iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 11 **Issue:** I **Month of publication:** January 2023

DOI: https://doi.org/10.22214/ijraset.2023.48794

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 11 Issue I Jan 2023- Available at www.ijraset.com

Scene Classification for Sports Video Using Transfer Learning

Prakhar Singh[1], Nilava Dutta[2], Rajvivek Sundriyal[3], Pulkit Suri[4], Kanika Malik[5]
[1, 2, 3, 4, 5]*IMSEC, Ghaziabad*

Abstract: ***The quality of scene classification is particularly crucial because it is a key step in both video summarization and video analysis. This article proposes a strategy to accomplish high-quality scene classification and focuses on several practical implementation gaps over the previous methodologies. We identify five scene types, including batting, bowling, boundary, crowd, and close-up, using cricket as a case study. For scene categorization, we use the ResNet152 model which is pre-trained on imagenet. The proposed method employs new, fully connected layers in an encoder fashion. We also prepared a dataset of cricket video and manually labelled them into five classes which was later used to train the model. The model achieved mean accuracy of 99.15 upon training on the dataset.***

Keywords: Deep Learning, Transfer Learning, ResNet, Neural Network, Labelled Dataset

I. INTRODUCTION

Video processing, monitoring, summarising, and broadcasting are expanding quickly as a result of the introduction of live broadcast apps and the availability of live streaming on top video hosting websites. Video content searches, video descriptions, and video summaries are currently trending topics. Researchers are actively working on automatic video summarization. Scene classification is the key element for the sports video summary. Each sport has specific predefined camera angles, zooming processes, and action replay techniques. Sports videos are known for having specialised scene protocols and known criteria for video capture.

We used recorded footage from a publicly accessible YouTube source and used our suggested model's focus on cricket as a case study. The decreased classification error rate guarantees the high dependability of the classification and, consequently, the high reliability of the dependent processes that use the building blocks of scene categorization.

With the development of machine learning and other forms of artificial intelligence (AI), computers are aiding humans more promisingly than ever before. Scene categorization is a crucial building element, thus it needs a lot of work to outperform current research state-of-the-art performance. There is still a need to address the quality gap, and practical implementation constraints, in machine learning. Improvement is required to achieve high-quality classification on a smaller dataset in practically less time. This is because labelled sports datasets are not available in this area and context. The proposed method improves upon the existing state-of-the-art technique. Our cricket sports dataset is created using YouTube cricket sports recorded recordings of previous occurrences. We identify the five most common and crucial scenes from the videos—batting, bowling, the boundary, a close-up, and the crowd.

II. LITERATURE SURVEY

Sports videos are often classified using shot classification and scene classification methodologies. Recent approaches used a variety of techniques and models, including deep learning models like AlexNet CNN and Recurrent Neural Network. Deep structural methods guarantee more accuracy but at the expense of greater computing complexity.

A categorization of soccer video frames using multi-class kernel SVM was presented by Rahul et al. [1]. A shot-detection system based on windows is modified. The difference between a given frame and each subsequent frame is calculated for a predetermined window size. It is possible to categorise the shot as a Bowler Runup, Batsman Stroke, Player Close-Up, Umpire, Ground, Crowd, Animations, or Miscellaneous.

Zhou L. [2] proposed a moving sports video moving object recognition and tracking system. Students' physical education could be changed by including the hidden Markov model. Throughout the tracking procedure, the used approach maintains incredibly high precision and stability. The system uses a top-view camera to monitor players in the ground as well as an onboard computer to monitor their position.

Ling-Yu et al. [3] proposed a unified framework for semantic scene classification in sports videos.The proposed framework achieved a classification accuracy of 85% to 95% on five different sports, i.e., volleyball, basketball, tennis, soccer, and table tennis. Camera shots were classified using different supervised learning algorithms, like decision trees, neural networks, support vectors machine (SVM).

Ashok et al. [4], proposed an optical flow approach for different sorts of batsman stroke detection in cricket sports videos.An optical flow approach for different sorts of batsman stroke detection in cricket videos has been proposed. A histogram-based technique is being adopted to detect the shot-boundary. Low-level features (i.e., grass pixel ratio, distance, etc.) were used to train a hybrid classifier to classify seven classes.

A study by Minhas et al. [5] presented a deep-learning-based shot classification model and presented good accuracy over different shot classes consisting of long, medium, close-ups, and crowd/out-of-field shots. The adopted methodology demonstrated good accuracy against baseline methods on the subjected dataset.

Russo et al. [6] proposed a model by combining deep learning and transfer learning, i.e., combining the functionality of VGGNET16, RNN,and GRU with transfer learning for final classification.Mixed sports videos were used to train a deep neural network to classify 10 and 15 target classes using CNN as feature extractor and then combining with temporal information from RNN for model formulation. The model demonstrated a good accuracy of 94% for 10 and 92% for 15 sports classes.

Jungheon et al. [7] proposed a multi-Modality classification system for video events using visual and audio relationships extracted from the video.Inception-V3 algorithms are employed for image feature vector extraction. Mel Frequency CepstralCoefficients (MFCCs) were used to extract the audio feature vector. YLI-MED dataset with 1823 videos and a self-collected YouTube dataset comprising 1369 videos were used for evaluation. Normalization of audio and visual feature vectors, as well as correlation integration, achieved better performance results.

Sozykin K et al. [8] proposed a 3D CNN-based multi-label deep Human Action Recognition (HAR) system for sports video summarization for the sport of Hockey and presented more than ten classes.A study has presented a 3D CNN-based multi-label deep Human Action Recognition system for sports video summarization for the sport of Hockey. Data pre-processing techniques like resizing, normalization, windowing, and sequence labeling were used. The dataset used for evaluation contains 36 grayscale videos recorded with a static camera position.

Bhalla et al. [9] proposed a multimodal approach for automatic cricket video summarization. They employed Optical Character Recognition (OCR), sound detection, and replay detection techniques. The proposed system achieved an accuracy of 89.45% for event detection.

Tejero-De-Prablos [10] presented a scheme for user-generated sports video summarization and employed the LSTM model to identify various user actions in the video.Videos were classified into interesting or non-interesting categories based on the user's choice of key features such as camera angle, frame rate and noise level.

An optical flow tracing method for shot-by-shot analysis of cricket has been developed by Karmaker et al. [11] A filter trained over six videos to recognize the actions was employed. Decision trees and the ELMAN neural network were used to classify the shots.

III. DATASET

To train the model for performance evaluation, labelled dataset of cricket sport video was required. We were dealing with scene classification for five classes namely: Bowling, batting, close-up, boundary, crowd which is shown in Fig 1. To create the dataset we first needed frames which was later labelled into the 5 classes. We have selected various sports videos from YouTube, considering different series and different lighting conditions.

Fig. 1 Scene classes, (a) Bowling, (b)Boundary, (c)Batting, (d) Close-up, and (e) Crowd.

Total six thousand relevant frames were identified and classified into 5 classes of bowling, batting, boundary, closeup and crowd. There was an increased number of frames of class 3 which is closeup in comparison to other classes. The frames distribution in respective classes is shown in Fig 2.

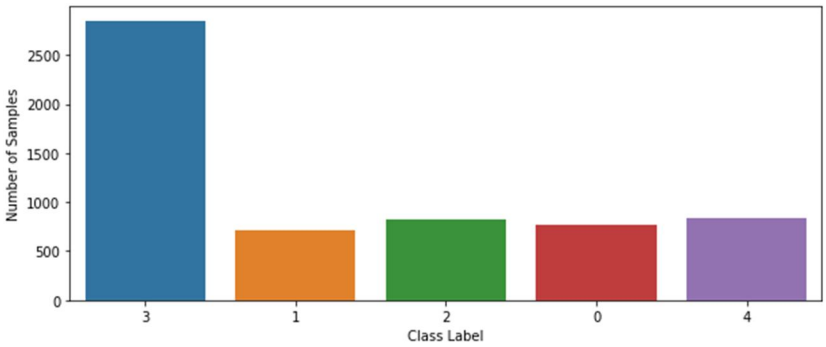


Fig. 2 Labeled dataset of each classes

IV. ARCHITECTURE

The model used is ResNet152 which has a total of 152 layers. This model is pretrained on imagenet weights and we applied three transfer learning layers at the end of the ResNet model and trained the model on the labelled dataset shown in Fig 3.

A. Transfer Learning

Every neuron in the input tensor is cross-connected to every neuron in the output tensor by a fully connected layer, which has an activation function at the output layer. The input tensor is flattened by the fully connected layer if the input shape rank is greater than 2. As a result, a completely linked layer is the dot product of the applied kernel and the input tensor.

Fully connected layers generate probabilities in a wide range; therefore, to reduce the noise level, it is essential to eliminate the weaker predictions. To perform this operation, a dropout process is carried out with a certain threshold of usually 0.5. This less densifies the neurons after removing the value of lesser probabilities. The dropout layer helps to avoid a model overfitting and significantly improves the validation accuracy. A dropout layer reduces the functional size of the process by removing unproductive neuron outputs. It is a regularization technique and filters out complex co-adaptation during the training phase.

B. Activation function

A tensor output activation function called the Rectified Linear Unit (ReLU) makes the model training process non-linear. Tensor output from the convolution process could have both positive and negative values; as a result, an activation function is used before sending the output to the following layer.

Positive values remain untouched while values higher than zero are converted to zero using a ReLU function. This action is known as correction. When given a range of negative and positive values, the non-saturating function f(x) = max(0, x) returns zero or a positive result. Negative values are removed from the final feature map. During the convolution process, it increases the non-linearity of the model without degrading the accuracy of classification in receptive fields.

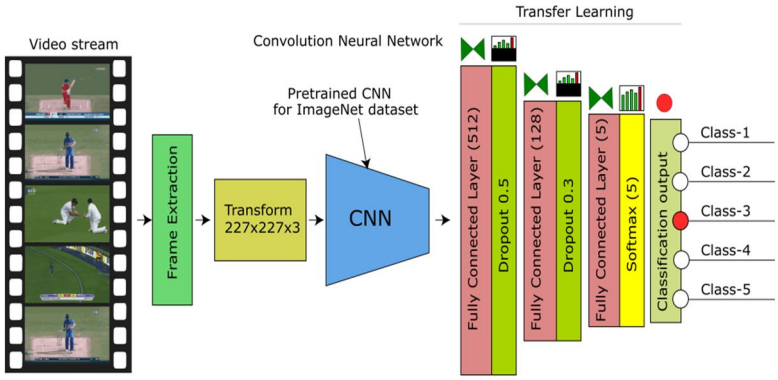


Fig. 3 The proposed model architecture

C. Libraries and Tools

For the training model we made use of Google Colaboratory and various libraries of python and keras. The Tools and Libraries used are mentioned below:

1) *Google Colaboratory*: Colaboratory, sometimes known as "Colab," is a data analysis and machine learning tool that enables you to integrate rich text, charts, photos, executable Python code, HTML, LaTeX, and more into a single Google Drive document during the convolution process, of classification in the receptive fields.
2) *Python:* Python is a high-level, interpreted language that is renowned for its readability, versatility, and simplicity. Since it is a general-purpose language, it may be used to create almost any kind of software, including desktop, web, and scientific applications.
3) *Keras:* Python-based Keras is a high-level, open-source neural network library. It was created to make the construction and prototyping of deep learning models simpler for researchers and practitioners. It utilises well-known deep learning packages like TensorFlow, Theano, and CNTK as its foundation.
4) *TensorFlow:* An open-source software library for artificial intelligence and machine learning is called TensorFlow (AI). It was created by Google and is extensively utilised in both academics and business for activities like language and picture processing. The foundation of TensorFlow lies in the idea of data flow graphs, in which the nodes of the graph stand in for mathematical processes and the edges for the data that passes through them.
5) *Matplotlib:* Matplotlib is a data visualization library for Python. It allows users to create a wide range of static, animated, and interactive visualizations in Python. With Matplotlib, you can create line plots, scatter plots, bar plots, error bars, bar plots, histograms, bar plots, box plots, pie charts, and many other types of visualizations. It is particularly useful for creating plots and charts to visualize data.
6) *Scikit-learn:* Scikit-learn is built on top of NumPy and SciPy, two libraries for scientific computing in Python. This means that it is fast and efficient, and can easily be integrated into larger projects. It is also well-documented, with a comprehensive user guide and numerous examples.

V. RESULT AND DISCUSSION

We have presented results for images and manually labeled dataset of sports videos. We evaluated their precision, recall, accuracy, error rate, and F1-score for all available categories on the dataset of videos. The result section represents the experimental evaluations of the performance of the proposed methodology.

A. Training Procedure

We employed TensorFlowTM backend with Keras applications for training purposes and incorporated dataset generators to handle the large dataset of images. Training achieved a robust accuracy of above 99% in very initial epoch. As the model was pre trained on imagenet weights, it converges very fast hence we used 5 epochs. Fig 4 shows the training and validation accuracy and loss respectively.

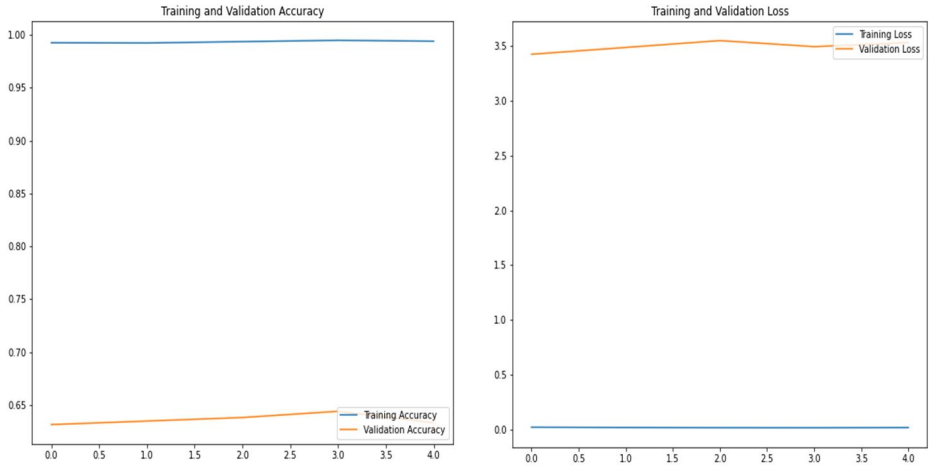


Fig. 4 (a) Training and Validation Accuracy, (b) Training and Validation Loss

B. Learning Rate Selection

Different learning rates were tested during the trials in order to train and evaluate the model. Under Adam Optimizer, the learning rate that performed best was 0.0001.

C. Performance Evaluation

Evaluations of the proposed model show the variations within and between classes.Although there are substantial similarities between boundaries and crowd because of the audience seen in the long-distance camera view, the proposed model classified it precisely even if there are obvious inter-class similarities in bowling and batting.Performance indicator F1-score shows that the proposed model outperformed other deep CNN models as well as SVM.

VI. CONCLUSION

We propose a model for sports video scene classification with the particular intention of sports video summarization. Our findings demonstrated that our proposed model performed best against state-of-the-art models and existing recent research level. We evaluated our model on cricket as a case study and obtained videos from YouTube.

The methodology could be used for medical image analysis, agriculture, biological organism classification, self-driving vehicles and other fields where data requires higher accuracy and error rates.

VII. FUTURE WORK

The model proposed is highly accurate in scene classification and can be used for cricket sports video summarization. The model will work equally well for other sports videos.

The proposed methodology is tested on sports video data to serve as a building block for video summarization, but it can also be applied to other fields such as medical image analysis, agriculture classification, biological organism classification, self-driving cars, and many others were classification calls for higher accuracy and the error rate is a major limiting factor.

The proposed model outperformed prior presented models in terms of performance measures, making it a great option to serve as a reusable building block. Research on video description, medical video analysis, and event description can be expanded to include scene classification. Similar to this, sports video to text summary is a current hot topic that calls for researchers to enhance the task's quality.

VIII. CODE

We have modified the ResNet152 model by applying three transfer layers over the ResNet model. The code for the entire project is hosted on github and linked below:

https://github.com/prakr97/scene_classify

IX. CONTRIBUTION

A. Prakhar Singh.

1) Managing Project
2) Dataset Labelling from 0 to 2k
3) Code and implementation

B. Nilava Dutta.

1) Models
2) Dataset Labelling from 2k to 4k
3) Transfer Learning layer implementation

C. Rajvivek Sundriyal

1) Video to Frames conversion
2) Dataset Labelling from 4k to 6k
3) Synopsis of project

D. Pulkit Suri.

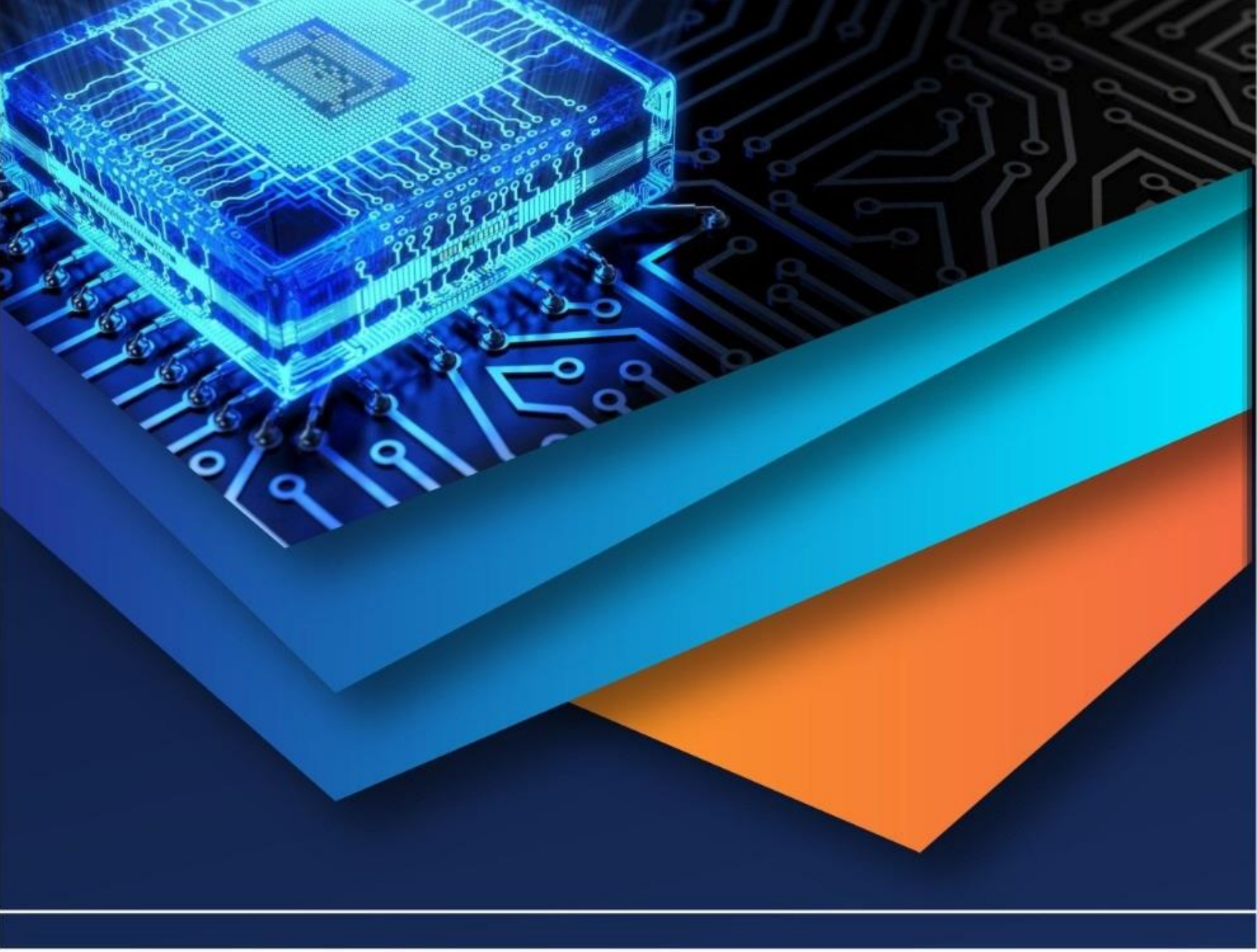
1) Report
2) Dataset Labelling from 6k to 8k
3) Graph implementation using matplotlib

X. ACKNOWLEDGMENT

We are really thankful to Assistant Professor Ms. Kanika Malik from the IMS Engineering College in Ghaziabad's Computer Science and Engineering department for his assistance in assisting us with the application of our research to the real world. Its our privilege to express our sincere regards to our project guide, Ms. Kanika Malik for his valuable inputs, able guidance, encouragement, cooperation and constructive criticism throughout the duration of our project. We sincerely thank the Project Assessment Committee members for their support and for enabling us to present the project on the topic.
"Scene Classification of Sports Video Using Transfer Learning."

REFERENCES

[1] Sharma, R.A.; Pramod Sankar, K.; Jawahar, C.V. Fine-grain annotation of cricket videos. In Proceedings of the 3rd IAPR Asian Conference on Pattern Recognition (ACPR 2015), Kuala Lumpur, Malaysia, 3–6 November 2015; pp. 421–425, doi:10.1109/ACPR.2015.7486538.
[2] Zhou, L.Sports video motion target detection and tracking based on hidden markov model. In Proceedings of the 17th International Conference on Scientometrics and Informetrics, Qiqihar, China, 28–29 April 2019; pp. 825–829, doi:10.1109/ICMTMA.2019.00186.
[3] Duan, Ling Yu; Xu, Min; Tian, Qi; Xu, Chang Sheng; Jin, Jesse S. A unified framework for semantic shot classification in sports video. IEEE Trans. Multimedia 2005, 7, 1066–1083.
[4] A. Kumar, J. Garg and A. Mukerjee Cricket activity detection. In Proceedings of the International Image Processing, Applications and Systems Conference, Sfax, Tunisia, 5–7 November 2014, pp. 1–6..
[5] Minhas, R.A.; Javed, A.; Irtaza, A.; Mahmood, M.T.; Joo, Y.B. Shot classification of field sports videos using AlexNet Convolutional Neural Network. Appl. Sci. 2019, 9(3), 483, doi:10.3390/app9030483.
[6] Russo, M.A.; Kurnianggoro, L.; Jo, K.H. Classification of sports videos with combination of deep learning models and transfer learning. In Proceedings of the 2nd International Conference on Electrical, Computer and Communication Engineering, Cox'sBazar, Bangladesh, 7–9 February 2019, doi:10.1109/ECACE.2019.8679371.
[7] Lee, J.; Koh, Y.; Yang, J. A deep learning based video classification system using multimodality correlation approach. In Proceedings of the International Conference on Control, Automation and Systems, Jeju, South Korea, 18–21 October 2017; pp. 2021–2025, doi:10.23919/ICCAS.2017.8204286.
[8] Sozykin, K.; Protasov, S.; Khan, A.; Hussain, R.; Lee, J. Multi-label class-imbalanced action recognition in hockey videos via 3D convolutional neural networks. In Proceedings of the 2018 IEEE/ACIS 19th International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing, Busan, South Korea, 27–29 June 2018; pp. 146–151, doi:10.1109/SNPD.2018.8441034.
[9] Bhalla, A.; Ahuja, A.; Pant, P.; Mittal, A. A Multimodal Approach for Automatic Cricket Video Summarization. In Proceedings of the 2019 6th International Conference on Signal Processing and Integrated Networks, Noida, India, 7–8 March 2019; pp. 146–150, doi:10.1109/SPIN.2019.8711625.
[10] Tejero-De-Pablos, A.; Nakashima, Y.; Sato, T.; Yokoya, N.; Linna, M.; Rahtu, E. Summarization of User-Generated Sports Video by Using Deep Action Recognition Features. IEEE Trans. Multimedia 2018, 20, 2000–2011, doi:10.1109/TMM.2018.2794265
[11] Karmaker, D.; Chowdhury, A.Z.; Miah, M.S.; Imran, M.A.; Rahman, M.H. Cricket shot classification using motion vector. In Proceedings of the 2015 2nd International Conference on Computing Technology and Information Management, Johor, Malaysia, 21–23 April 2015; pp. 125–129, doi:10.1109/ICCTIM.2015.7224605.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY