Deep Learning-Based Behavior Analysis in Basketball Video: A Spatiotemporal Approach

Jingyi Wang*
Department of Physical Education, Graduate School Kunshan National University Gunsan, South Korea, 54150

***Abstract*—The study of sports movement analysis technologies based on video has significant practical applications. Digital video footage, human-computer communication, as well as additional technologies can greatly improve the effectiveness of sports training. This research looks at the players' technical proficiency in a basketball contest footage and suggests a behaviour assessment technique inspired by the use of deep learning and attention mechanisms. First, we develop an approach for effortlessly obtaining the marking lines from the basketball arena and stadium. After that, the most significant frames of the footage have been shot using a spatial and temporal ranking technique. Next, we design a behaviour comprehension and prediction technique by implementing an autoencoder design. The results of the study may be sent to instructors and data scientists instantly to support them in determining their strategies and professional decisions. An extensive dataset of basketball films is used to test the proposed method. The outcomes demonstrate that the recommended attention mechanism-based strategy competently recognises the movement of video individuals while attaining substantial behavioural assessment efficiency.**

***Keywords*—Basketball; player movement analysis; player technique analysis; deep learning; attention mechanism**

I. Introduction

Today's Olympic Games are more than just an athletic competition. It has developed into an extensive competition between countries for technical advancement. The breaking of several records during the Olympics represents both technological advancements in sports and human progress beyond physiological limits. As a result, the cross-disciplinary field of sports science has been receiving more attention. Sports science includes biomechanics, sports medicine, and computer science. The primary goal is to raise athletes' competitive skill sets. Sports scientists operate in two domains: (1) physiology, health, and medical sports experts test players to ensure that the training regimen is efficient; and (2) experts apply contemporary technological innovations to relevant game studies by developing a range of supportive training aids.

Regarding computer and engineering technology, it is necessary to automatically gather a number of technical parameters during athletes' training in order to enhance scientific and technological analysis in sports training. The conventional approach includes sensors for the athletes. The disadvantage of this technique is that the athlete's performance during competition might be impacted by the sensor. Activities like sports professionals agree that the application of multimedia analysis in activities may greatly improve training efficiency due to the rising popularity of video capture equipment and the ongoing advancement of computer vision technologies in recent decades. In sports training, digital video technology was used [1] to record the training and competing procedure using a camera and automatically evaluate data on the athletes' postures, movements, etc. In order to achieve a type of human-computer interaction (HCI), the analytic findings are presented to the coaches and players in an understandable manner [2]. This may significantly reduce the possibility of injuries to players while also allowing coaches and athletes to accomplish the goals of quick feedback and intuitive instruction.

Unlike the traditional approach of affixing detectors to the sportsman's body, technological video-based activities training gear functions as a wireless method that is conducted without causing any discomfort to players and can instantly collect the most precise data regarding their activity postures. Thus, it has tremendous significance and a wide range of potential applications for raising athletes' training effectiveness and level of competition. Human-computer interface and sports action recognition are two examples of contemporary intelligent applications that make extensive use of human action analysis. Numerous action detection algorithms have been put out, and their results have been impressive. Ji et al. [3] developed a 3D CNN paradigm while using a standard CNN model to derive traits from 3D footage is not feasible. Another method used to identify human behaviour is to examine the joints in the human skeleton. Histograms of 3D joints are used by Xia et al. [4] to recognize human actions. An effective HCI assessment system, human mobility tech assessment, and player activity video assessment performance can all be achieved by using an advanced motions analysis approach employing the footage keyframe. The computer tracks the subject's action orientation and activity pattern in real-time when watching the action footage to determine the location and shape of an individual's body component. The computer then analyses the technical components of the move and informs the instructor or player of its results.

Basketball is a popular group sport with millions of supporters worldwide and widespread public affection. A competitive basketball video is used as the investigation's subject in this study, which also proposes an activity analysis method for analysing and predicting the players' movements, including dribbling, passing, shooting, and so on. Our suggested method's pipeline is illustrated in Fig. 1. In order to improve the ability to represent motion, we first developed a keyframe retrieval method for activity videos that rely on spatiotemporal characteristics. As can be observed in Fig. 2, the playing surface in the contest video will show up in the footage, therefore it has to be eliminated immediately to eliminate the auditorium's distraction before the player's position is tracked. As a result, the range of potential regions for player monitoring in the future might be decreased. Finally, a CNN-

[*] Corresponding authors.

RNN framework is used to classify the player's behaviour based on the video keyframe feature sequence. In conclusion, the paper's primary innovations consist of the following:

- The study uses a spatiotemporal ranking scheme to identify keyframes in video content, focusing on their stability, meaningfulness, and ability to be distinguished over time.
- A clustering-based technique is used to isolate the court area and remove auditorium disturbances, narrowing potential region ranges for future player monitoring. The starting cluster variety and cluster center are chosen based on trait disparities of the visual color histogram optimum point, reducing tolerance to original group numbers and center while increasing precision and effectiveness. The straight line is fitted using the least-squares approach, and the line parameter is adjusted for camera tuning.
- A comprehensive analysis of players' behavior is conducted using an encoder and decoder-based design, which improves location and movement prediction accuracy.

The following sections are organized as follows. Section II provides an overview of relevant work. The suggested methods are thoroughly discussed in Section III. Section IV presents the experiment's results along with a thorough analysis. Section V provides a summary of the paper's conclusion and future directions.

II. Related Work

A. Human Movement Analysis

Deep learning has already been applied in recognition domains such as the classification of images [5], [6], face recognition [7], and human location prediction [8]. Since video character motion recognition may be thought of as a long-term picture classification issue, research on video character motion recognition also frequently uses the deep learning approach to picture identification [9], [10], [11]. When it comes to motion recognition, convolutional neural networks (CNN) are not as common as they are in other areas of computer vision. It has always been challenging to utilise CNN to identify human movements in a video. CNNs are more appropriate for extracting characteristics from just one still picture because they are less susceptible to chronological data. However, the development of CNN for stationary images has greatly facilitated the progress of image recognition. Many CNN designs have already been created lately that enable CNN to use visual time-series data to some degree. The paper claims that there are methods for altering CNN input such that its initial layer can pick up the footage's spatiotemporal characteristics.

A predetermined number of sequential video frames is used as a CNN source in [12]. Amin et al. [13] proposed video frame sampled integration in several temporal realms, which was a step farther than the simple stacking of frames from videos in [12]. Late fusion combines the CNN's fully connected layer, which translates to a visual frame, with a predetermined temporal domain duration length; the initial fusion process is the same as that proposed in [12], and gradual fusion entails raising the network's input temporal domain duration tier by tier. It appears that the research technique does not completely employ the footage's chronological data because the reliability of the preceding video recognition strategy only slightly increases when contrasted with a particular spatial space CNN. A method based on the structure of 3D CNN is published in [14]. By expanding the original 2D network in the context of time-space, this design enables the system to learn the footage's attributes in the context of time. A 3D filter and many sequential video frames are used as input by this sort of framework to learn the spatial and temporal traits of the video.

Experimental results show that this framework outperforms visual frame fusion considering additional inputs, although it is more complicated and requires additional facilities for both training and evaluation. Two parallel designs were presented in [15] as a space and time dual-flow topology to make use of the temporal features of the video. Additionally, the framework shows that the majority of behaviors of characters in the UCF 101 database can be identified using only the optical flow insight. The two CNNs, individually, use a number of optical flow visuals of the footage's frame and the footage as their input. They subsequently combine each element of the data and gather time and spatial details regarding the subject's motion to identify the activity characteristic of the footage character. The identification accuracy remains low even though the structure partially exploits the video's temporal features.

B. Retrieval of Video Key Frames

The key frame extraction technique, which is extensively utilized in motion capture, human behaviour, and motion identification, is a hotspot for pattern recognition research [12]. Nevertheless, no universal keyframe technique has been identified since motion video is extremely complicated and nonlinear. The authors of [16] select an important frame set with a high limiting rate by setting a specific threshold based on a comparison of its entropy measurements of colour histograms within the nearest footage frames. Although the threshold needs to be preset, it is easy to achieve overlap or missing keyframes, which results in limited flexibility when the movement of objects in the video shifts substantially. In [17], the complex K-means clustering based on its kernel and neighbourhood data is used to continually filter the keyframes and optimise the picture's noise and clarity.

However, as there are currently no space-time constraints, the selected chronological frame sets possess a lower potential for temporal representation, making them unsuitable for real-time HCI. In [18], the footage is separated into moving objects and a changing background. The unsupervised clustering approach analyzes the object's movement and shifts in shape to identify the keyframes. The retrieved keyframe sets are short, the motion properties are well-defined, and they might potentially meet live video recognition criteria since the source video is analysed and understood at the stage of semantics. When using the key extraction methods described in [18], it is frequently necessary to develop the object recognition and kinematic trait descriptor algorithm in line with the usage backdrop. Determining how to construct an individual's motion framework for the transition video is so crucial.

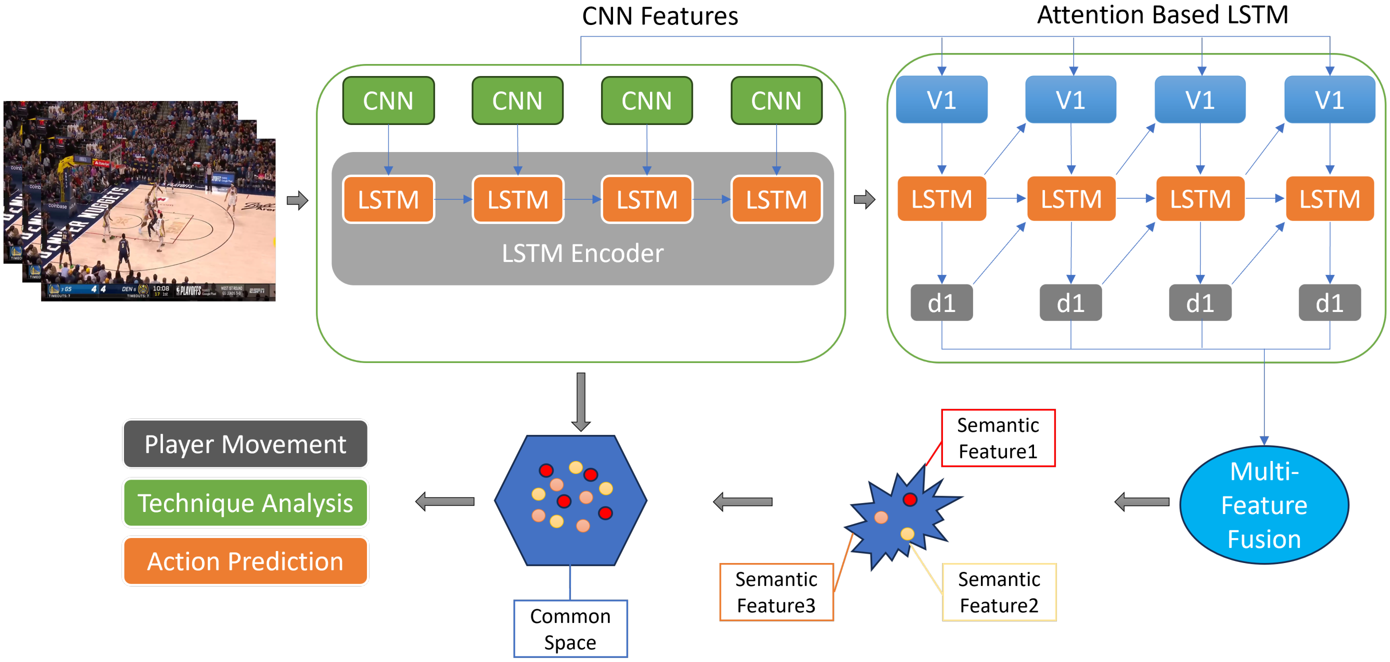

Fig. 1. The Systematic pipeline of the proposed framework.

In order to collect motion features in both time and space from multiple successive footage frames, the researchers of [12] use a 3D CNN and select key frames using multiple channels insight fusion. This framework is better at accurately detecting physique motion and maintaining both the temporal and location components of individual motion. The individual's activity feature approach is not suitable for activity footage keyframe retrieval because the motion features vary greatly as a result of the motion footage's routine mistakes or irregular movements. To find the image's human-sized bounding box, the authors of [19] employ the histogram of gradients (HoG) human body classifiers. In order to determine the crucial frame, they use an anatomical template to divide the individual physique bounding box into 16 different weighted motion areas. Then, they evaluate the relative motion pattern difference within each motion patch. Physique template size restrictions reduce critical frame retrieval accuracy in complex situations or whenever the degree of arena changes dramatically, and they also make it easy to cause errors in human physique motion detection.

III. Methodology

The proposed framework's structured pipeline, which is illustrated in Fig. 1, describes the sequential method for evaluating and estimating player behaviour in basketball footage. This pipeline addresses the difficulties of player motion analysis in games videos by combining artificial intelligence-based behaviour prediction, keyframe extraction, and spatiotemporal analysis in a coherent manner.

A. Extraction of Court and Marking Lines

The present work uses the extraction of court and identifying lines as its initial research challenge. On the contrary, it can efficiently filter out the presence of the audience outside the perimeter of the court and reduce the number of computations for player tracking that follows; conversely, the efficiency of the extraction will have an impact on the players' behaviour prediction. To divide up the court area, we decide to use the K-means clustering technique. Initially, the trait difference of the visual component, the colour histogram optimum point, is computed in order to choose the starting cluster size and the cluster centre. Following the mean clustering technique's segmentation of the visuals, the estimated court area is calculated based on the pertinent judgment criteria. The full-court space and free-throw box are then obtained using morphology. The marker lines in the acquired greyscale picture of the court area are segmented using the edge detection function, and the trajectory characteristics of the court line are extracted using the method of the Hough transformation. The resultant line characteristic is then adjusted for further camera calibration once the line is calculated using the least-squares approach. Both Fig. 3 and 4 illustrate the impact of the marker line and court detection algorithms.

B. Motion Video Key Frame Extraction

The spatial as well as temporal impact of every frame in the movie is predicted using spatial and temporal attention methods. The implications of every scene are then obtained by fusing these significant scores. Here, we use the sparse weighting feature W in our technique to represent the significance of each frame. Individuals often focus on regions that contrast more in terms of time and space. The spatial attention algorithm's primary goal is to locate every object in every scenario. The temporal attention algorithm's primary job is to identify the motion-rich regions of the footage. As a result, both of these approaches may readily replicate the significance of human vision for each media frame. We

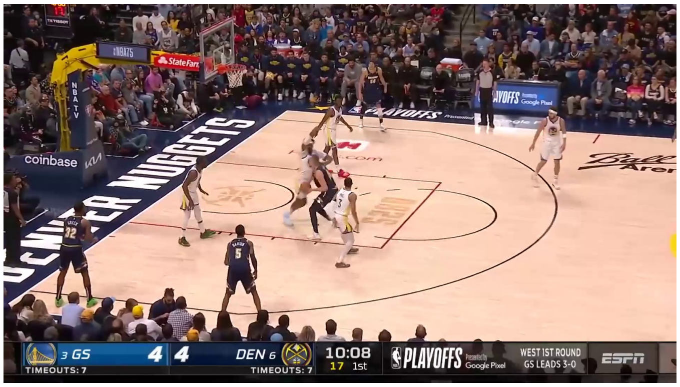

Fig. 2. A Scene taken from a footage of a basketball match.

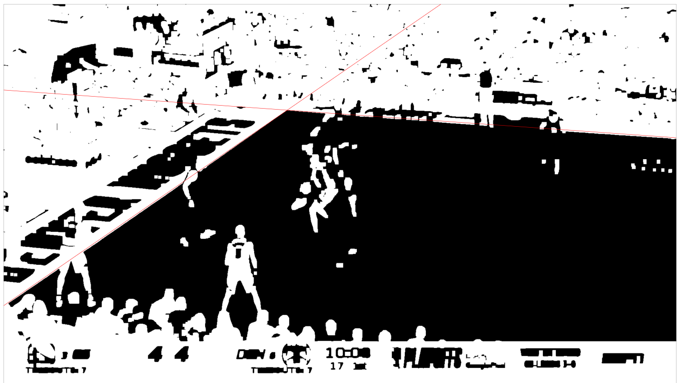

Fig. 3. A diagram showing the arena and mark line that were discovered.

provide a spatial attention system that uses image descriptors. This technique, which might be employed to determine an image's visual saliency, is known as a picture signal. This procedure is determined by the sign function of the specific cosine transform, as seen in [20]. It has been demonstrated that the image's signal method can roughly identify the foreground items in the image. First, we enlarge the frame to 64×48 for a certain frame f in the clip. Each colour component in the image is defined by the image signal procedure as follows:

$$IS(f_c) = \text{sign}(\text{DCT}(f_c)) \tag{1}$$

where f_c is the colour vector of frame f, DCT is the specific cosine transform functionality, and the sign expression is dependent on the metaphorical operation that follows the input. The reconstructed image f'_c in the spatial realm is then projected to the converted signal after it has undergone an inverse offline cosine transformation:

$$f'_c = \text{IDCT}(IS(f_c)) \tag{2}$$

The following formula is used to determine the resulting static feature map $S(f)$:

$$S(f) = G \times \sum_c f'_c \odot f'_c \tag{3}$$

where the Gaussian kernel is represented by G, the operation of convolution by $\times$, and the Hadamard product operator by $\odot$. We normalise every score in $S(f)$ to [0,1] by dividing it by the highest value after generating the static feature map $S(f)$. For every frame$f$, the static attentive weight A_S is determined by taking a mean of the non-zero items in $S(f)$. If the image frame f has a static attentive weight A_S value around 1, it is deemed noteworthy. In contrast, a frame f is deemed insignificant if its value is around 0.

Researchers integrate many attention values in numerous algorithms using a linear fusion approach, which produces a unified attention result [21], [22]. In the event when n attentive values need to be merged, the linear fusion method's general structure looks like this:

$$A_L = \sum_{i=1}^{n} w_i A_i, \quad \sum_{i=1}^{n} w_i = 1 \tag{4}$$

while A_L is the attentive value following the linear's merging of the various attention outcomes, and w_i is the weighting of the attentive value A_i. The above-mentioned spatial and temporal weights are then fused as a sparse weight W using a nonlinear fusion approach. The temporal feature map $TS(f)$ of the image frame f determines the weight value:

$$w_T = \alpha e^{1-\alpha} \tag{5}$$

$$\alpha = \max(TS(f)) - \min(TS(f)) \tag{6}$$

$$w_S = 1 - w_T \tag{7}$$

where the spatial significance weight is denoted by w_S. A greater alpha value corresponds to a larger weight of the temporal attention weight w_T of the image frame f if the temporal feature map $TS(f)$ contains significant activity facts, and vice versa. For example, let A_S represent spatial attentive weights and A_T represent temporal attentive weights. We declare $w = [w_S, w_T]$ along with $A = [A_S, A_T]$. The subsequent nonlinear fusion approach allows us to obtain the resulting sparse weight W:

$$W = \frac{w \cdot A + 1}{2(1+\rho)} \left(\|2w_S A_S - w \cdot A\| + \|2w_T A_T - w \cdot A\| \right) W_D \tag{8}$$

$$W_D = 1 + \frac{1}{2(1+\rho)} \left(\|1 - 2w_S\| + \|1 - 2w_T\| \right) \tag{9}$$

In this case, the weight's relevance in the attention weight fusion mechanism is represented by the specified constant ρ.

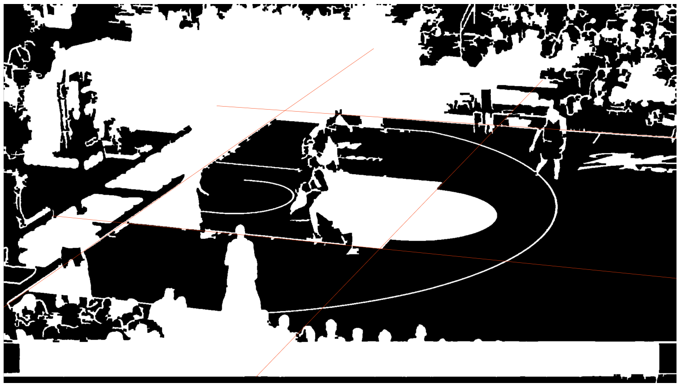

Fig. 4. An example of the identified free-throw basket.

C. Player Behavior Analysis and Prediction

We use an encoder-decoder architecture to develop an approach for analysing and predicting player behaviour. A specific source video x is encoded into an uninterrupted map of space determined by the encoder structure ϕ_E:

$$V = \{v_1, \ldots, v_N\} = \phi_E(x) \tag{10}$$

where the convolutional neural network (CNN) is represented by ϕ_E. There are a total of N keyframe vectors of traits, and the ith frame's M-dimensional pattern vector is $v_i \in \mathbb{R}^M$. To convert the video characteristics into a vector, we choose LSTM for our decoder system ϕ_D:

$$(h_t, z_t) = \phi_D(y_t, h_{t-1}, V) \tag{11}$$

In this case, the LSTM adjust its concealed state. h_t by using the feature V, the present input y_t, and the prior concealed state h_{t-1}. We add a method of attention to the LSTM foundation to enhance action recognition efficiency:

$$i_t = \sigma(W_i y_t + U_i h_{t-1} + A_i c_t + b_i) \tag{12}$$
$$f_t = \sigma(W_f y_t + U_f h_{t-1} + A_f c_t + b_f) \tag{13}$$
$$o_t = \sigma(W_o y_t + U_o h_{t-1} + A_o c_t + b_o) \tag{14}$$
$$g_t = \tanh(W_g y_t + U_g h_{t-1} + A_g c_t + b_g) \tag{15}$$
$$m_t = f_t \odot m_{t-1} + i_t \odot g_t \tag{16}$$
$$h_t = o_t \odot \tanh(m_t) \tag{17}$$

The parameters that need to be learnt by LSTM are represented by W, U, A, and b. The input data used in LSTM at every step t is represented by y_t, the function used for Sigmoid activation is represented by σ, and the context vector is represented by c_t. An essential component is context vector facts. A straightforward method for addressing the unpredictability in video length is to average all video features, then enter the resulting vector into the framework at each point in time:

$$c_t = \frac{1}{n}\sum_{i=1}^{n} v_i \tag{18}$$

The internal temporal framework of the video is ignored by this technique, which leads to information loss even if it successfully condenses all of the important frame data into a single vector. To facilitate motion detection, our approach uses global temporal knowledge in order to intelligently focus on a subset of the video's important frames throughout the entire decoding procedure. The model avoids mixing several events across the whole video segment by just taking into account a section of the media sequence, allowing it to distinguish objects and activities throughout the stream. Our method also enables the system to concentrate on the video's most important components, which may be many critical frames in a row. Weights are dynamically added to each frame characteristic in the video:

$$c_t = \frac{1}{n}\sum_{i=1}^{n} a_i^t v_i \tag{19}$$

where

$$\sum_{i=1}^{n} a_i^t = 1. \tag{20}$$

The importance of attention value at time t, a_i^t, must be determined at each stage of the LSTM decoders. The attention value a_i^t represents the correlation value of the ith frame characteristic in the source video, given all the recognised movements, such as $\{z_1, \ldots, z_{t-1}\}$. In order to decode the prior concealed state h_{t-1} within the LSTM, we create a function. To calculate the un-normalised results, this concealed state concurrently receives the clip frame characteristics V and h_{t-1} summarises all of the prior motions:

$$\epsilon_t = w^T \tanh(W_a h_{t-1} + U_a V + b_a) \tag{21}$$

In the decoding procedure. w^T, W_a, U_a, and b_a are learnt alongside the LSTM attributes. The significance weight is obtained by normalising the appropriate score ϵ_t.

$$\alpha_t = \text{softmax}(\epsilon_t) \tag{22}$$

The attention mechanism is the method by which the pertinent score and attentive weight are determined. By raising the keen weight of the pertinent frame throughout the decoding procedure, the system of attention only pays tribute to partial frame details in the entire video. Nevertheless, we allow the attention method to comprehend the temporal pattern in the movie through the LSTM, rather than overtly forcing the option to concentrate on a certain portion of the content. In conclusion, Algorithm 1 may be used to characterise the suggested approach.

IV. Experimental Results and Discussion

In this article, we employ a basketball media dataset to test human movement recognition.

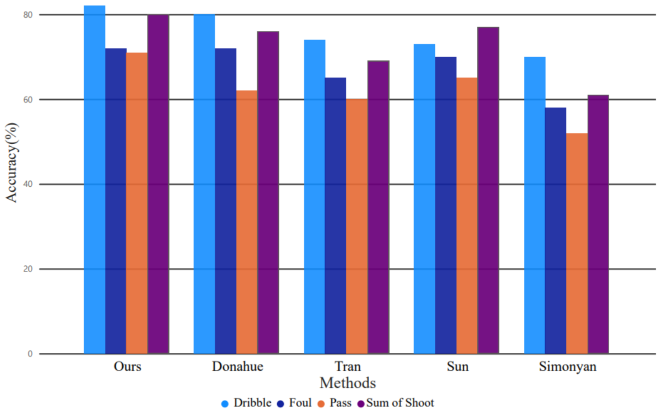

Fig. 5. The evaluation assessed the test set's recognition efficiency for various approaches' motions.

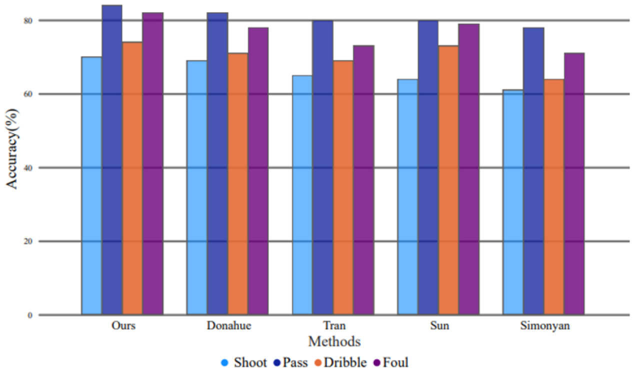

Fig. 6. The precision of predicting the motions of various techniques on a given set.

TABLE I. An Explanation of the Action Recognition Database for Basketball Players

Motion type	Foul	Dribble	Pass	Shoot
Number/Training set	649	711	1038	924
Number/Valid set	147	191	227	208
Number/Testing set	153	172	261	214

A. Dataset Characteristics

This section describes the specific traits and attributes of the collection of data that was utilized in this investigation.

```
Algorithm 1 Analysis of Human Behaviour in Relation to Volleyball Videos
Require: A Volleyball video
Ensure: Player mobility, approach evaluation, and action prediction
1: Using the Hough transformation and K-means clustering, extract the court and marker lines.
2: Utilising the created image signal, calculate visual saliency and take out important video frames.
3: Create an encoder-decoder framework-based system for analysing and predicting player behaviour.
```

TABLE II. Outcomes From the Professional Basketball Player Action Recognition Dataset

Motion type	Foul (%)	Dribble (%)	Pass (%)	Shoot (%)
Accuracy/Valid set	80	90	75	84
Accuracy/Testing set	74	85	70	82

TABLE III. The Activity Recognition Database of Basketball Players Yielded the Activity Prediction Outcome

Motion type	Foul (%)	Dribble (%)	Pass (%)	Shoot (%)
Accuracy/Valid set	77	87	73	83
Accuracy/Testing set	72	83	71	80

There are 10,311 video clips in the dataset that were taken from 51 NBA basketball games that were televised by sports media. Cameras often used in sports coverage are used to record all videos through a third-person viewpoint [28]. The first step is to classify the footage videos into Four different action classes: Dribble, Foul, Pass, and Shoot. Fig. 5 shows how these action types are distributed. The experimental setup section presents a detailed experimental investigation of these groups. To maintain uniformity in quality and frame rate, all video samples have been standardized and converted to RGB format. Furthermore, the same models and parameters that were used to process the RGB dataset were also used to analyze an optical flow dataset. Every clip is labeled using a specified nomenclature that contains the title, video number, and timestamp, which indicates the start time of the accompanying shot, to enable appropriate experimentation. Table I summarizes how many of each kind of movement there are in each set:

B. Deep Learning Training

The following section presents our employed encoder–decoder-based action analysis and prediction system's training state. In our method, every frame is enlarged to 64 × 48 before being input into the feature extraction architecture that has been created. We make use of the deep learning framework Caffe. To train the system's parameters, we employ the SGD gradient descent technique. Specifically, after 10,000 steps, we raise the learning rate from 0.1 - 0.001. The framework is trained until the training loss converges, with the momentum value set to 0.9 and the weighted decay set at 0.0002.

C. Analysis and Comparison

1) Recognition accuracy: The findings from the tests for activity detection and prediction are shown in Tables II and III. The degree to which the individual's expected next action and the ground truth coincide is known as the prediction accuracy. Tables II and III demonstrate that the proposed method has a success rate of over 80% in predicting and recognising shot and dribble actions. Nevertheless, the accuracy of pass/foul recognition and prediction is lower. Furthermore, it is noted that the evaluation set's identification and prediction precision are somewhat worse than the valid set's, indicating that

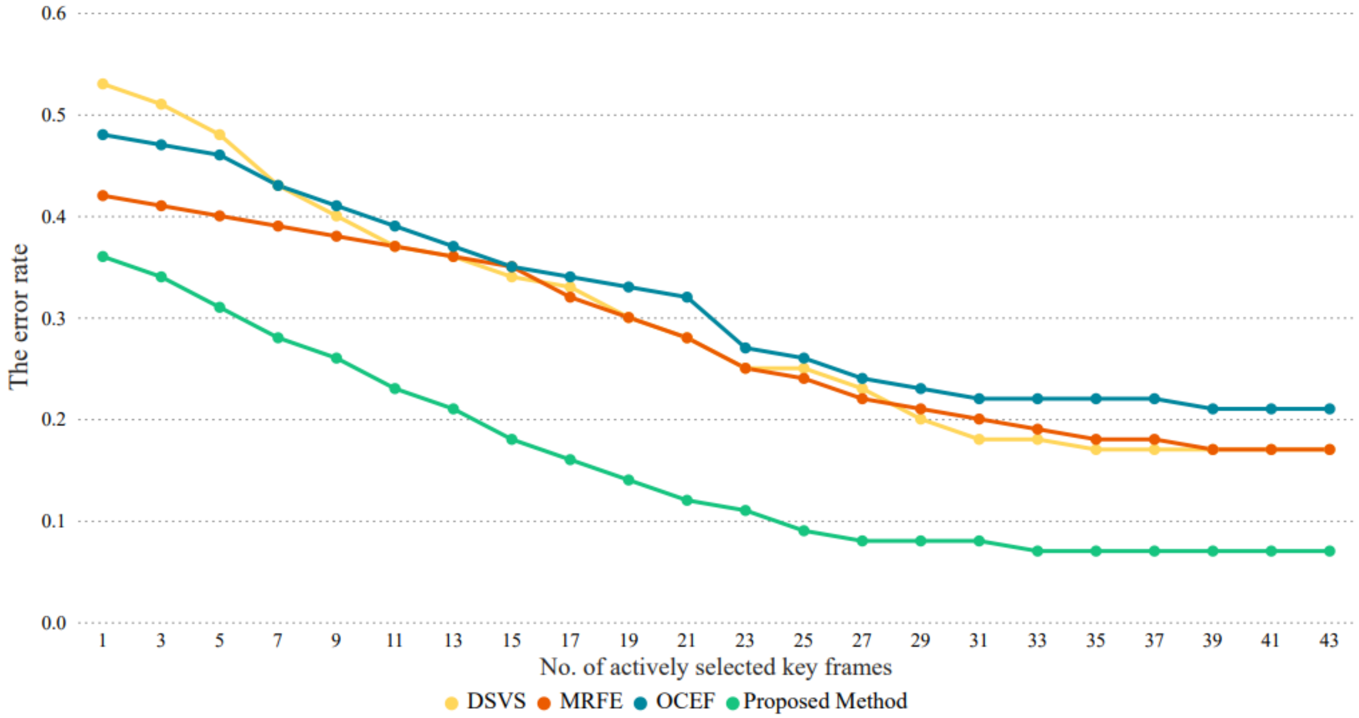

Fig. 7. The rate of error for various key frame grabbing techniques.

more distinct training examples might improve the model's efficiency even more. The proposed method was also contrasted with several other methods that were previously used to analyse the dataset, including Simonyan et al. [15], Tran et al. [14], Donahue et al. [23], and Sun et al. [24]. All of these methods are based on advanced neural networks. Fig. 5 and Fig. 6 show the outcomes of several methods for recognising and predicting the four movements on the test set. It is shown that the approach recommended in this study can more accurately recognise a player's mobility in the basketball footage than previous methods. Additionally, it makes more accurate predictions about the individual's next act than earlier methods. In summary, when compared to contemporary methods, our approach has shown improved accuracy results for both movement recognition and future movement prediction.

2) Contrast of key frame selections: By employing three well-known keyframe selection methods, online clustering key frame selection (OCFE) [25], motion-based crucial frame selection (MKFE) [26], and dictionary-based valuable frame selection (DSVS) [27], we test the efficacy of the approach we propose experimentally. The initial strategy groups a large number of frames into many centres using the K-means strategy. These centres are effortlessly used to categorise the remaining frames. The MKFE approach, which produces an action descriptor, focuses primarily on the dynamics of the subjects in the media clip. The DSVS method guides the picking of important frames by turning a clip into a dictionary by using sparsity constancy. In Fig. 7, the outcome contrast across multiple keyframe selection techniques is displayed. The error rate is a measure that we use to quantitatively assess the effectiveness. The difference between the chosen clip frames and the ground truth, which is determined by trained video experts, is measured in this particular case by the error rate. It is clear that the key frame identification approach we created works best because our strategy has the smallest error rate compared to other methods.

D. An Examination of Key Parameters

We do experiments to examine the important factors. Action analysis in our work relies heavily on the selection of important frames since human activities in basketball recordings may be correctly and effectively reflected in a variety of representative video frames. The weight's relevance is represented by the preset constant ρ in Eq. (6). Action analysis and prediction are impacted by the performance of key frame selection, which is influenced by the value of ρ. As a consequence, we compare the results under various ρ parameters. Since our dataset contains four basketball activities, we use each ρ value to evaluate the recognition rate of four actions. The final result is shown in Fig. 8. The best recognition accuracy, 76.5%, can be achieved by setting $\rho = 0.5$, which is 0.5% greater than setting $\rho = 0.4$, according to the average value.

V. Conclusion and Future Direction

In the field of computer vision, human activity detection has been a popular study area. Instructors and data scientists may be able to quickly determine the health of the athlete through human-computer interaction with the use of automated human motion capture and identification from athletic sporting

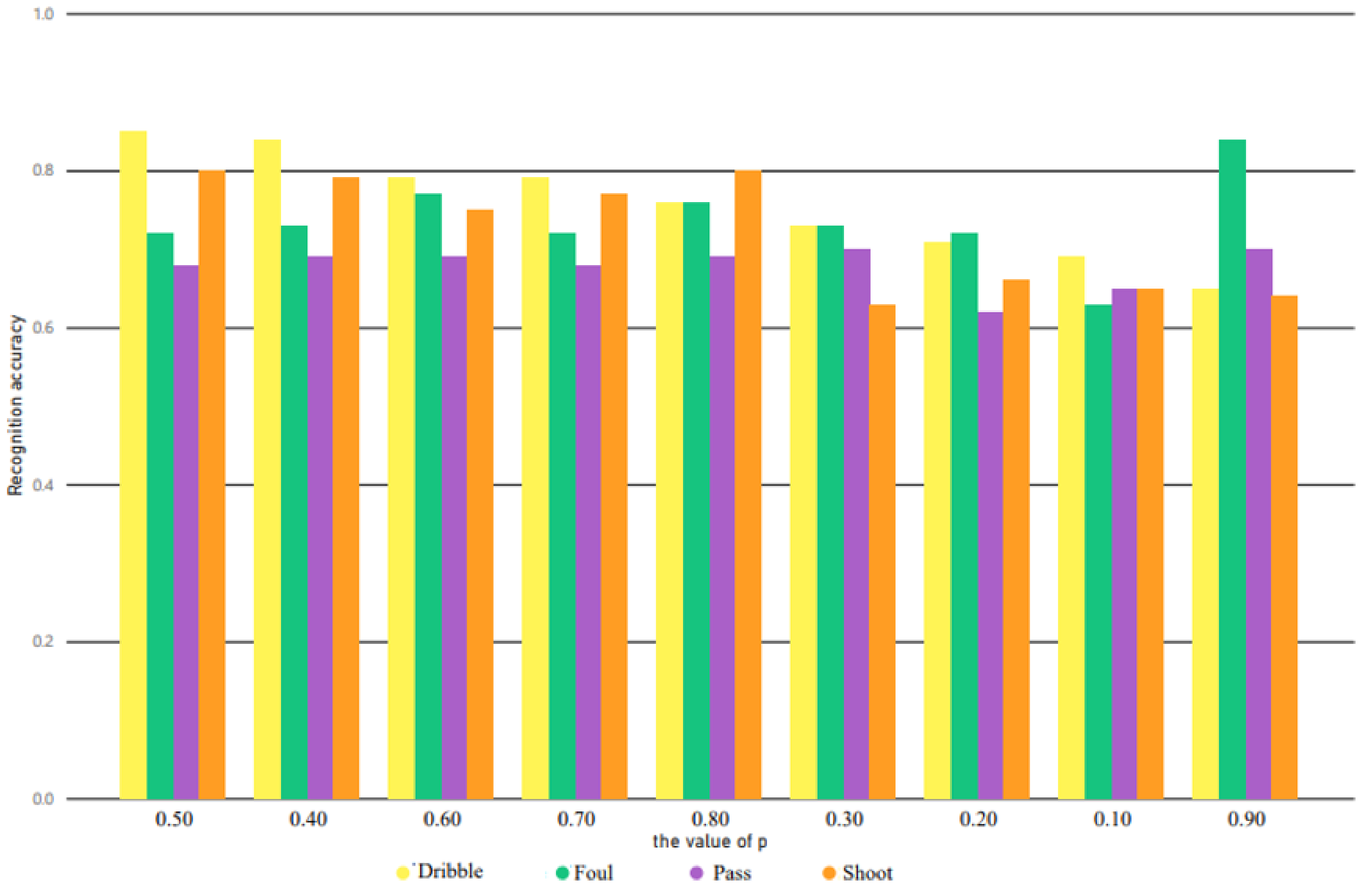

Fig. 8. The precision of behaviour recognition under various parameter conditions as determined by experiment.

movies. The technical traits of basketball players in sporting events are examined in this research, which also suggests a practical technique for movement identification and predictions in basketball movies. To extract important footage frames based on their corresponding value, a spatiotemporal ranking technique has been presented. The basketball field and marked line are then identified to remove any ambiguity in athlete tracking and positioning. Lastly, an encoder-decoder architecture is created for predicting and identifying player movements. Trainers and data scientists may use the analysis findings in real-time to assist them in analysing the technical decisions and approaches. The suggested approach is tested on a big sample of basketball videos. The findings demonstrate that the suggested approach can accurately and successfully identify player motions in-game footage.

Future research aims to expand the applicability of a proposed method in sports beyond basketball, such as soccer, volleyball, and tennis. Testing the method on publicly available action recognition datasets and incorporating multimodal data could improve prediction accuracy. Optimizing the framework for real-world scenarios, enhancing attention mechanisms, and addressing class imbalance are also areas of focus. Implementing the framework in real-time applications could enhance its practical utility for coaches and athletes. These directions aim to strengthen the versatility, accuracy, and applicability of the proposed approach.

References

[1] D. Yow, B.-L. Yeo, M. Yeung, B. Liu, *Analysis and presentation of soccer highlights from digital video*, in: Proc. ACCV, Vol. 95, 1995, pp. 11-20.

[2] F. Quek, D. McNeill, R. Bryll, S. Duncan, X.-F. Ma, C. Kirbas, K. E. McCullough, R. Ansari, *Multimodal human discourse: gesture and speech*, *ACM Trans. Comput.-Hum. Interact.*, vol. 9, no. 3, 2002, pp. 171–193.

[3] S. Ji, et al., *3D convolutional neural networks for human action recognition*, *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 1, 2012, pp. 221–231.

[4] L. Xia, C.-C. Chen, J. K. Aggarwal, *View invariant human action recognition using histograms of 3d joints*, in: 2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, IEEE, 2012, pp. 20–27.

[5] H. Xiong, W. Yu, X. Yang, M. N. S. Swamy, Q. Yu, *Learning the conformal transformation kernel for image recognition*, *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 28, no. 1, 2015, pp. 149–163.

[6] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich, *Going deeper with convolutions*, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 1-9.

[7] F. Schroff, D. Kalenichenko, J. Philbin, *Facenet: A unified embedding for face recognition and clustering*, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 815-823.

[8] J. Tompson, R. Goroshin, A. Jain, Y. LeCun, C. Bregler, *Efficient object localization using convolutional networks*, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 648-656.

[9] J. Zhang, Y. Han, J. Tang, Q. Hu, J. Jiang, *Semi-supervised image-to-video adaptation for video action recognition*, *IEEE Trans. Cybern.*, vol. 47, no. 4, 2016, pp. 960–973.

[10] S. Ul Amin, M. Ullah, M. Sajjad, F. A. Cheikh, M. Hijji, A. Hijji, K. Muhammad, *EADN: An Efficient Deep Learning Model for Anomaly Detection in Videos*, *Mathematics*, vol. 10, no. 9, 2022, p. 1555. doi:10.3390/math10091555.

[11] F. Husain, B. Dellen, C. Torras, *Action recognition based on efficient deep feature learning in the spatio-temporal domain*, *IEEE Robot. Autom. Lett.*, vol. 1, no. 2, 2016, pp. 984–991.

[12] S. Ji, W. Xu, M. Yang, K. Yu, *3D convolutional neural networks for human action recognition*, *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 1, 2012, pp. 221–231.

[13] S. Ul Amin, B. Kim, Y. Jung, S. Seo, S. Park, *Video Anomaly Detection Utilizing Efficient Spatiotemporal Feature Fusion with 3D Convolutions and Long Short-Term Memory Modules*, *Adv. Intell. Syst.*, vol. 6, no. 7, 2024, p. 2300706. doi:10.1002/aisy.202300706.

[14] D. Tran, L. Bourdev, R. Fergus, L. Torresani, M. Paluri, *Learning spatiotemporal features with 3d convolutional networks*, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 4489-4497.

[15] K. Simonyan, A. Zisserman, *Two-stream convolutional networks for action recognition in videos*, in: Advances in Neural Information Processing Systems, 2014, pp. 568–576.

[16] S. Ul Amin, Y. Kim, I. Sami, S. Park, S. Seo, *An Efficient Attention-Based Strategy for Anomaly Detection in Surveillance Video*, *Comput. Syst. Sci. Eng.*, vol. 46, no. 3, 2023.

[17] S. Wang, D. I. Lan, J. Liang, *Multi-dimensional fuzzy clustering image segmentation algorithm based on kernel metric and local information*, *Electron. Lett.*, vol. 51, 2015, pp. 693–695.

[18] N. J. Janwe, K. K. Bhoyar, *Video key-frame extraction using unsupervised clustering and mutual comparison*, *Int. J. Image Process. (IJIP)*, vol. 10, no. 2, 2016, pp. 73–84.

[19] P. A. N. G. Ya-jun, *Key frames extraction of motion video based on prior knowledge*, *J. Henan Polytech. Univ. (Natl. Sci.)*, vol. 35, no. 6, 2016, pp. 862–868.

[20] X. Hou, J. Harel, C. Koch, *Image signature: Highlighting sparse salient regions*, *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 1, 2011, pp. 194–201.

[21] G. Guan, Z. Wang, S. Lu, J. D. Deng, D. D. Feng, *Keypoint-based keyframe selection*, *IEEE Trans. Circuits Syst. Video Technol.*, vol. 23, no. 4, 2012, pp. 729–734.

[22] S. Chakraborty, O. Tickoo, R. Iyer, *Adaptive keyframe selection for video summarization*, in: 2015 IEEE Winter Conference on Applications of Computer Vision, IEEE, 2015, pp. 702–709.

[23] J. Donahue, L. A. Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, T. Darrell, *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.

[24] L. Sun, K. Jia, D.-Y. Yeung, B. E. Shi, *Human action recognition using factorized spatio-temporal convolutional networks*, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 4597-4605.

[25] A. Bouguettaya, *On-line clustering*, *IEEE Trans. Knowl. Data Eng.*, vol. 8, no. 2, 1996, pp. 333–339.

[26] J. Luo, C. Papin, K. Costello, *Towards extracting semantically meaningful key frames from personal video clips: From humans to computers*, *IEEE Trans. Circuits Syst. Video Technol.*, vol. 19, no. 2, 2009, pp. 289–301.

[27] Y. Cong, J. Yuan, J. Luo, *Towards scalable summarization of consumer videos via sparse dictionary selection*, *IEEE Trans. Multimedia*, vol. 14, no. 1, 2012, pp. 66–75.

[28] S. R. Shakya, C. Zhang, Z. Zhou, *Basketball-51: A video dataset for activity recognition in the basketball game*, *CS & IT Conference Proceedings*, vol. 11, no. 7, 2021.