BASKETBALL-51: A VIDEO DATASET FOR ACTIVITY RECOGNITION IN THE BASKETBALL GAME

Sarbagya Ratna Shakya,, Chaoyang Zhang and Zhaoxian Zhou

School of Computing Sciences and Computer Engineering,
University of Southern Mississippi, Hattiesburg, USA

ABSTRACT

In recent years, there has been an increase in the association of technology in sports and live sports broadcasting networks. From score updates, broadcasting commercials, assisting referees for decision making, and minimizing errors, the adoption of technology has been used for fair play and improve results. This has been possible with the advancement in video analysis, classification techniques, and the availability of resources. This paper introduces a new labelled video dataset collected from a live basketball game broadcasted on live TV to determine the type of basket scored in the basketball game. Among different shots, the points the player can score are basically of three types: 3 points, 2 points, which depends on the range of shots taken and 1 point which is the free shots taken after a foul. This dataset consists of labelled video clips collected from the live broadcast of the game from the broadcasting medium to classify different scoring activities. This paper also gives the preliminary analysis of the dataset for different class labels using 3D ConvNet and two-stream 3D ConvNet methods to show the complexity of the dataset.

KEYWORDS

Basketball dataset, 3D ConvNet, two-stream 3D ConvNet.

1. INTRODUCTION

With recent development in the field of Artificial Intelligence (AI), computer vision, and innovations in deep learning and neural network algorithms, the application of these techniques in different fields especially in sports has been increasing at an incredible pace. Areas of sports such as monitoring player fitness, player injury detection, sports marketing, broadcasting, wearable technology, have been using AI and computer vision in the past few years. Sports personnel, TV broadcasters have implemented AI in automated journalism to enhance sports coverage and spectator experience. Also, AI-powered Wearable devices worn by players provide data that can be used in player tracking [1], player performance analyzing, and optimizing training and player efficiency. With the use of sensors and tracking devices, only limited data about the player and game can be obtained and will be hard to analyze these data in real-time. Also, the necessity to place sensors on the body of players while playing has made it unrealistic to collect data in a real-world scenario. These drawbacks with the sensor data can be resolved if videos can be used to extract information rather than using sensors. But with the large video data collected for many hours and numerous replays to extract the required information from it can make it tedious for people in analyzing the footage. With the application of computer vision, an automated system can be developed that can analyze the videos, players, game situation and gather important insights and valuable analysis from them. Ball tracking, player tracking is some of the major application of computer vision implemented that has provided coaches better understandings of the formation of teams and given instant analysis to better the performance.

David C. Wyld et al. (Eds): CCSIT, SIPP, PDCTA, AISC, NLPCL, BIGML, NCWMC - 2021
pp. 141-153, 2021. CS & IT - CSCP 2021 DOI: 10.5121/csit.2021.110712

The application of AI-based techniques has already been implemented in different sports[2]. AIautomated video highlights generation and broadcasting which picked key moments of the game has been announced and on development. In sports like tennis the ball tracking system has been used to identify whether the ball has landed in or out of bounds. It uses computer vision to construct the trajectory of the detected ball using multiple frames from multiple camera angles. Also, in soccer, the goal-line technology has been adopted that uses multiple camera systems where it uses the computer vision technology system to determine whether the ball has crossed the line on the goal line or not. The analysis gives the referee enough information to make decision in a quick time. This implementation of technology has helped the officials to take the quick and right decision and minimize human errors in sports. Although the application in different sports has provided encouraging results, it has been a challenge due to different factors, to fully automate these systems by using the video.

In this paper, we introduce a new video-based basketball dataset derived from live video broadcast TV for classifying scoring activity in the basketball game. The main objective of building this dataset is to develop a well-labeled dataset dedicated to the activities related to basketball as there is a lack of such datasets for applying automated computer vision techniques for activity recognition. Only a few in other sports like Volleyball[3] where specific sports-related action classes have been used for classification. Our dataset may help researchers to develop and test different architecture and algorithms to evaluate real-time basket and score recognition from live video of basketball games without any human interventions.

Most of the datasets currently existing for basketball have been the dataset that shows the statistics of players and coaches. Most of them include statistics like the number of games played, number of games won, offensive-defensive efficiency, field goal percentage, and many more. Although with the successful application of machine learning and deep learning in the field like face recognition, video segmentation, and its increased application in the video analysis field, the lack of large visual based basketball dataset has been a problem to apply classification techniques in this sport. This dataset provides a short 6 second length clips labelled with different scoring activities as the classification classes sourced from the live videos broadcast on TV. The dataset contains the clips of the scoring attempts from the players throughout the games divided into 8 different categories. Each clip contains information about whether it is attempted from long-range, mid-range, or short-range shots and if it has made or miss the basket. Hence it will help to gain information about the scoring points scored by the player whether it is three-point shots, two-point shots or a free throw for one-points. This should be helpful to develop an automated scoring system during a live game without any human involvement. Also, this will provide information to assist the referee during an unclear condition for better decision making. Hence, the dataset can provide necessary information for research to develop a fully automated system using only the commercial broadcasted video data without using any special camera or camera setup during the games. But the dynamic background of the clips with audience movement, the movement of the camera, change of camera angle, multiple movements of players, presence of flashing information/advertisement, and difficulty in tracking the ball in the background have been some of the challenges this dataset brings in classifying the videos.

Therefore, the main contributions of this paper can be summed up as follows:

a. A new labeled basketball-related action video dataset for activities related to scoring in the basketball games derived from the broadcasted video for a real-world scenario.
b. A baseline reference model and analysis with state of art action recognition models is done to set benchmarks on this dataset.

The organization of the rest of the paper is as follows. Section 2 shows some previous work, summarize existing benchmark sports dataset, as well as a literature review of some of the other sports-related dataset. Section 3 gives a brief explanation and details of the dataset, the data collection process, dataset features, and annotation framework. Section 4 presents the explanation of the baseline models being used, Section 5 shows some experimental results obtained for evaluating the performance of the deep learning models, while finally Section 6 contains the conclusions and future work lines.

2. LITERATURE REVIEW

Many video-based benchmarks human activity recognition dataset such as HMDB51[4], UCF101[5], ACTIVITYNET[6], KINETICS[7] has been published which contains sports-related actions such as catching or throwing a baseball, juggling a soccer ball, playing cricket, shooting a basketball, etc. and has been included as human action classes. These datasets contain videos and images of different activities performed by the subject inside the video and have contributed greatly towards the video classification, with the study of human action classification from clips of humans performing different sports-related activities.

Not only activity related but also some sports-specific video dataset has been collected for video analysis to extract some critical and beneficial information from the video content. Some benchmark datasets Wang et al.[8], UCF Sports[9], Olympic Sports[10], Sports -1M [11], SVW[12] for sports video analysis collected from different sources such as YouTube, TV broadcast, smartphone, and tablets have been developed. These datasets contain images and videos of different sports such as baseball, basketball, tennis, badminton, football, horse riding, running, etc., and have been used to classify different sports. But the lack of inclusion of all the actions related to the specific sports has made it difficult to apply in real-world application and classify the actions related to a variety of actions performed on a single sport. As per example, in dataset UCF101, there are only the activities related to the dunking action as Basketball Dunk and basketball shooting. Rather in the real game, many actions like dribbling, passing, and other different scoring actions are present which is lacking in the previously published dataset. The most published dataset consists of several videos range from around 100 to millions of videos of variable length. The class categories also range from a few to around 500 different sports classes focused on sports actions.

Some specific basketball-related datasets such as SportUV which is explained in this link (https://www.nbastuffer.com/analytics101/sportvu-data/) has been developed where the automated ID and tracking technology system records the tracking of the spatial position of the ball, players, and referee on the court 25 times per second during the game. This dataset indicates when the three-point shot is taken and whether the shot is successful. This data was joined with the play-by-play data from the NBA and is kept that are in both datasets. Since the 2017-2018 seasons, the Second spectrum has been the official player tracking technology provider that collects 3D spatial data of movements of ball, players, referee locations from cameras installed on NBA arenas. In APIDIS dataset [13] the video is collected from 2MPixels color cameras installed around the basketball court (four on each side). In [14] the authors present methods to predict the behavior of the basketball player from the first-person videos (10.5 hours) collected by the University team at Northwestern Polytechnical University.

In earlier times research has also been done to detect scores from the broadcast video in real-time. In [15], the author has proposed a real-time approach to detect score region and recognize the score in broadcast basketball video using frame difference and texture information. They have shown that their approach achieves high accuracy compared to traditional text recognition methods.

Many machines learning and deep learning architecture has been used for action recognition using these video datasets. Most video-based architecture used 2D and 3D content with LSTM[16][17][18][19][20][21] for transferring information across frames and capture long-range dependencies. In recent years, two-stream networks[22][23][24][25] which uses two different types of stream, RGB and flow data, are fused for prediction. For our analysis, we used the 3D ConvNet to train our model from scratch for both our RGB and optical flow video data along with two-stream 3D ConvNet with early fusion techniques. The detail of the dataset is explained in section 3 and the methods we used are explained in section 4.

3. DATA COLLECTION AND DATA CHARACTERISTICS

In this section, we describe how the data was collected, processed, and prepared.

3.1. Data Collection

Step1: Collection of live video games

The videos are collected from the live video of NBA basketball games broadcasted from the different broadcasting channels. Apart from players actions, the videos include scores, flashing information, replays, highlights, interviews, advertisements, and other graphics displayed during the game as it is being broadcasted on live TV. The quality of the video recorded at first is in HD.

Step 2: Manually store the label and timestamp.

Once the live video of the game is recorded, we then manually list the timestamp of the video at the point when the ball reaches near the rim after the player makes a shot towards the basket. The list consists of the hour (HH), minutes (MM), and second (SS) of the video at the time the ball is near the rim from the starting of the video. Then it is manually annotated whether the shot has made the basket or miss the basket and the range from which the shot was taken for labeling the clips.

Step3: Generate clips

Once we have the timestamp from the video, a 6-sec video clip was generated where the point of action is in the middle of the clips. This time duration is chosen so that the clips have complete information of the action such as from where the shot was taken, make or miss of the shot, and the information if there are any rebounds or multiple attempts to make the basket after the first shot. We generate the clips from videos of 51 full basketball games. The average number of clips generated from each video is about 200. To make it more appropriate for experimental purposes, the dimension of the clips is also reduced from HD to the dimension of Quarter Video Graphics display (QVGA) (320×240) pixels value. Sample video frames of the collected dataset are given in Figure 1. The figure shows the sample frame of videos classified into 8 different labels and three specific timestamps of the activities like the point of the throw, the point of the ball on the rim and the point after the ball make or miss the basket.

Step 4: Optical flow dataset

From the RGB video clips, for experimental analysis using temporal data, we also generate the optical flow videos for optical flow data. We find the relationships between the consecutive frames using the optical flow concept which was first proposed by[26] and generate the optical

Figure 1: Example of frames from the video. From left to right represents frames of videos at the time of the shot, ball on the rim, and the ball after the shot and from top to bottom represents the action for 2p0,2p1,3p0,3p1, ft0,ft1, mp0, and mp1.

flow video clips from the RGB video clips. For that, we used the open-source library OpenCV with the Gunnar Farneback optical flow technique [27]. This method detects the pixel intensity changes between the two consecutive frames and gives the highlighted pixels. The optical flow clips were generated and labelled from the original RGB clips. The objective of generating the optical flow clips is to learn the temporal flow of the information in the video.

3.2. Dataset Characteristics

This section describes the detailed characteristics and features of the dataset. The Dataset has a total of 10,311 video clips generated from 51 NBA basketball games broadcasted in the media. The videos are entirely from the third person view captured from the camera used from sports broadcasting media. The clips are initially labelled into 8 class labels: defined as two-point miss(2p0), two-point make(2p1), three-point miss(3p0), three-point make(3p1), free throw miss(ft0), free throw make(ft1), mid-range shot miss(mp0), and mid-range shot make(mp1). The make and miss of the shot taken by players has been represented by 1 and 0 in the activity labels. The distribution of the number of class labels is represented in figure 2. The highest number is that of the three-point miss (about 20% of total data) and the lower is for the mid-range make and miss (around 5.4% of total data). The highest different between the make and miss is in free-throw i.e., there is more free-throws make in games than free throw miss as shown in figure 2. Also, the difference between make and miss is lowest in the mid-range shot which has somewhat equal number of make and miss. To study the characteristics of the dataset, the analysis has been studies based on different groups. The dataset has been grouped on 4-class based on the range of shots taken. The four class categories are two-point, three-point, mid-point, and free throw. Again, to analyze the miss/make of the shots the whole dataset is divided into 2-label dataset of make and miss. The details about the experimental analysis of these groups are described in the experimental setup section of the paper. For proper grouping, the nomenclature of the clips is given to provide information about the label, video number, and the timestamp of the clip which represents the 3rd second of the clip. The optical flow clips were names similar to its RGD video clips and have the same characteristics. Identical models and parameters were used on the optical flow dataset to analyze the performance as has been used for the RGB dataset.

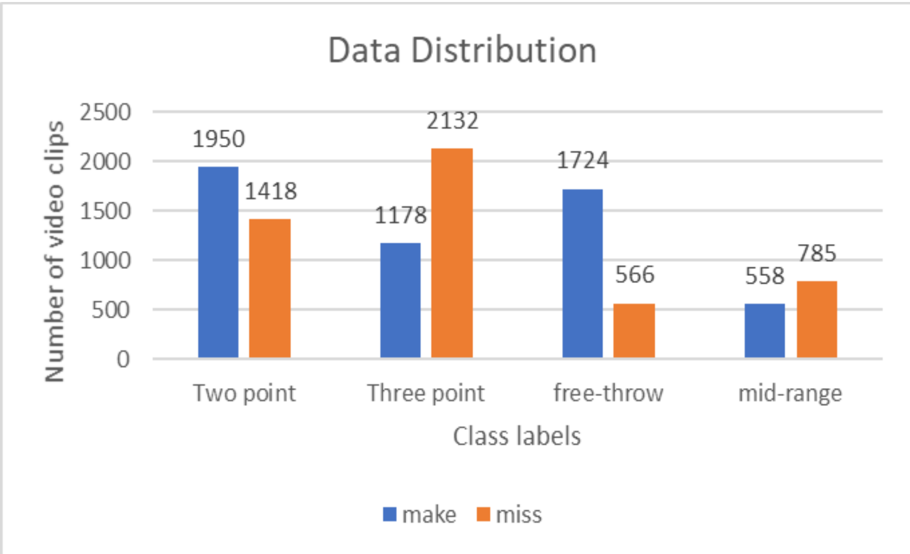


Figure 2: Example of data distribution of the dataset based on different groups.

4. BASELINE PERFORMANCE MODEL

For our evaluation, we used two state-of-art deep learning models for video classification: i) 3D convolution network [28] for video segmentation on both RGB and optical flow video datasets and ii) two-stream ConvNets[22]. The details of the approach are described as follows. In this section, we describe the basic overview of this architecture and explain how we apply this architecture in our experiments.

4.1. 3D ConvNet

3D ConvNets are like the 2D ConvNets but with three-dimensional convolutional kernels which can make segmentation prediction for a volumetric patch. Because of its 3D nature, it seems to be the ideal approach to video modeling and has been used for many video segmentation[29][30] approaches. In addition to the height and weight of the 2D convolutional kernel, the 3D ConvNet has third kernels representing the spatial-temporal filters. Hence it will help to analyze the spatial and spectral features of action between frames of the clips in time dimensions.

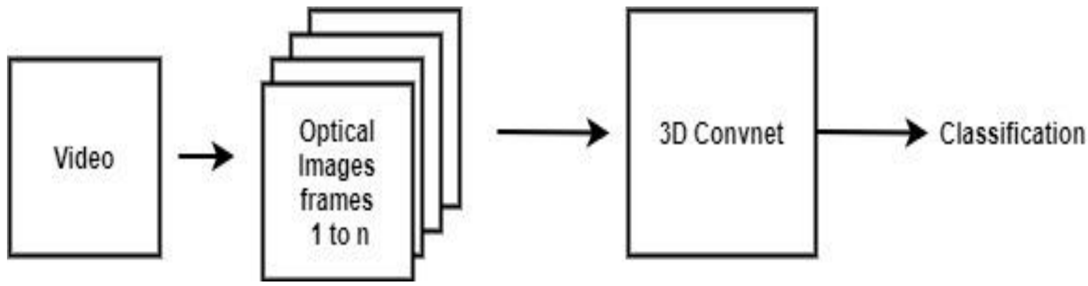


Figure 3: Example of the 3D ConvNet

For our analysis, we implemented a 3D convolutional neural network similar to C3D [31], which has 4 convolutional layers block which includes 3D CNN layer followed by max-pooling layers, dropout layers, and batch normalization layers. Then it is followed by 1 global average pooling layer and 3 dense layers. The model has the input of video clips with 50 frames with a pixel size reduced to size 80×80. The size of the frame is chosen to transform the high-dimensional data with minimum size videos to reduce the memory requirement and computational complexity of the model. The filter numbers range from 16, 64, 256, 512 for 3D CNN and 256 and 32 for fully connected layers. The SoftMax function is used to predict the output classes. The loss function used is categorical cross-entropy. For all our experiments we use the initial learning rate of 0.1 with Nadam optimizer with its default arguments parameters, batch size of 20, and train for 100 epochs. All models are trained on an Nvidia Titan Xp GPU.

4.2. Two stream 3D network

The main idea behind two-stream networks is to train two-stream of CNN networks, one RGB

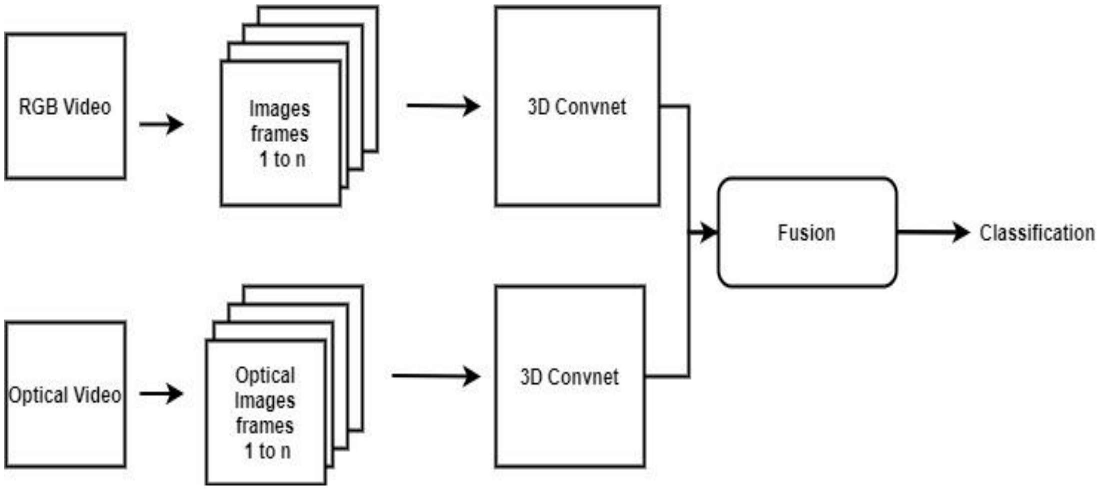


Figure 4: Example of two stream 3D ConvNet

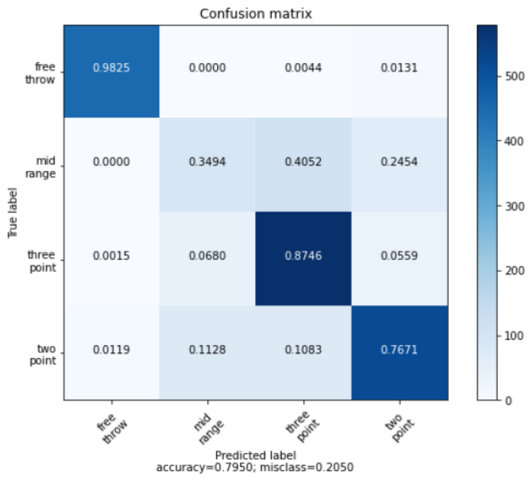
data to get spatial information and another stream with optical flow data for temporal information. These two networks are then fused in some part of the model before classification. In Earlier approaches, a two-stream network is implemented by using short temporal snapshots of video by averaging the predictions from a single RGB frame and a stack of 10 extremely computed optical flow frames [22]. In our case, we generated the optical flow clips from our RGB data clips and used the optical flow clips as our input video data. For our experiment, we used the identical 3D ConvNet network with same input parameters for both the stream with RGB video as input data on one stream and the corresponding optical flow video on the second stream. We then fuse the output taking average after the 3D ConvNet layer and then pass it to the fully connected layers. The configuration of the two-stream network is as shown in figure 4.

5. EXPERIMENTS AND ANALYSIS

This section presents some evaluations using this basketball dataset to illustrate the characteristics and challenges for action recognition. We used one of the most used deep learning architectures, 3D ConvNet for video classification. The task aims to correctly classify the scoring label of the video clips that contain the scoring activity of the player when they take a shot. Here we use the clips derived from the broadcasted videos to train the classifier and evaluate the performance based on different classifier algorithms. For our evaluation, we test our model into two classification types: Subject dependent and subject independent. Here the subject represents the video of the basketball game. The objective of analyzing in different classification type is to study the impact in the performance of the model during training and testing because of the difference in the background audience, court color, players jersey, player movement and camera orientations that can have different features in different videos from one game to another.

5.1. Subject Dependent

In Subject Dependent classification, the division of the dataset is done randomly to training and testing dataset where both contains samples from all the subjects. If there are samples for n number of subjects, both training and testing dataset contains certain percentage of samples from all n number of subjects. For the subject dependent, the total



Classification Report

	precision	recall	f1-score
0	0.98	0.98	0.98
1	0.44	0.35	0.39
2	0.76	0.87	0.81
3	0.83	0.77	0.80
accuracy			0.79
macro avg	0.75	0.74	0.74
weighted avg	0.79	0.79	0.79

Figure 5: Confusion Matrix and classification report for the RGB basketball dataset using 3D ConvNet for 4 class subject dependent classification.

clips collected from 51 games is divided into (80/20) training and testing datasets based on their class labels where clips from each video can be in the training and testing data. This random division of the total data into training and testing data, have samples corresponding to the same video and, most likely the samples of all subjects. The confusion matrix and classification report for the RGB dataset for 4-class subject dependent classification is shown in figure 5. The 0,1 2 and 3 in classification reports represents the classification class of free throw, mid-range, three-points and two-points. Here, the mid-range has the lowest recall value of about 35% whereas the free-throw activity has the highest classification recall value of about 98%. The overall accuracy is 79%. Most of the mid-range shots has been misclassified as three-point and two-point shot as there is a small margin of range between mid-range with two-point and three-point range. Also, the low number of training data for mid-range can be the reason for low classification accuracy.

5.2. Subject Independent

In subject independent classification, the division of dataset is done to training and testing dataset such that samples from the subject included in training are not included in testing data. That isfor n number of subjects, samples from n-k subjects are used in training data whereas samples from the remaining k subjects are used in testing. The system classifier will be tested with testing data having completely new features than the training data. For subject independent classification, the clips from 41 different games were used as training while the clips from the next 10 games were used as testing datasets. This is done to make the subject and features of the testing data completely unknown to the classifier trained on entirely different training video data. The objective is to analyze the effect on the classification performance due to differences in a court appearance, player movement, background change, camera orientation,

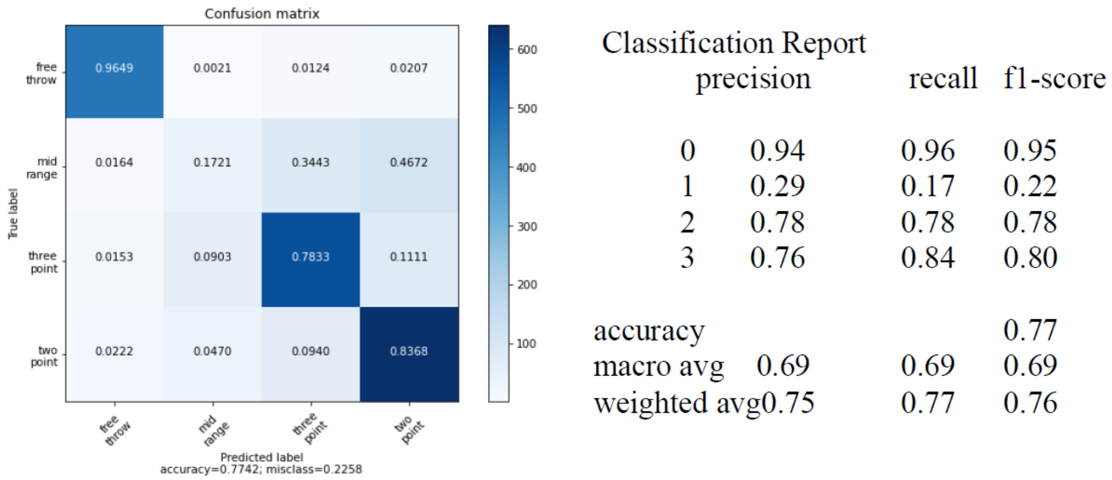


Figure 6: Confusion Matrix and classification report for the RGB basketball dataset using 3D ConvNet for 4 class subject independent classification.

and other factors in testing data as compared to the dataset used for training. The confusion matrix and classification report for the RGB basketball dataset using 3D ConvNet for 4 class classification is shown in figure 6. Here also the mid-range has been misclassified mostly as two-point range and three-point range and has the lowest recall value of 17%. The overall accuracy is 77% which is less than the subject-dependent classification accuracy.

In both cases, 20% of the training dataset was used as a validation dataset for evaluating the performance of the model during training. With each epoch, the model tests the performance on the validation dataset and will save the model if the performance betters (validation loss improves) than the previous saved best model. We evaluate our training for 100 epochs. Thus, at the end of the 100 epochs, the saved model will be the best one with the lowest validation loss throughout the training process. Then the performance is evaluated in the testing dataset. For diverse analysis, we did our experiment for different groups of data. First, we analyze our experiments for our original 8-class labels. Then to analyze the range of the shots, we group our dataset into four different class labels(4-class) (three-point, two-point, mid-range, and free shot) and finally to analyze the scoring of the shots we analyze the model for two-class labels(2-class) (make/miss) of the shots.

Table 1: Accuracy comparison of different model for different class group

Methods	**Accuracy**					
	Subject Dependent			**Subject Independent**		
	8-class	**4-class**	**2-class**	**8-class**	**4-class**	**2-class**
3D ConvNet	59.48%	79.50%	77.31%	56.78%	77.42%	75.38%
3D optical ConvNet	51.72%	73.39%	72.47%	52.08%	74.44%	70.78%
Two stream 3D ConvNet	58.94%	74.79%	76.44%	51.85%	73.13%	76.06%

Table 1 shows the comparison of the accuracy for the 3D ConvNet for RGB data, optical flow data, and two-stream model with average fusion. We observe that among different input data in CNN models, the accuracy is mostly high with 3D ConvNet with RGB videos as compared to 3D ConvNet with optical video. The 3D ConvNet model has higher accuracy performance from RGB videos than optical flow videos in all cases. The performance using two-stream networks has not improved the result significantly than using 3D ConvNet with RGB data. Also, among different groups of data, the model has higher accuracy in range classification(4-class) than in the 8-class group or 2-class group. Only that using two-stream 3D ConvNet the make and miss have shown higher performance compared to in 4-class in both subject dependent and subject independent classification than using single 3D ConvNet for RGB and optical dataset separately. We also observe that the accuracy is higher in most cases with subject-dependent analysis than subject independent analysis. Figures 4 and 5 present the confusion matrix based on recall and classification reports for the 4- class 3D ConvNet methods for subject dependent and subject independent classification methods. As described in section 5.1 and section 5.2, most of the misclassification is for mid-range which is mostly misclassified as a two-point or three-point shot. This can be due to the imbalanced nature of the dataset where there is a smaller number of mid-range data clips compared to 2-point and 3-point. Also, the identical similarity of the interclass activity and lack of specific range distinction between different ranges can be the reason for low performance. Also, we can see that the model can capture features of the free-throw which has significantly higher accuracy as compared to another group. During a free throw, the less movement of the players, constant camera angle, and low camera movement than in other activity groups can be the reason for the higher classification accuracy.

6. CONCLUSION AND FUTURE WORKS

In this study, we try to develop a dataset to classify the scoring action from a basketball game broadcasting video. Also, to learn the spatial and temporal features we used the state-of-the-art classification method of 3D ConvNet. Future work includes adding more activities of players in the game like dribbling, fouls, and scoring types for full automation of the activity recognition in the live videos. In most of the other dataset when deriving the optical flow dataset, the background will be static but in our case with the movement of the camera, results in the dynamic movement of the background. This has increased the challenges for improving the results using two-stream model and has also results in low classification performance with temporal information from optical data than considering only RGB spatial information.

Some of the other challenge it faces is the similar nature of the videos between different classes. The temporal information on the training will be not only from the movement of the ball or the player with the ball but it also tracks the movement of all the players on the court along with the referee and background audience. This dynamic nature of the dataset has added more challenges for classification. The experiments used for analyzing the performance on this dataset explained in this paper uses only basic parameters and features of the dataset. Analysis can be made by using higher dimensions input data with other deep learning networks considering the challenges the dataset presents. This can certainly help in increase the performance for classifying the scoring activities. This shows a higher possibility and opportunity to increase the performance using this dataset for further research.

REFERENCES

[1] W.-L. Lu, J.-A. Ting, J. J. Little, and K. P. Murphy, "Learning to track and identify players from broadcast sports videos," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 7, pp. 1704–1716, 2013.

[2] G. Thomas, R. Gade, T. B. Moeslund, P. Carr, and A. Hilton, "Computer vision for sports: Current applications and research topics," *Comput. Vis. Image Underst.*, vol. 159, pp. 3–18, 2017.

[3] M. S. Ibrahim, S. Muralidharan, Z. Deng, A. Vahdat, and G. Mori, "A Hierarchical Deep Temporal Model for Group Activity Recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[4] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre, "HMDB51: A Large Video Database for Human Motion Recognition," in *Proceedings of the IEEE International Conference on Computer Vision*, 2011, pp. 2556–2563.

[5] K. Soomro, A. R. Zamir, and M. Shah, "{UCF101:} {A} Dataset of 101 Human Actions Classes From Videos in The Wild," *CoRR*, vol. abs/1212.0, 2012.

[6] B. G. Fabian Caba Heilbron Victor Escorcia and J. C. Niebles, "ActivityNet: A Large-Scale Video Benchmark for Human Activity Understanding," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 961–970.

[7] W. Kay *et al.*, "The Kinetics Human Action Video Dataset," *CoRR*, vol. abs/1705.0, 2017.

[8] Yang Wang, Hao Jiang, M. S. Drew, Ze-Nian Li, and G. Mori, "Unsupervised Discovery of Action Classes," in *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, 2006, vol. 2, pp. 1654–1661.

[9] M. D. Rodriguez, J. Ahmed, and M. Shah, "Action MACH a spatio-temporal Maximum Average Correlation Height filter for action recognition," 2008.

[10] J. C. Niebles, C.-W. Chen, and L. Fei-Fei, "Modeling Temporal Structure of Decomposable Motion Segments for Activity Classification," in *Computer Vision -- ECCV 2010*, 2010, pp. 392–405.

[11] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei, "Large-scale Video Classification with Convolutional Neural Networks," in *CVPR*, 2014.

[12] S. M. Safdarnejad, X. Liu, L. Udpa, B. Andrus, J. Wood, and D. Craven, "Sports Videos in the Wild (SVW): A Video Dataset for Sports Analysis," in *Proc. International Conference on Automatic Face and Gesture Recognition*, 2015.

[13] F. Chen, D. Delannay, and C. De Vleeschouwer, "An autonomous framework to produce and distribute personalized team-sport video summaries: A basketball case study," *IEEE Trans. Multimed.*, vol. 13, no. 6, pp. 1381–1394, 2011.

[14] S. Su, J. Pyo Hong, J. Shi, and H. Soo Park, "Predicting Behaviors of Basketball Players From First Person Videos," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[15] G. Miao, G. Zhu, S. Jiang, C. Xu, and W. Gao, "A Real-Time Score Detection and Recognition Approach for Broadcast Basketball Video," 2007, pp. 1691–1694.

[16] N. Liu and J. Han, "A deep spatial contextual long-term recurrent convolutional network for saliency detection," *IEEE Trans. Image Process.*, vol. 27, no. 7, pp. 3264–3274, 2018.

[17] J. Donahue *et al.*, "Long-term recurrent convolutional networks for visual recognition and description," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 2625–2634.

[18] M. Abdullah, M. Ahmad, and D. Han, "Facial Expression Recognition in Videos: An CNN-LSTM based Model for Video Classification," in *2020 International Conference on Electronics, Information, and Communication (ICEIC)*, 2020, pp. 1–3.

[19] J. You and J. Korhonen, "Attention Boosted Deep Networks For Video Classification," in *2020 IEEE International Conference on Image Processing (ICIP)*, 2020, pp. 1761–1765.

[20] T. Akilan, Q. J. Wu, A. Safaei, J. Huo, and Y. Yang, "A 3D CNN-LSTM-based image-to-image foreground segmentation," *IEEE Trans. Intell. Transp. Syst.*, vol. 21, no. 3, pp. 959–971, 2019.

[21] S. H. Bae, I. Choi, and N. S. Kim, "Acoustic scene classification using parallel combination of LSTM and CNN," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2016 Workshop (DCASE2016)*, 2016, pp. 11–15.

[22] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," *arXiv Prepr. arXiv1406.2199*, 2014.

[23] X. Peng and C. Schmid, "Multi-region two-stream R-CNN for action detection," in *European conference on computer vision*, 2016, pp. 744–759.

[24] Q. Xiong, J. Zhang, P. Wang, D. Liu, and R. X. Gao, "Transferable two-stream convolutional neural network for human action recognition," *J. Manuf. Syst.*, vol. 56, pp. 605–614, 2020.

[25] P. Thiam, H. A. Kestler, and F. Schwenker, "Two-stream attention network for pain recognition from video sequences," *Sensors*, vol. 20, no. 3, p. 839, 2020.

[26] R. Hetherington, "The Perception of the Visual World. By James J. Gibson. USA: Houghton Mifflin Company, 1950 (George Allen & Unwin, Ltd., London). Price 35s.," *J. Ment. Sci.*, vol. 98, no. 413, p. 717, 1952.

[27] G. Farnebäck, "Two-frame motion estimation based on polynomial expansion," in *Scandinavian conference on Image analysis*, 2003, pp. 363–370.

[28] K. Hara, H. Kataoka, and Y. Satoh, “Can Spatiotemporal 3D CNNs Retrace the History of 2D CNNs and ImageNet?,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.
[29] S. Ji, W. Xu, M. Yang, and K. Yu, “3D convolutional neural networks for human action recognition,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 1, pp. 221–231, 2012.
[30] G. W. Taylor, R. Fergus, Y. LeCun, and C. Bregler, “Convolutional learning of spatio-temporal features,” in *European conference on computer vision*, 2010, pp. 140–153.
[31] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, “Learning Spatiotemporal Features with 3D Convolutional Networks.” 2015.

AUTHORS

Sarbagya Ratna Shakya received the B. Eng. in Electronics Engineering from National College of Engineering, Tribhuvan University of Nepal in 2009; M. Eng. in Computer Engineering from Nepal College of Information Technology, Pokhara University of Nepal in 2014. Currently he is a PhD student in School of Computing Sciences and Computer Engineering, University of Southern Mississippi since 2016. His current research interests include machine learning, deep learning, and high-performance computing.

Chaoyang Zhang received his MS degree in computer science and PhD degree in computational analysis and modelling from Louisiana Tech University in 2001. He is currently a Professor of Computer Science in the School of Computing Sciences and Computer Engineering at the University of Southern Mississippi. He has published more than seventy papers in academic journals and conference proceedings. His research interests include data mining, machine learning, bioinformatics, image processing and high-performance computing.

Zhaoxian Zhou received the B. Eng. from the University of Science and Technology of China in 1991; M. Eng. from the National University of Singapore in 1999 and the PhD degree from the University of New Mexico in 2005. All His degrees are in Electrical Engineering. From 1991 to 1997, he was an electrical engineer in China Research Institute of Radio wave Propagation. He joined the University of Southern Mississippi in 2005. He has published more than fifty papers in academic journals and conference proceedings. His current research interests include computational science and electrical engineering.