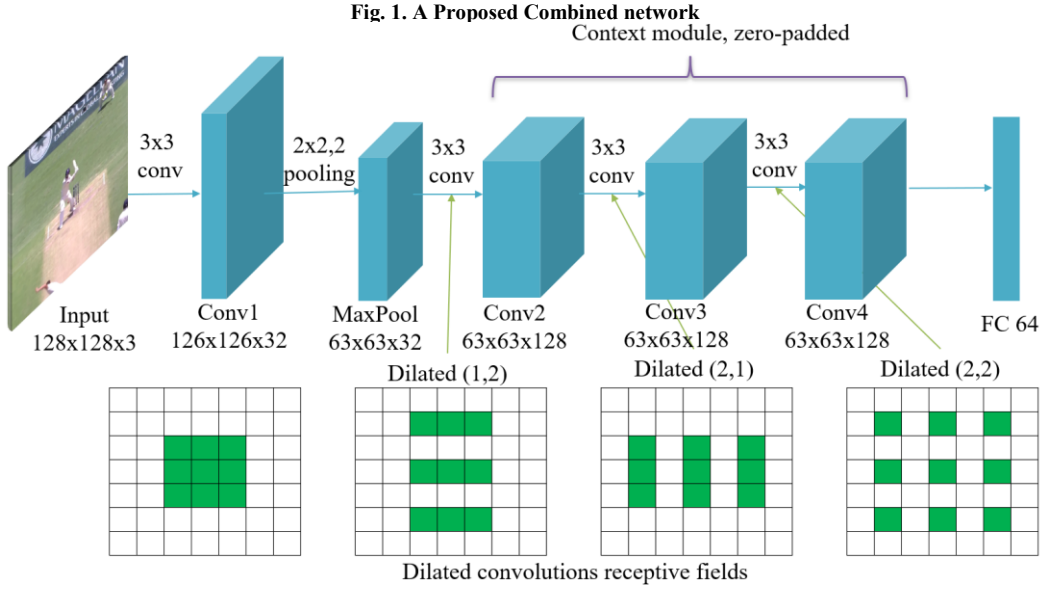

Fig. 2. Architecture of the Convolutional neural network

TABLE 1. Description of the Convolutional part

Layer	1	2	3	4
Filter size	3x3	3x3	3x3	3x3
Dilation(width,height)	-	(1,2)	(2,1)	(2,2)
Feature maps	32	128	128	128
Non-linearity	ReLU	ReLU	ReLU	ReLU
Resizing	Max pooling	-	-	-

Batch Normalization was done after each activation. RMSProp was used as the optimizer. The learning rate was kept constant as '0.0001' throughout all experiments. Categorical Cross-entropy was used as the loss function. The number of epoch was set to 60.

For the recurrent part, the gated recurrent units (GRU) [8] was used. To determine what is the best number of frames in a sequence to feed the network, many tests were run, and the comparison is shown in Fig.4.

IV. Experiments

Hardware configurations for the experiments: AMD ryzen 7 1800x 3.6GHz processor, 32GB Ram, Nvidia 1080ti of 12GB Ram GPU. Keras was used to implement the network and was run on Ubuntu 16.04.

The sports classification dataset contains a total of 300 video sequence derived from 50 different videos of recent sports events taken from YouTube. Each sequence contains 64 frames. To create the dataset, five sports classes were selected: basketball, cricket, football, ice hockey, tennis. For each class, 60 sequences were taken. All the videos have 720 pixels resolution and contain 25 to 30 frames per second. Fig.3 shows five classes' example for 10 frames in a sequence. The input size of the frames was 128x128.

80% of the dataset was chosen for training and 20% were chosen for testing randomly. Test was done for 2 frames in a sequence to 21 frames in a sequence and compared to get the best possible model. More than 21 frames in a sequence (22-64) could not be tested as the current GPU memory of 12 GB was insufficient. The best accuracy for training data could be achieved for 8 frames in a sequence of 99.58%, and for test data, the highest accuracy was 96.66% for 6 and 13 frames in a sequence. The accuracy here can be described as numbers of classes correctly classified/total number of classes.

Comparing the results with the present best results [3] who combines audio and video features to classify three classes of sports of 180 sequence, their highest accuracy was 96.11% while the proposed method shows up to 96.66%.

Fig. 3 Example of 5 sports classes, each row contains 10 sequential frames of a class

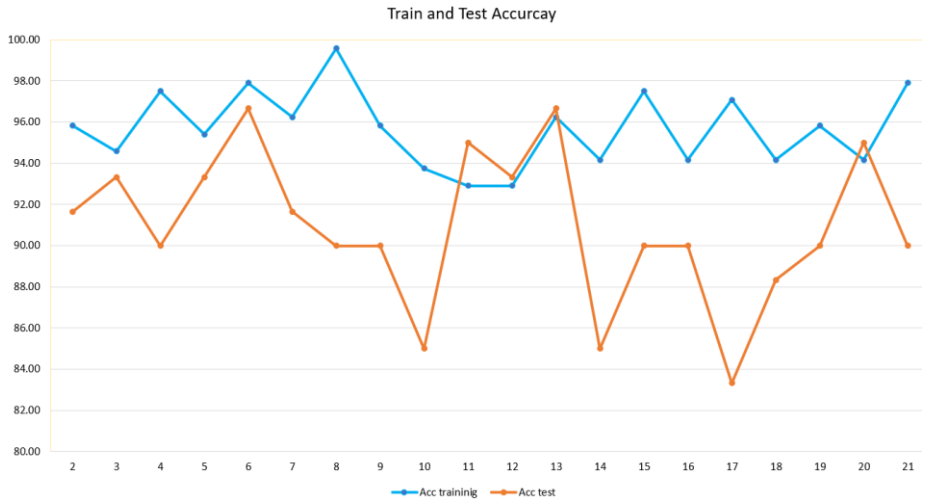

Fig. 4 Comparison of trained and test accuracy from 2 to 21 frames in a sequence

V. CONCLUSION

High-level features are required for the network to classify sports based on human actions and environmental scene context. Two ways to learn high-level features are – the input image size which needs to be large enough and the network which needs to be deep as well. However, it is difficult to implement both because of hardware limitations such as memory. If the image size is set to 64x64 the network can be made deeper as well as it is possible to test even 64 frames in a sequence, but performance drops in consequence. Still, this network performs well for fairly simple dataset like ours. In the future, when more classes are considered, and the dataset is large enough it can be determined what kind of improvement the network requires.

REFERENCES

[1] K. Messer, W. Christmas, J. Kittler, "Automatic sports classification", Proc. IEEE Int. Conf. Pattern Recognition, pp. 1005-1008, 2002.

[2] R. Gade, T. Moeslund, "Sports type classification using signature heatmaps", Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 999-1004, 2013.

[3] R. Gade, M. Abou-Zleikha, M. G. Christensen, T. B. Moeslund, "Audio-visual classification of sports types", 2015 IEEE International Conference on Computer Vision Workshop (ICCVW), pp. 768-773, Dec 2015.

[4] P. Campr, M. Herbig, J. Vaněk and J. Psutka, "Sports video classification in continuous TV broadcasts," 2014 12th International Conference on Signal Processing (ICSP), Hangzhou, 2014, pp. 648-652.

[5] V. Ellappan and R. Rajasekaran, "Event Recognition and Classification in Sports Video," 2017 Second International Conference on Recent Trends and Challenges in Computational Models (ICRTCCM), Tindivanam, 2017, pp. 182-187.

[6] F. Cricri et al., "Sport Type Classification of Mobile Videos," in IEEE Transactions on Multimedia, vol. 16, no. 4, pp. 917-932, June 2014.

[7] Alexander Filonenko, Laksono Kurnianggoro, Kang-Hyun Jo, "Smoke Detection on Video Sequences Using Convolutional and Recurrent Neural Networks", International Conference on Computational Collective Intelligence, pp. 558-566, 2017.

[8] Cho, K., van Merrienboer, B., Bahdanau, D., Bengio, Y., "On the properties of neural machine translation: encoder-decoder approaches", computing research repository (2014).