Tuijin Jishu/Journal of Propulsion Technology
ISSN: 1001-4055
Vol. 45 No. 1 (2024)

Advancing Activity Recognition in Tennis: Employing Bag of Words Approach for Enhanced Video Analysis

Shama P. S[1], Prakash Pattan[2]

[1]*Research Scholar, PDACE, KalaburagiPDACE, Kalaburagi, Karnataka, India*

[2] *Assoc. Prof. Dept. of Computer Science & EngineeringPDACE, Kalaburagi,Karnataka, India*

Abstract: -The study presents a novel video representation technique for activity recognition, focusing on modeling video dynamics with activity attributes. The video sequence is divided into short-term segments characterized by its dynamic features. These segments are then represented using a dictionary of attribute dynamics templates based on a generative model known as the binary dynamic system (BDS). The process involves learning the dictionary of BDSs from a training dataset and quantizing attribute sequences extracted from videos into BDS codewords, resulting in a histogram known as the bag-of-words for attribute dynamics (BoWAD). Extensive experimental evaluation demonstrates the superiority of the BoWAD representation compared to other state-of-the-art methods in capturing temporal structure for complex activity recognition in videos. The proposed approach offers a robust and effective means to model video dynamics, thereby enhancing the accuracy and performance of activity recognition systems. The experimental analysis highlighted the impressive performance of the proposed approach in accurately identifying the tennis events' Bounce, Net, and Hit. The model achieved outstanding accuracy (87.92%), recall (92.08%), and precision (87.92%).

Keywords*: Activity Recognition, Attributes of Activities, Short-Term Segments, Dictionary Of Attribute Dynamics Templates, Binary Dynamic System (BDS), Learning Dictionary, Quantizing Attribute Sequences, Bag-Of-Words For Attribute Dynamics (Bowad), Temporal Structure Modelling.*

1. Introduction

Recognizing human activities and events poses a crucial challenge in computer vision research. Two main research directions have gained significant attention in this area. The first direction focuses on modelling the temporal composition of activities, which involves using low-level video representations. Various methods have been proposed to model the temporal structure of low-level features extracted from videos, including both discriminative and generative models [1]. The second direction represents activities as collections of semantic attributes, enabling a higher level of abstraction where features denote occurrences of semantic concepts like scene types, actions, and objects [2]. This intermediate representation fosters better generalization, facilitates semantic reasoning, and allows knowledge transfer across different instances.

Figure 1 illustrates the challenges in modelling the dynamics of attributes for complex activities, using the example of a "tennis-serve" activity and its associated trajectory [3]. Figure 1 illustrates the complexities in capturing the dynamics of attributes related to intricate activities. The top section depicts a "tennis-serve" activity. In contrast, the bottom section shows the corresponding trajectory, color-coded for different motions: red for "arm motion," green for "foot motion," and blue for "ball motion." It's important to note the intricate nature of the trajectory and that only a brief segment (highlighted in red) is crucial to the action of interest.

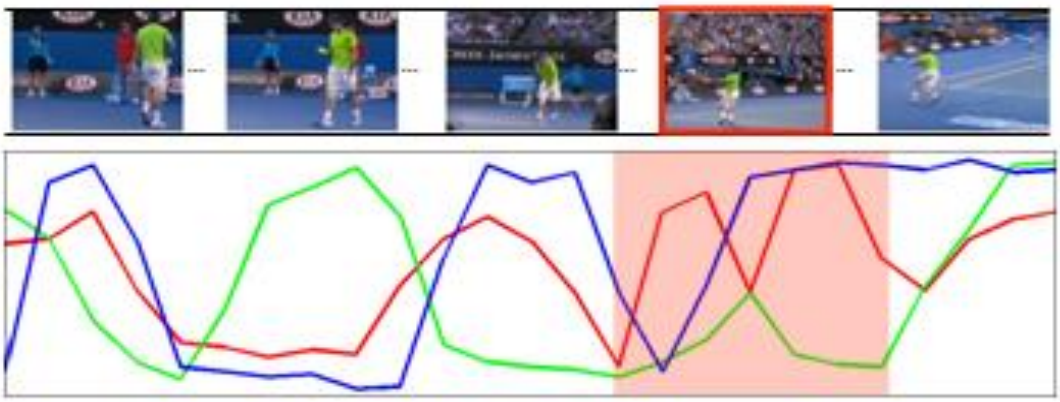


Fig. 1. Challenges In Modeling The Dynamics Of Attributes Of Complex Activities.(Top)"Tennis-Serve" Activity. (Bottom)Associated Trajectory (Red For "Arm Motion," Green For "Foot Motion," And Blue For "Ball Motion").

While both research directions have their merits, they also face limitations in recognizing complex activities. For instance, solely characterizing the temporal structure based on low-level features may not suffice for such activities. On the other hand, representing videos as an order with a smaller set of attributes lacks fine-grained activity discrimination, as it cannot distinguish activities expressing similar characteristics in different orders. To address these challenges, recent work [3] proposed unifying both directions by modelling the temporal structure of video projection in attribute space using binary dynamic systems (BDS). While BDS achieved state-of-the-art performance, it still faces challenges, such as handling videos with multiple events of interest and fitting complex attribute space trajectories.

To overcome these limitations, this study introduces a novel video representation called the bag-of-words for attribute dynamics (BoWAD) [4]. BoWAD is an extension of the bag-of-visual words (BoVW) popular in image classification. Unlike BoVW, which utilizes visual appearance templates, BoWAD relies on attribute dynamics templates, specifically temporally localized BDSs. This approach represents activities as collections of characteristic short-term behaviours, eliminating the need for a single BDS to model overly complex attribute trajectories. The study proposes a procedure for learning a dictionary of BDSs and quantizing videos based on this dictionary, demonstrating superior performance compared to state-of-the-art temporal structure modelling approaches in challenging datasets. Overall, BoWAD provides a unified and effective solution for recognizing complex activities in videos.

2. Relatedwork

In recent years, action recognition has emerged as a crucial problem in computer vision, and the bag-of-features (BoF) H. Wang et al. [5] representation has gained significant popularity in this field. The BoF approach involves representing videos as collections of feature vectors, enabling the modelling of temporal activity structures. Several models have been developed based on this representation, such as Laptev et al. [6] use of spatiotemporal binning pyramids to match vector-quantized histograms from different video regions. Additionally, Niebles et al. [7] and Gaidon et al. [1] have represented activities with many decomposable parts or atomic actions, exploring the potential of generative models in this context. Moreover, Laxton et al. [8], V. Kellokumpu et al. [9], B. Li et al. [10] and R. Chaudhry et al. [11] integrated confidence about objects and sub-actions over time using dynamic Bayesian networks, while various active systems have been utilized to represent the evolution of human activity, employing features like local binary patterns, tracked parts, or frame-wise motion histograms.

Recent advancements in image analysis research have revealed the advantages of semantics or attribute-based representations over the BoF approach for action recognition. Liu et al. [12]. proposed using attributes as latent variables for support vector machines to enhance action recognition. Sadanand et al. [13] demonstrated substantial improvements over standard benchmarks by employing a bank of action detectors sampled across semantic and viewpoint spaces. Similarly, Rohrbach et al. augmented videos with text-script data, modelling activities as standard sets of attributes defined in terms of basic actions and objects. The introduction of the

binary dynamic system (BDS) model by Li and Vasconcelos [14] emphasized the significance of modelling video trajectories in attribute space to understand human behaviour better.

In this study, the authors expand on attribute dynamics modelling by proposing to learn dictionaries of models for attribute dynamics [14]. The approach of G. Doretto et al. [15] builds upon the bag-of-systems framework, where dynamic textures (DTs) were previously employed to characterize emotional scenes. However, A. Ravichandran et al. [16], the main challenge lies in identifying the "centroid" of a collection of dynamic textures due to the non-Euclidean nature of the space of linear dynamic systems. To overcome this, the authors propose an alternative and principled solution explicitly designed for clustering attribute sequences, offering several advantages over the MDS-kM (multi-dimensional scaling and k-means) approach.

Through their proposed method, they achieve superior recognition and accuracy in modelling temporal structures for complex activity recognition, further advancing the state-of-the-art in this important domain of computer vision research.

3. Methodology

In the context of activity recognition, we present a novel representation known as the "bag-of-words for attribute dynamics" (BoWADs). This new approach aims to capture and characterize the dynamics of attributes associated with various activities. By representing videos as collections of binary attributes, BoWADs offer a higher level of semantics, enabling better generalization and improved recognition of complex activities. The key idea behind BoWADs is to model video segments' short-term dynamics, providing a more effective way of inferring activities and discriminating between different actions. Figure 2 shows the proposed methodology.

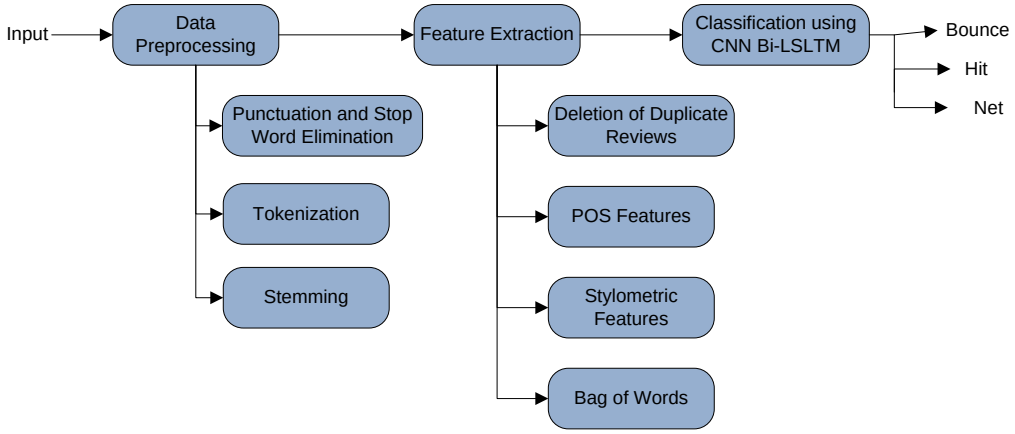


Fig. 2. Proposedworkflowdiagram

A. Words and Attributes

In the context of activity recognition, we present a novel representation known as the "bag-of-words for attribute dynamics" (BoWADs). This new approach aims to capture and characterize the dynamics of attributes associated with various activities. By representing videos as collections of binary attributes, BoWADs offer a higher level of semantics, enabling better generalization and improved recognition of complex activities. The key idea behind BoWADs is to model video segments' short-term dynamics, providing a more effective way of inferring activities and discriminating between different actions. Figure 2 shows the proposed methodology.

The bag of visual words (BoVW) has gained popularity as a widely used representation for image classification and, more recently, for action recognition. It involves representing an image as a Bag of Features (BoF) by learning a dictionary of representative feature vectors, termed visual words, and quantizing the extracted features for classification purposes. The BoVW representation is a histogram of visual word counts, frequently employed as a feature vector for image and video classification tasks. However, despite its widespread use, alternative feature spaces have shown significant benefits in encoding higher-level semantics. These alternatives represent images or videos as collections of binary attributes, offering a more insightful representation of the data [17].

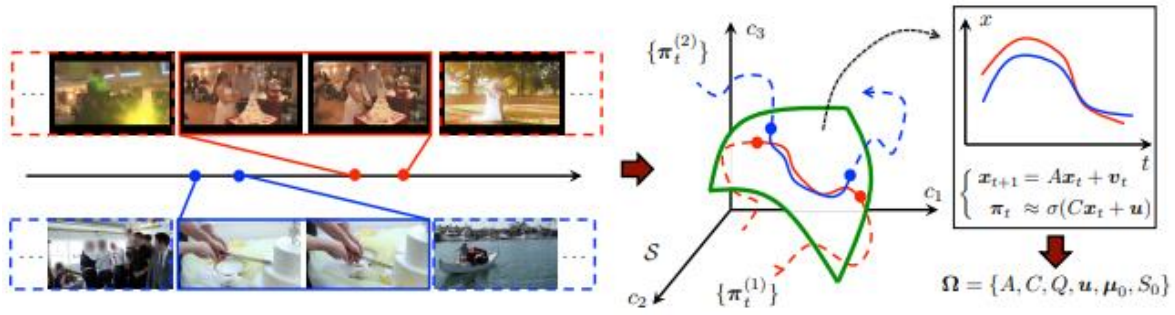


Fig. 3.Learninga BDS. Video sequences(left) Trajectories in attribute space (center) Trajectoryinthelatent state space(right).

The parameterized function B(y;p) represents a multivariate Bernoulli distribution with parameter p ∈ [0, 1]K, while the component-wise logistic transformation σ(θ) is defined as σi(θ) = (1 + e(-θi))(-1). The observation model can be interpreted as a binary principal component analysis (binary PCA) applied to the binary data {yt}. Binary PCA is a dimensionality reduction technique specifically designed for binary data. It takes a matrix Y = [y1, ..., yτ] ∈ {0, 1} (K×τ) as input and produces an L-dimensional embedding of the natural parameter, which serves as an attribute score vector. Each component πi(v) in this vector represents a confidence score quantifying the presence of the ith attribute in video v. In this context, these scores are treated as posterior probabilities πc(v) = p(c|v) of attribute c given a low-level representation of video v, such as a Bag of Features (BoF) histogram of spatiotemporal descriptors.

Figure 3 illustrates learning a BDS (Behavioral Dynamic System). Video sequences, shown on the left, are transformed into trajectories within an attribute space, as depicted in the center. Sequences with comparable semantics follow similar trajectories. The BDS then employs binary PCA to map these video trajectories into a lower-dimensional space, represented in green. Finally, a Gauss-Markov process is learned to describe the corresponding trajectories in this latent state space, as shown on the right.

A. Bag of Words for Attribute Dynamics

In recognizing tennis gameplay, the Binary Dynamic System (BDS) provides a more comprehensive model of video dynamics than the holistic attribute model. However, the BDS still presents two significant limitations illustrated in Figure 3. Firstly, there is no guarantee that a video sequence solely depicts the desired tennis activity, such as a particular shot or serve. Often, informative segments (e.g., a backhand stroke) may be surrounded by less relevant elements (e.g., players moving between shots), leading to parameter estimates that do not precisely represent the tennis event of interest. Secondly, as tennis matches involve various complex actions and movements, some of which may occur at different times, these state trajectories are unlikely to conform to the Gauss-Markov process. Nonetheless, these limitations are less likely to arise when the BDS is applied to short-term video segments.

On the other hand, most tennis actions can be effectively identified by characterizing the short-term segments that make up the gameplay. For example, the attribute sequence "swing-hit," "run-approach," and "ball-land" may sufficiently differentiate a forehand shot from a backhand shot, which could be characterized by the attribute sequence "swing-hit," "run-backward," and "ball-land." Tennis actions can be effectively distinguished by examining the presence or absence of certain attributes in video segments. Based on these observations, the approach proposes extending the Bag of Visual Words (BoVW) representation to capture the short-term dynamics of attribute sequences in tennis gameplay videos.

To implement this approach for recognizing tennis gameplay, tennis match videos are divided into temporal overlapping segments. Each segment represents a short portion of the gameplay and may contain a specific tennis action or movement. These segments then undergo attribute mapping to extract relevant attributes representing the tennis activities within each segment. The attribute sequence for a segment can be denoted as$\Pi^{(i)} = \left\{\pi_t^{(i)}\right\}_{t=1}^{\tau i}$Πt (i) represents the attribute score vector at a time ‘t’ in the segment ‘i’.

By creating attribute sequences for each segment, the short-term dynamics of tennis gameplay can be effectively captured in a structured manner, allowing for more accurate recognition and classification of tennis actions.

B. Learning and Recognition with BoWADs

In the upcoming section (Section 5), we will demonstrate that BoWADs, when used in conjunction with standard histogram-based classifiers like support vector machines (SVMs) equipped with the histogram intersection kernel (HIK), serve as a highly effective representation for recognizing complex activities. However, before delving into the recognition aspect, let us focus on the initial challenge of quantizing attribute sequences. To do so, we commence by tackling the task of learning a dictionary for Attribute Dynamics (WAD dictionary).

Traditional clustering methods, like k-means, aim to find prototypes within the space of training samples. For instance, in k-means, a cluster prototype corresponds to the centroid of the data points within the cluster, and the Euclidean distance metric is commonly used. However, extending this approach to the clustering of BoAS is not straightforward due to several challenges:

Attribute sequences can have varying lengths, making direct comparison and clustering difficult. The attribute space of these sequences exhibits non-Euclidean geometry, further complicating the clustering process. The search for optimal prototypes under this nonlinear geometry may lead to intractable nonlinear optimization.Given our interest in characterizing the appearance and dynamics of attribute sequences, finding a prototype Binary Dynamic System (BDS) is more desirable than a set of prototype sequences. We propose a solution based on learning a Bag-of-Models (BoM) dictionary to address this. Let {zi} with zi∈, Z for i = 1 to N be a set of training samples. The goal is to learn a collection of representative models {Mi} in a model space M. This involves two essential mappings:

The first mapping fM: Z ↦ M({zi}) ∈ M maps a collection of examples {zi} from the data space D to a model M({zi}) in the model space M.

$$f_{\mathcal{M}} : Z \supseteq \{z_i\} \rangle \rightarrow M(\{z_i\}) \in \mathcal{M} \quad (1)$$

The second mapping, M × M ↦dM (M1, M2) ∈ R+, measures the distance between two models, M1 and M2.

$$\mathcal{M} \times \mathcal{M} \ni (M1, M2) \rangle \rightarrow d_{\mathcal{M}}(M1, M2) \in R+ \quad (2)$$

We utilize the above mappings to create a model M(zi) for each training example zi. The training samples are then clustered at the model level using two alternating steps:

In the assignment step, each zi is assigned to the cluster whose model is closest to M(zi) based on the metric dM (M1, M2).In the model refinement step, the model associated with each cluster is relearned from the training samples assigned using the mapping fM.This clustering process, summarized above as Bag-of-Models Clustering (BMC), generalizes the standard k-means algorithm. The learned Words for Attribute Dynamics (WAD) from the training BoAS are then utilized to quantize the BoAS extracted from the video sequence for classification. The resulting histogram of WAD counts is denoted as the Bag of Words for Attribute Dynamics (BoWAD) and serves as the feature vector for video classification. This representation is summarized as BoWAD-BMC.

Furthermore, we extend the BDS learning process with a two-step decomposition, as discussed in Section 3.2. We first apply binary PCA to all attribute score vectors and then learn the parameters of the hidden Gauss-Markov process through a least squares problem involving all latent state sequences obtained from binary PCA. This approach enables each BDS learned per cluster to effectively characterize the appearance and dynamics of all attribute sequences in that cluster.

Finally, the assignment step uses the Bi-LSTM to classify between three BDSs. Initially proposed as a measure of dissimilarity between infinite output sequences of three Linear Dynamic Systems (LDSs), the Bi-LSTM has been adapted for distinguishing the outputs of three activity structures like Bi-Net Chuny (BC) Kernal [18].

All experiments used a 5-dimensional state space for both BDSs and BoWADs. The low-level representation was the Bag of Features (BoF) of spatial-temporal interest points (STIPs), quantized into a vocabulary learned from the training set.

C. ClassificationphaseusingBiLSTM

In this section, the classification phase utilizes Bidirectional Long Short-Term Memory (BiLSTM) networks. The objective is to identify and categorize tennis activities effectively and accurately. The proposed approach involves the BiLSTM model, which will be further explained later. The primary goal is to distinguish between different activities with high precision and improve the recognition system's overall accuracy.

1) Bidirectional Long Short-Term Memory (BiLSTM) Networks: In the research, the Bidirectional Long Short-Term Memory (BiLSTM) network was utilized to construct the model. BiLSTM is regarded as one of the most prominent Recurrent Neural Networks (RNNs) due to its capability to retain essential information while discarding transient data. The memory cell in BiLSTM incorporates a data gating system, enabling effective handling of long-term sequences and addressing gradient-related issues, thus facilitating feature extraction. Figure 4 illustrates the configuration of input-output and forget gates in BiLSTM, allowing the network to detect interdependencies using multiple cells. The operations and equations of the three gates are further elaborated in the following section. The bidirectional nature of the BiLSTM allows the network to capture both past and future contextual information, which is beneficial for understanding complex activity patterns.

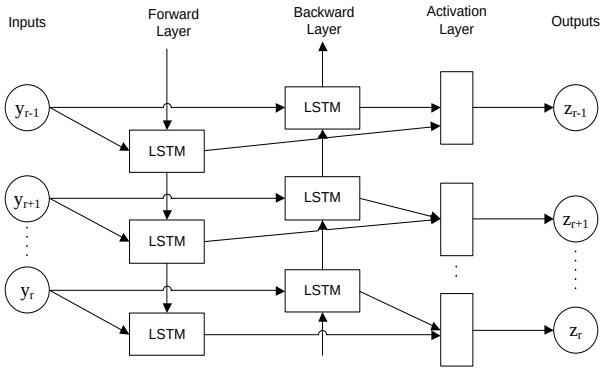


Fig. 4. Bi-LSTM structure.

2) Classification and Output: The final step of the classification phase involves making predictions based on the output of the BiLSTM network. We use a fully connected layer on the BiLSTM to classify activity classes. The model is trained using labeled video data, and during the inference phase, it predicts the activity label for a given input video.The BiLSTM networks enable our model to effectively recognize video activities by exploiting spatial and temporal information. This approach has demonstrated strong performance in various activity recognition benchmarks and real-world applications.

4. Experimental Results

This section presents the results obtained from extensive testing and simulations conducted to evaluate a fictitious product review system's performance thoroughly. The experiments were performed on the MATLAB 2019b platform, ensuring compatibility with the Windows operating system.

We employed several essential metrics to assess the system's effectiveness, including accuracy, specificity, sensitivity, and precision. These metrics allow us to evaluate the simulator's performance comprehensively.Accuracy measures the overall correctness of the system's predictions, providing insight into how well it can classify reviews accurately. On the other hand, specificity evaluates the system's ability to correctly identify genuine reviews, ensuring counterfeit reviews are not incorrectly classified as genuine. Sensitivity, also known as recall or true positive rate, focuses on the system's capacity to detect counterfeit reviews accurately.On the contrary, precision focuses on the system's precision in correctly classifying counterfeit reviews, minimizing the possibility of false positives. By considering these metrics together, we

understand the simulator's performance and capability to distinguish between genuine and counterfeit product reviews effectively.

Comparing the various outcomes and analyzing the performance metrics will provide valuable insights into the efficiency and reliability of the fictitious product review system. These evaluations will further refine the system and enhance its overall performance, ensuring a more accurate and dependable classification of product reviews.

1) Specifications of parameters: In our study, we employed a random number generator to select the parameters for the CNN-BiLSTM algorithm. This approach was adopted to ensure unbiased and diverse parameter configurations for conducting comprehensive experiments. Table 1 presents a detailed listing of the specific parameter values, such as learning rates, batch sizes, activation functions, and other relevant settings. Each parameter was chosen from a predefined range to encompass a diverse set of values, allowing for a robust evaluation of the CNN-BiLSTM algorithm's performance across various configurations. This random selection process helps gain insights into the algorithm's behavior and its effectiveness in recognizing activities from textual data.

TABLE I. PARAMETERSPECIFICATIONS.

Methods	Parameter metrics	Values
CNN-Bi-LSTM	The Network's Configuration	Fully connected
	The Total Number of Epochs	100
	Hidden Units Equipped with Bi-LSTM Technology	200
	Rate of Learning	0.010
	Activation	Softmax
	StepSize	100
	FullyConnectedLayer	Dropout
	The Size of the Pool	Max-pooling

Several comprehensive experiments were conducted to thoroughly evaluate the effectiveness of the proposed methods for tennis activity recognition. The test dataset consisted of 25 video sequences captured from a tennis tournament, each lasting one minute. These videos were recorded at a frame rate of 30 frames per second with a resolution of 640 x 480 pixels, resulting in approximately 1800 frames per video. Figure 5 displays examples of video frames.

Fig. 5. Examples of Video frames.

The tennis ball's position and trajectory were visualized throughout the game's timeline to gain deeper insights into the dynamics of tennis gameplay. Figure 6 illustrates this visualization, with the yellow circle representing the ball's current location at different frames and the green line depicting its trajectory over time. Analyzing the ball's movement and interactions with the players and the court provided valuable information about the patterns and characteristics of different tennis activities.

These extensive experiments and evaluations aimed to validate the effectiveness and reliability of the proposed methods in tennis activity recognition. The findings from these experiments have significant implications for enhancing the algorithms and their performance in real-world tennis video analysis scenarios.

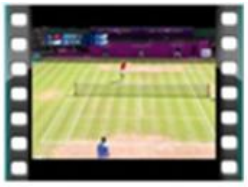
Fig. 6: Ball tracking Trajectory.

In Figure 6, the players' positions and trajectories in the current frame number are visualized. The blue circle represents the position of the players, and Figure 7 further illustrates the movement of the players. The yellow line denotes the trajectory of the lower half of the player, while the green line shows the trajectory of the upper half of the player. This visualization provides valuable insights into the players' movements and interactions during the tennis gameplay, contributing to a comprehensive understanding of the dynamics of the match [19].

Fig.7. Player Tracking Trajectory.

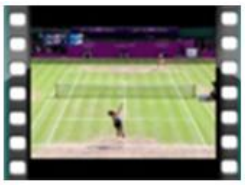
(a) (b)

Fig. 8. Dataset Contents: (a) Annotated –Action: Short Video Aligned With A Verb Phrase. (b) Video Commentary Dataset: Game Videos



Upper Player: Smashes Down the Line.
Down Player: Waits for the Ball

In Winner: Serena Williams Arrow a Good Serve At T, Sharapova Is Unable to Return It.



Upper Player: Struggles to Reach the Ball.
Down Player: Massive Serve.

In Winner: Serena Williams Arrows a Good Serve At T, Sharapova Is Unable to Return It.

Aligned With Commentary.

TABLE II. DEPICTIONS OF THE VIDEO INPUT SEQUENCE.

Input with Ground Truth		
	In Winner: Serena Williams arrows a good serve at T, but Sharapova cannot return it.	**In Winner:** Federer'sgood serve in the middle, Fedrer crafts a forehand return, short rally Delpotro cross-court forehand fails to land inside the court.
Phrases	(waits for the ball), (waits for the ball), ….. (Prepares for serve), …., (hits a good serve), (sizzling serve), ….	(Prepares for serve), …., (tosses ball for serve), (hits a good serve) …… (Waits for the ball) …… (returns a quick forehand return) ……., (sprays a forehand) …….
Descriptions (Top 2 retrievals)	1. In Winner: Serena Williams hits a good service; Sharapova struggles. 2. In Winner: Serena Williams hits a good service; Sharapova struggles.	1. In Winner: Delpotro Fine serve, Delpotro works a forehand return, brief rally, Delpotro rushes to net and punches a forehand volley winner. 2. In Winner: Federer's Quick serve, Delpotro returns a quick forehand return, a couple of shorts are exchanged, andDelpotro nets a forehand down the line.

Figure 8 (a) presents the contents of the Dataset, showcasing the alignment of annotated actions with corresponding verb phrases in short video clips. The video content captures the game's various stages, such as preparing for service, waiting for the ball, and executing a powerful serve that strikes the opponent player. The descriptions provide additional insights into the players' actions, specifically mentioning Winner Serena Williams hitting a good service and her opponent Sharapova struggling to respond.

Figure 8 (b) also introduces the Video commentary dataset, where game videos are aligned with commentary phrases. The phrases correspond to various actions, such as preparing, tossing the ball for serve, and hitting a well-executed serve. The descriptions elaborate on the player's performance, with Winner Delpotro making a

fine serve and working a forehand return. Delpotro's rush to the net and execution of a forehand volley winner is also highlighted [20].

Table 2 demonstrates the process of translating the input sequence of videos into a set of phrases. These phrases are then utilized to generate the final description of the top two retrievals. This method allows for a comprehensive and informative representation of the video content, enabling accurate and efficient recognition and retrieval of important actions and moments during the tennis gameplay.

2) Simulation Measures: This section comprehensively evaluates the model's performance by analyzing its accuracy, specificity, sensitivity, and precision. These evaluation metrics provide crucial insights into the model's ability to correctly classify and recognize different patterns and actions.
The model's accuracy is calculated using Equation (3). TP represents the number of true positives, TN indicates the number of true negatives, FP represents the number of false positives, and FN denotes the number of false negatives. The formula evaluates the proportion of correctly classified instances out of the total instances.

As represented by Equation (4), sensitivity calculates the true positive rate, the proportion of correctly identified positive instances (actions) out of all the positive instances. It gives us an understanding of the model's ability to detect positive events correctly.

On the other hand, specificity, as shown in Equation (5), measures the true negative rate, representing the proportion of correctly identified negative instances (non-actions) out of all the actual negative instances. This metric provides insights into the model's ability to classify non-action events accurately.

Precision, indicated by Equation (6), calculates the proportion of true positive instances out of all that the model predicted as positive. It is a crucial metric to determine the correctness of the model's positive predictions.

To summarize, the accuracy, sensitivity, specificity, and precision are represented as Ac, Se, Sp, and Pr, respectively. These evaluation metrics help comprehensively assess the model's performance [21] and determine its effectiveness in accurately recognizing and classifying tennis-related actions and patterns.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (3)$$

$$Sensitivity = \frac{TP}{TP + FN} \quad (4)$$

$$Specificity = \frac{TN}{TN + FP} \quad (5)$$

$$Precision = \frac{TP}{TP + FP} \quad (6)$$

From Equations (3) – (4), the accuracy, sensitivity, specificity, and precision were compared asAc, Se, Sp, andPr.

3) Performance Evaluation: The proposed model's performance for event identification can be assessed by plotting the ground truth values. Figure 9 displays the effectiveness of different types of event detection for Bounce, Hit, and Net. The plot demonstrates the model's ability to identify these events accurately, providing valuable insights into its performance across various event categories. This evaluation allows for accurately detecting all three events in a play and provides comprehensive assessment results. Figure 10 illustrates the model's accuracy, recall, and precision, showcasing the effectiveness of event identification.
To illustrate the effectiveness of different event detections further, Table 3 presents a confusion matrix. This matrix showcases the model's classification results, highlighting true positives, false positives, and false negatives, which aid in understanding the model's overall accuracy and misclassifications. These evaluations gain a comprehensive understanding of the proposed model's effectiveness in event identification, enabling informed decisions and optimizations to enhance its performance.

TABLE III. CONFUSION MATRIX OF EVENT DETECTION.

	Bounce	Hit	Net	Total
Bounce	31	2	2	35
Hit	2	36	4	42
Net	1	3	34	38

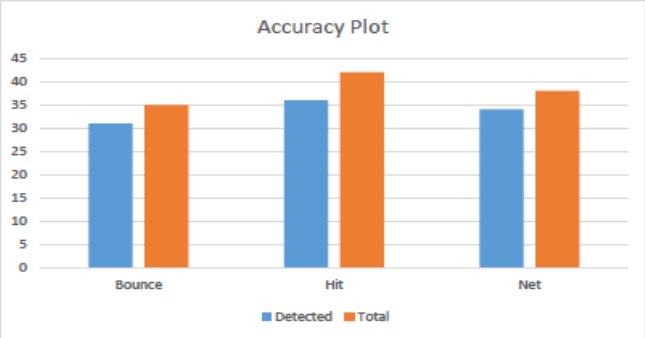


Fig. 9. Effectiveness of Different Types of Event Detection.

The accuracy results for different event detections in tennis are as follows:

- Bounce Detection Accuracy: 88.57% (31 correct predictions out of 35).

- Hit Detection Accuracy: 85.71% (36 correct predictions out of 42).

- Net Detection Accuracy: 89.47% (34 correct predictions out of 38).

Overall, Event Detection Accuracy across all action types is 87.92%. These accuracy values indicate the model's ability to classify and identify events during a tennis match.

Accuracy Bounce Detection: (31/35) *100% = 88.57%

Accuracy Hit Detection: (36/42) *100% = 85.71%

Accuracy Net Detection: (34/38) *100% = 89.47%

Overall, Event Detection Accuracy is 87.92%.

Table 4 presents a comprehensive classification of action types in tennis, encompassing total actions, correctly predicted ball hits, false predictions, and missed ball hits on the tennis court by the players. The performance evaluation matrix results, including precision and recall, are then computed to assess the model's accuracy and effectiveness in classifying various actions.

TABLE IV. CLASSIFICATION RESULTS WERE OBTAINED WITH 115 SEQUENCES.

Action Type	Total Actions	Correct	False	Missed	Precision	Recall
Bounce	35	29	4	2	87.87	93.55
Hit	42	31	7	4	81.58	88.57
Net	38	33	3	2	91.67	94.29
Total	115	93	14	8	87.92	92.08

The evaluation metrics used in the simulation include accuracy, recall, and precision, each playing a crucial role in assessing the method's effectiveness (details provided in the text). The experimental investigation demonstrates impressive results, achieving high rates. These findings validate the approach's efficacy in identifying fraudulent product reviews with considerable accuracy and precision.

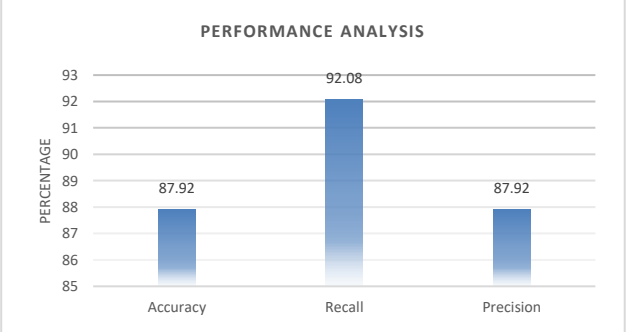


Fig. 10. Performance evaluations were conducted on a variety of simulation measures.

TABLE V. STATE OF ART COMPARISON

Algorithms	Accuracy
SVM	78.28%
LSTM	82.11%
SVM+BoW	84.33%
CNN+Bi-LSTM	87.92%

Table 5 showcases the comparative performance of machine learning and deep learning algorithms measured by accuracy. The Support Vector Machine (SVM) registers an accuracy of 78.28%, indicative of its proficiency in classification tasks via optimal hyperplane identification. On the other hand, the Long Short-Term Memory (LSTM) architecture, designed for handling sequential data due to its capacity to learn long-term dependencies, clocks in at 82.11%. An intriguing blend of SVM with the Bag of Words (BoW) model, which transforms text into a frequency-based representation, boosts the accuracy to 84.33%. This amalgamation implies a potential text classification task. The standout performer in the table is the hybrid model, CNN+Bi-LSTM, recording an impressive 87.92% accuracy. This combination marries the spatial feature extraction capabilities of Convolutional Neural Networks (CNNs) with the sequential prowess of Bidirectional LSTMs, hinting at its utility in tasks like text classification. Overall, the table accentuates the merit of integrated models, especially the CNN+Bi-LSTM, underscoring the benefits of meshing different neural architectures to enhance model efficacy. The specific utility of these models would be contingent on the dataset and problem specifics.

5. Conclusion

This study presents a novel approach to address the activity recognition challenge by modeling attributes and dynamics. The method combines the advantages of histogram-based representations and the power of BDSs to capture video attribute dynamics effectively. Novel algorithms were developed to learn BDS dictionaries and quantify video data to achieve this.

The proposed representation surpasses other state-of-the-art methods relying on attributes or temporal structures to recognize complex activities. In the context of tennis event detection, the model demonstrated high accuracy for different types of actions, with bounce detection achieving 88.57% accuracy, hit detection achieving 85.71% accuracy, and net detection achieving 89.47% accuracy. The overall event detection accuracy for tennis was 87.92%, effectively identifying different events during a tennis match.

Furthermore, the experimental analysis highlighted the impressive performance of the proposed approach in accurately identifying the tennis events' Bounce, Net, and Hit. The model achieved outstanding accuracy (87.92%), recall (92.08%), and precision (87.92%).

References

[1] K. Tang, L. Fei-Fei, and D. Koller. Learning latent temporal structure for complex event detection. CVPR, 2012.

[2] M. Rohrbach, M. Regneri, M. Andriluka, S. Amin, M. Pinkal, and B. Schiele. Script data for attribute-based recognition of composite activities. ECCV, 2012.
[3] W. Li and N. Vasconcelos. Recognizing activities by attribute dynamics. NIPS, 2012.
[4] J. Zhang, M. Marszałek, S. Lazebnik, and C. Schmid. Local features and kernels for classification of texture and object categories: A comprehensive study. IJCV, 73(2):213 – 238, 2007.
[5] H. Wang, M. Ullah, A. Kl¨aser, I. Laptev, and C. Schmid. Evaluation of local spatiotemporal features for action recognition. BMVC, 2009.
[6] I. Laptev, M. Marszałek, C. Schmid, and B. Rozenfeld. Learning realistic human actions from movies. CVPR, 2008.
[7] J. Niebles, C. Chen, and L. Fei-Fei. Modeling temporal structure of decomposable motion segments for activity classification. ECCV, 2010.
[8] B. Laxton, J. Lim, and D. Kriegman. Leveraging temporal, contextual, and ordering constraints for recognizing complex activities in video. CVPR, 2007.
[9] V. Kellokumpu, G. Zhao, and M. Pietik¨ainen. Human activity recognition using a dynamic texture-based method. BMVC, 2008.
[10] B. Li, M. Ayazoglu, T. Mao, O. Camps, and M. Sznaier. Activity recognition using dynamic subspace angles. CVPR, 2011.
[11] R. Chaudhry, A. Ravichandran, G. Hager, and R. Vidal. Histograms of oriented optical flow and binet-cauchy kernels on nonlinear dynamical systems for recognizing human actions. CVPR, 2009.
[12] J. Liu, B. Kuipers, and S. Savarese. Recognizing human actions by attributes. CVPR, 2011.
[13] S. Sadanand and J. J. Corso. Action bank: A high-level representation of activity in video. CVPR, 2012.
[14] W. Li and N. Vasconcelos. Recognizing activities by attribute dynamics. NIPS, 2012. B. Afsari, R. Chaudhry, A. Ravichandran, and R. Vidal. Group action-induced distances for averaging and clustering linear dynamical systems with applications to analyzing dynamic scenes. CVPR, 2012.
[15] G. Doretto, A. Chiuso, Y. Wu, and S. Soatto. Dynamic textures. IJCV, 51(2):91–109, 2003.
[16] A. Ravichandran, R. Chaudhry, and R. Vidal. Categorizing dynamic textures using a bag of dynamical systems. IEEE TPAMI, 35(2):342–353, 2012.
[17] N. Rasiwasi, P. J. Moreno, and N. Vasconcelos. Bridging the gap: Query by semantic example. IEEE Trans. Multimedia, 9(5):923–938, 2007.
[18] A. Ravichandran, R. Chaudhry, and R. Vidal. Categorizing dynamic textures using a bag of dynamical systems. IEEE TPAMI, 35(2):342– 353, 2012. 2, 5, 6, 7, 8
[19] Archana, M., and M. Kalaisevi Geetha. "Object detection and tracking based on trajectory in broadcast tennis video." Procedia Computer Science 58 (2015): 225-232.
[20] Sukhwani, Mohak, and C. V. Jawahar. "Tennisvid2text: Fine-grained descriptions for domain-specific videos." arXiv preprint arXiv:1511.08522 (2015).
[21] Hiremath, S.S., Hiremath, J., Kulkarni, V.V., Harshit, B.C., Kumar, S., Hiremath, M.S. (2023). Facial Expression Recognition Using Transfer Learning with ResNet50. In: Suma, V., Lorenz, P., Baig, Z. (eds) Inventive Systems and Control. Lecture Notes in Networks and Systems, vol 672. Springer, Singapore. https://doi.org/10.1007/978-981-99-1624-5_21