Hindawi
Wireless Communications and Mobile Computing
Volume 2021, Article ID 2055703, 11 pages
https://doi.org/10.1155/2021/2055703

Research Article

Classification of Tennis Video Types Based on Machine Learning Technology

Xun Gong[1] **and Fucheng Wang**[2]

[1]*Physical Education Department, Heilongjiang Bayi Agricultural University, Daqing, 163319 Heilongjiang, China*
[2]*Engineering College, Heilongjiang Bayi Agricultural University, Daqing, 163319 Heilongjiang, China*

Correspondence should be addressed to Fucheng Wang; wangfucheng@byau.edu.cn

Received 15 April 2021; Revised 6 May 2021; Accepted 22 May 2021; Published 7 June 2021

Academic Editor: Wenqing Wu

With the rapid development of online video data, how to find the required information has become an urgent problem to be solved. This article focuses on sports videos and studies video classification and content-based retrieval techniques. Its purpose is to establish a mark and index of video content and to promote user acquisition through computer processing, analysis, and understanding of video content. Video tennis classification has high research and application value. This article focuses on video tennis based on the selection of the basic frame of each shot and proposes an algorithm for classification of shots based on average grouping. Based on this, we use a color-coded spatial detection method to detect the type of tennis match. Then, it integrates the results of audiovisual analysis to identify and classify exciting events in tennis matches. According to statistics, although the number of people participating in tennis cannot enter the top ten, the number of spectators ranks fourth. Four tennis tournaments, masters, and crown tournaments are held every year around the world. Watching large-scale international tennis matches has become a pillar of leisure and vacation for many people. Tennis matches last from two hours to four hours or more, and there are countless large and small tennis matches around the world every year, so the number of tennis records created is staggering. And artificial intelligence technology is rarely used in tennis in the sports world (5%), but football has reached 50%. Therefore, when dealing with such a large amount of data, we urgently need to find a fast and effective video retrieval classification method to find the required information. The experiment of tennis video classification research based on machine learning technology proves that the accuracy of tennis video classification reaches 98%, so this system has high feasibility.

1. Introduction

1.1. Background. Tennis video data contains a lot of information, such as characters, scenes, objects, actions, and stories. According to statistics, 80% of the information received by humans is received visually, while tennis video information is intuitive and vivid, making it the most effective way of communication in human life. In today's fast-changing, complex information age, the development of computer and network technology, and the promotion and application of multimedia, various tennis video materials are constantly being created, and more and more tennis videos and digital databases are emerging: on-demand tennis video, mobile, web TV, and many other new and tennis video streaming media. In the face of a large number of tennis videos, how to quickly obtain the required information from them is very important. Machine learning has a wide range of applications in human behavior recognition, mainly focusing on smart video surveillance, patient monitoring systems, human-computer interaction, virtual reality, smart home, smart security, athlete-assisted training, and content-based video retrieval and smart image compression. It has broad application prospects and potential economic and social value, and many methods of behavior identification are also used. So machine learning provides technical support for tennis video classification.

1.2. Significance. Sports tennis video content analysis technology is individually evaluated by researchers due to its wide application prospects and significant academic value. The

basic technology of sports tennis video content analysis is actually the analysis of events and their relationships. By highlighting and organizing these events, the query needs of most users can be met. From an academic point of view, event detection and recognition are typical problems of computer vision and pattern recognition. On the one hand, content-based tennis video search technology creates an effective semantic index of tennis video information, allowing users to quickly and easily view and retrieve tennis video content. At the same time, the development of content-based tennis video search technology provides more user-friendly and personalized functions. New tennis video services have changed the way users watch tennis videos from passive acceptance to more active choices. It can be said that related technologies such as content analysis have changed the consumption of information. Users can not only view differentiated high-resolution professional tennis video data but also capture a large amount of effective information. Therefore, finding a way to solve these problems and successfully creating a sport tennis video content analysis system will provide important insights into other similar target semantic analysis problems and ultimately video analysis and search, thereby promoting the development of various fields.

1.3. Related Work. Rovithakis et al. proposed a hybrid neural network/genetic algorithm technology to design feature extractors to achieve highly separable classes in the feature space. The purpose of this system is to identify the condition of human tissues in the surrounding blood vessels (i.e., normal, fibrous, and calcified). In order to distinguish normal levels from normal cells and cells affected by acute lymphoblastic leukemia, the system was further tested and classified by the range of blood sample nucleus measurement. As an advantage of the proposed technique, you may encounter the fact that the algorithmic nature of the design process, the result of optimized classification, and the performance of the system are less dependent on the type of classifier used [1]. However, due to the uncertainty of the experimental process, there are still gaps in the experimental results. Zupanc and Bosnic believe that essays are considered to be the most useful tool for evaluating student learning outcomes, guiding the learning process, and measuring progress. Grading students manually is a time-consuming process, but it is still necessary. Automatic composition evaluation is a practical solution to this task, but its main disadvantage is that it mainly focuses on vocabulary and text grammar, while semantic testing is very limited. They suggested extending the existing automated paper evaluation system by incorporating consistency and other semantic features of consistency [2]. However, their experimental process is not closed, so there is a certain deviation in the experimental results. Jenke et al. use EEG signals for emotion recognition, which can directly assess the user's "internal" state, which is considered an important factor in human-computer interaction. Many feature extraction methods have been studied, and appropriate features and electrode positions are usually selected based on neuroscience findings. However, a small number of different feature sets have been used and their suitability for emotion recognition has been tested on different data sets that are usually small. One major limitation is the comparison of systems with no features. Therefore, they reviewed the feature extraction methods for emotion recognition in EEG signals based on 33 studies. An experiment was conducted to compare these features using machine learning techniques to perform feature selection on self-recorded data sets. The performance of different feature selection methods, the usage of selected feature types, and the results of electrode position selection are given [3]. However, the factors selected by their multivariate method are slightly inferior to that of the univariate method, resulting in inaccurate results.

1.4. Innovation Points/Main Content. The innovation of this paper is (1) the use of video stream semantic analysis methods, which mainly include shot classification, player detection, and player tracking. On the basis of the existing lens classification algorithm, the characteristics of the lens in the tennis match video are fully studied, and a lens classification scheme based on Hough transform and SVM is proposed, which divides the shots into match shots and nonmatch shots. Then, use the frame difference method to achieve player detection and player tracking and give the experimental results. (2) Through the audio and video feature fusion technology, the detection of ACE balls, bottom line matches, and tennis balls in tennis matches is realized. (3) Introduce a continuous hidden Markov model (CHMM). The sounds that appear in a tennis match are divided into five categories, namely, batting, cheering, enthusiastic commentary by the narrator, ordinary commentary by the narrator, and noise. Calculate the audio feature values, and then train the sample parameters to obtain the continuous hidden Markov model of the five types of sounds, and then calculate the output probability of the sound to the various CHMM models, and use the maximum output probability to realize the automatic classification of the audio stream in the tennis match.

2. Tennis Video Field Lens Detection

The traditional video structure analysis method is to edit the segmentation in the video stream, basic frame extraction, and scene segmentation to obtain structured video information [4, 5]. The structure of the tennis video is very good. Shot changes usually occur after the end of the round, usually with the player's close-up (maybe the audience's shot) and the beginning of the next round. Due to the large amount of tennis video data, the method of the first one detecting the download limit, then selecting the basic frame, and then downloading the game download content by shooting classification contains a lot of unnecessary content, which must be [6, 7]. Therefore, it is more effective to extract court shots directly from the tennis video stream. Figure 1 shows a common video structured block diagram. It can be seen from the figure that the video structure is mainly composed of video key frames, video shots, video groups, and video scenes. We know that video is composed of out-frame images, so the video frame is the most basic element of the video structure [8, 9]. A continuous video frame constitutes a video shot. In a video frame, one or more frames of video that can

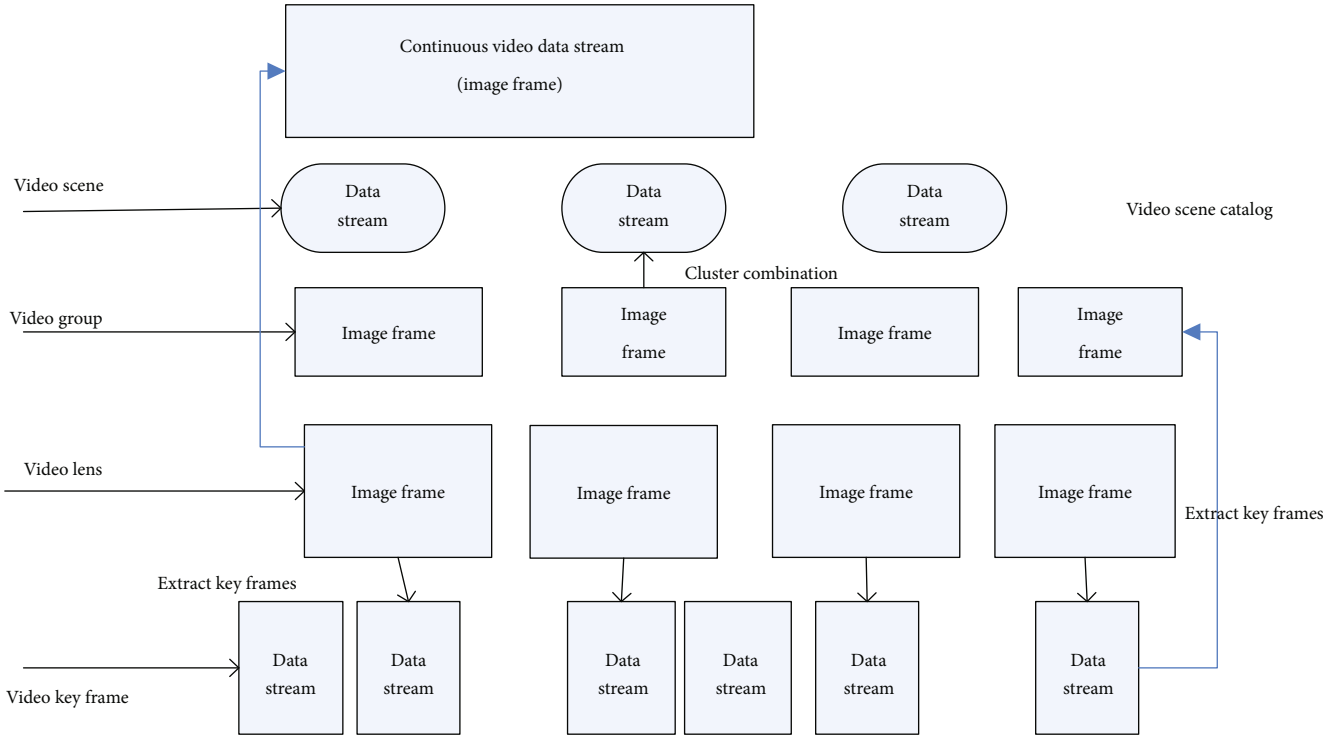

Figure 1: Structured block diagram of video stream.

represent the characteristics of a video shot constitute a key frame. Extracting key frames and segmenting video shots are the basis and key step of video structuring, which directly affects the efficiency of subsequent video detection [10]. Several video shots can constitute a video scene, and these shots are either semantically related or adjacent in time. The structure between the video scene and the video shot is defined as a group. The shots in the same group usually have similarity in content, similarity in background, and continuity in time [11], as shown in Figure 1.

2.1. Common Lens Detection Methods for Tennis Courts. The following briefly introduces several methods for detecting shots of tennis courts. The tennis match field shot referred to here is defined as a global shot containing the field [12, 13], because tennis video first needs to detect whether it is a tennis match, which is divided according to the color-coded spatial detection method of the tennis court.

(1) Detection method based on the statistical characteristics of the main color: there are four main types of tennis courts: clay courts, synthetic plastic courts, hard courts, and grass courts. These types can be distinguished by their color attributes [14]. Through the statistical analysis of some typical frames of tennis courts, Xu summarized the standard color attribute characteristics of four types of venues, as shown in Table 1.

This method first calculates the main color of the current frame image in the prescribed rectangular window and then calculates the Euclidean distance between the main color and the mean color value of various venue types (as shown in Table 1.1) and obtains the minimum Euclidean distance [15]. Which venue type has the smallest Euclidean distance, then court_class is recorded as the venue type, and court_color is recorded as the main color value corresponding to the venue type. The formula for calculating the Euclidean distance between pixel point (m, n) and point (j, k) is as follows:

$$d_e[(m,n),(j,k)] = \left[(m-j)^2 + (n-k)^2\right]^{1/2}. \quad (1)$$

For the color image of the color model, the formula for calculating the Euclidean distance is as follows:

$$d_e[(m,n),(j,k)] = \left[\left(B_{(x,y)} - B_{(j,k)}\right)^2 + \left(R_{(x,y)} - R_{(j,k)}\right)^2 + \left(G_{(x,y)} - G_{(j,k)}\right)^2\right]^{1/2}. \quad (2)$$

Table 1: Color statistical characteristics of tennis courts.

Class of count	Mean (R : G : B) colors	Class threshold	Fraction threshold
Carpet	111 : 154 : 123 68 : 69 : 71 59 : 53 : 52 135 : 112 : 152	25	40
Clay	125 : 56 : 54 167 : 56 : 68	25	25
Hard	46 : 56 : 25	25	25
Grass	95 : 152 : 65 68 : 54 : 152	25	40

If the minimum Euclidean distance obtained is not less than the marked field type threshold (Class Threshold), then, the frame image is considered to be not a frame containing a tennis court [16]. Otherwise, do the following for each pixel in the rectangular window area: calculate the Euclidean distance between the color of each pixel and the field color marked by count_color, if the distance is greater than the Fraction_Threshold threshold of the field type marked by count_class, then

$$\text{Court} = \text{court} + 1, \tag{3}$$

$$\text{Total} = \text{total} + 1. \tag{4}$$

Finally, the formula is used to calculate the proportion of pixels belonging to the court. If court_fraction is greater than 0.6 (determined as the set empirical value), the frame of image is considered to be a tennis court frame belonging to the current court type. Otherwise, the frame does not belong to the frame containing the playing field [17]. This method is more convenient and simple, easy to implement, and low computational complexity. However, there are two shortcomings. First, when calculating the Euclidean distance between a specific type of field frame and the color of the current frame, a fixed threshold is used to determine whether the current frame is a game scene frame. The problem is that the fixed closed value has certain limitations. If the selection is unreasonable, it will greatly affect the detection effect [18, 19]. Second, using color features alone cannot obtain good detection results. If the color of the clothing worn by the player is very close to the color of the field, the close-up shot of the player will also be considered a field shot [20].

(2) Detection method based on white pixels: Pan and Weijun observed that no matter what type of court, the boundary line of the tennis court is always white. In addition, the number of white pixels composing the boundary line of the field is a relatively constant [9]. Using these two characteristics, the algorithm is proposed.

(i) Detection of white pixels: normally, the boundary line of any tennis court is white. However, in actual situations, there is not only the field line, which is the only white object in an image [21]. Advertising icons, parts of the stadium, spectators, and sometimes even the clothes worn by the players are more or less white. Therefore, it is assumed in the literature that the width of the field line will not be wider than one pixel. It is set in the literature to select a candidate pixel and compare whether the brightness value of the pixel in the four directions of the pixel distance is greater than the brightness value of the candidate pixel. If the brightness value of the surrounding pixels is less than the brightness value of the candidate pixel, the candidate pixel is marked as a white pixel [22].

(ii) Detection of the playing field frame: set a rectangular window; the length is the length of the image, the width is one-half of the width of the image, and the center of the rectangle is the center of the image. The selection of the threshold of this method is based on experience, so it greatly affects the accuracy and effectiveness of the detection [23]. In addition, for the clay field, the white field line is not complete, and the accuracy of detection is low by relying solely on the number of white pixels on the field line [24].

3. Audio Stream Analysis and Recognition in Tennis Video

In recent years, the semantic analysis of audio streams has attracted the attention of researchers in related technical fields such as content-based video analysis and retrieval, speech recognition, and audio retrieval. By comparing and analyzing several representative audio recognition and classification algorithms, we found that the algorithm based on the HMM Hidden Markov Model has better recognition efficiency. In this paper, the HMM classification algorithm is used to divide the tennis game audio stream into five categories: batting sound, cheering sound, passionate commentary by the narrator, gentle definition by the narrator, and background noise [23].

3.1. Implementation of Audio Classifier. Currently, the implementation of audio classifiers is mainly based on the following methods:

(1) Rule-based audio classification: the basic idea of this method is you can select a feature that can be distinguished from other audio categories, then set a threshold for the function, and compare the calculated actual function value with the threshold according to the default rule to specify the audio category. This method is easy to use, but due to its simplicity, it is only suitable for audio types that have simple identification functions such as mute. This method has the following disadvantages:

(i) Decision rules and classification order are not necessarily optimal

(ii) Decision errors at higher levels will accumulate to the next level, forming a "snowball" effect

(iii) The classification error is large and requires prior human knowledge and experimental analysis, especially the threshold determination. Therefore, the classification accuracy of the rule classification method is low, and it is only suitable for simple voice classification that can be clearly distinguished, and it is difficult to support complex and multifunctional voice classification applications

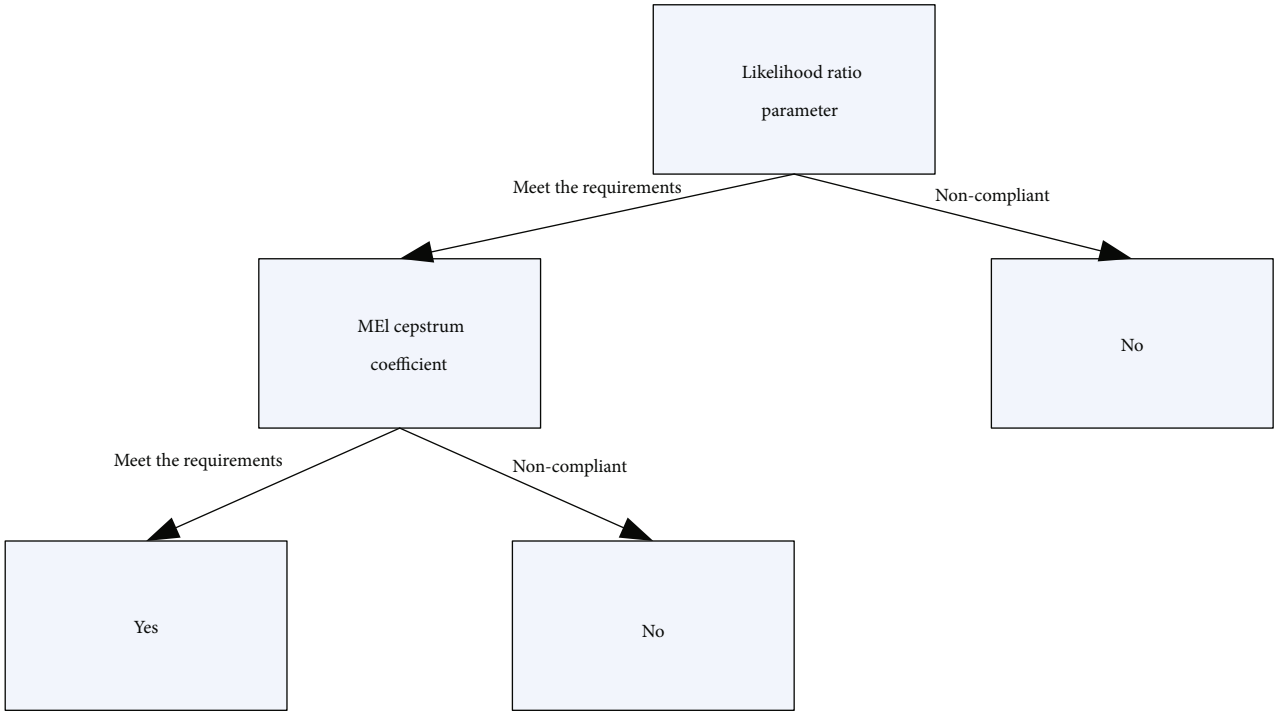

Figure 2: Schematic diagram of decision number.

(2) Minimum distance audio classification: this method uses the idea of matching templates to create a template for each audio type and then calculates the feature vector of the actual audio frame and uses the feature vector to match the template vector (usually calculating their distance from the vector space) to determine the type of sound.

(3) Audio classification based on statistical learning algorithm: this method requires specifying a batch of preclassified training samples, creating a classifier with guided learning and training and measuring the samples to be classified in the test set to measure the classification performance. Speech classification based on statistical learning algorithms is the focus of speech classification research. It provides an effective way to perform automatic and self-learning classification. This is the main research direction in this field now and in the future.

3.2. Typical Audio Classification Method. Typical audio classification methods include decision tree-based classification methods, KNN (nearest neighbor method) classification methods, Bayesian classification methods, neural network classification methods, SVM (support vector machine) classification methods, and hidden Markov model methods.

(1) Classification method based on decision tree: the so-called decision tree is a tree structure similar to a flowchart. Each node of the tree represents a test of an attribute (value), its branches represent the test results, and each node of the leaves of the tree represents a category. The highest node of the tree is the root node. The decision tree in Figure 2 describes whether the audio track is a speech signal and has been classified and predicted, as shown in Figure 2.

In order to sort and identify unknown data objects, the attribute values in the data set can be checked according to the structure of the decision tree. The path from the root node to the leaf node of the decision tree forms the class prediction of the corresponding object. Decision trees can be easily transformed into sorting rules. When building a decision tree, there are many branches in the data set that will generate noise or abnormal data. Pruning a tree is a method of locating and eliminating such branches to improve the classification accuracy of unknown objects.

(2) KNN (k-nearest neighbor method) classification method: KNN classification is a classification method that minimizes distance. KNN classifier is a classification calculation method based on learning rate. The training samples have n digital features, and each sample represents a node in a n-dimensional space, so all samples are stored in a n-dimensional space. When inputting an unknown data object (type), the KNN classifier will search the dimensionless space and find the k training samples closest to the unknown data object. When inputting an unknown data object (category), the KNN classifier will search the dimensionless space and find the k training samples closest to the unknown data object. These training samples k are the "k" of the unknown data object. The concept of "nearest" refers to the Euclidean shortest distance between two points in the e-dimensional space, and the two points in the n-dimensional

space $X = \{x_1, x_2, \cdots, x_n\}$ and $Y = \{y_1, y_2,$ the Euclidean distance between, $\cdots, y_n)$ are defined as

$$d(X, Y) = \sqrt{\sum_{i=1}^{n}(X_i - Y_i)}. \quad (5)$$

Data objects of unknown categories are classified as the category with the highest occurrence among the "k-nearest neighbors". When $k = 1$, the data object of the unknown category will be attributed to the category of the training sample closest to its target. The nearest neighbor classifier is a classification technique based on instance learning or lazy leaning.

(3) Bayesian classification method: Bayesian classification is a very mature statistical classification, mainly used to predict the possibility of relationships between class members. For example, the probability that a particular observation belongs to a particular category can be determined by the associated characteristics. The Bayesian classifier is based on Bayes' theorem. The basic Bayesian classifier assumes that the value of each attribute in the specified category is independent of each other. When using large databases, Bayesian classifiers can provide higher classification accuracy and computational performance.

(4) Neural network classification method: as an advanced artificial intelligence technology, a neural network is very suitable for processing nonlinear and those processing problems characterized by fuzzy, incomplete, and imprecise knowledge or data due to its own processing, distributed storage, and high fault tolerance. This feature of it is very suitable to solve the problem of data mining.

Similar methods to classify tennis videos using machine learning include the following:

(1) Use ConvNet to divide frames one by one

(2) Use a time-distributed ConvNet in a network and pass such features to RNN

(3) Use a 3D convolutional network

(4) Use ConvNet to extract features from each frame and pass this feature sequence to another RNN

(5) Use ConvNet to extract features from each frame and pass this feature sequence to another MLP

In summary, the HMM has a unique advantage if the research problem is based on sequence, such as time sequence or state sequence. There are two types of data in the research question. One type of sequence data is observable, that is, the observation sequence, while the other type of data cannot be observed, that is, the hidden state sequence, referred to as the state sequence.

4. Audio Stream Analysis of Tennis Match

4.1. Basic Principles of Hidden Markov. Audio classification is an important means to extract audio structure and content semantics and has great application value in content-based audio retrieval. Because the hidden Markov model (HMM) can well describe the temporal statistical characteristics of audio signals and many audios, it is often classified into multiple classes. This paper proposes an audio multilabel classification algorithm based on a hidden Markov model. The classification algorithm is used to classify so that an audio has multiple labels. HMM has a double random process. A Markov chain is a meditation process with a finite state. Another stochastic process describes the statistical correspondence based on observations and statements.

The parameters included in an HMM are as follows:

(1) E: the number of states is included in the Markov chain. Let $a_1, \cdots, a$ be e states, then $n, n(a_1 \cdots a_n)$ are the states of the Markov chain at time t

(2) F: the number of observations is corresponding to N states. Let $b_1, \cdots, b_f$ be F observation values, then o, $o \in (b_1, \cdots b_f)$ are the observation values at time t

(3) Q: the probability of the initial state is$Q = (Q_1, \cdots, Q)$, and

$$Q_1 = P(n_1 = a_i), \quad 1 \le i \le n. \quad (6)$$

Equation (6) describes the probability that any state in the HMM is taken as the initial state

(4) N: state transition probability matrix is $A = (C_{ij})_{n\times n}$, and

$$C_{ij}P\left(n_{t+1} = a_j\right), \quad 1 \le i, j \le N. \quad (7)$$

Equation (7) describes the probability of the state at time $t + 1$ and takes the state at time t as the condition

(5) B: observation probability matrix is $M = (a_{jk})_{n\times m}$, and

$$a_{jk} = P(o_t = b_k), \quad 1 \le j \le N, 1 \le k \le M. \quad (8)$$

Equation (8) describes the probability that the observed value is b_k when the state is a_j at time t.

Therefore, according to the description of each parameter above, one HMM can be denoted as

$$\lambda = (E, F, Q, N, B) \quad (9)$$

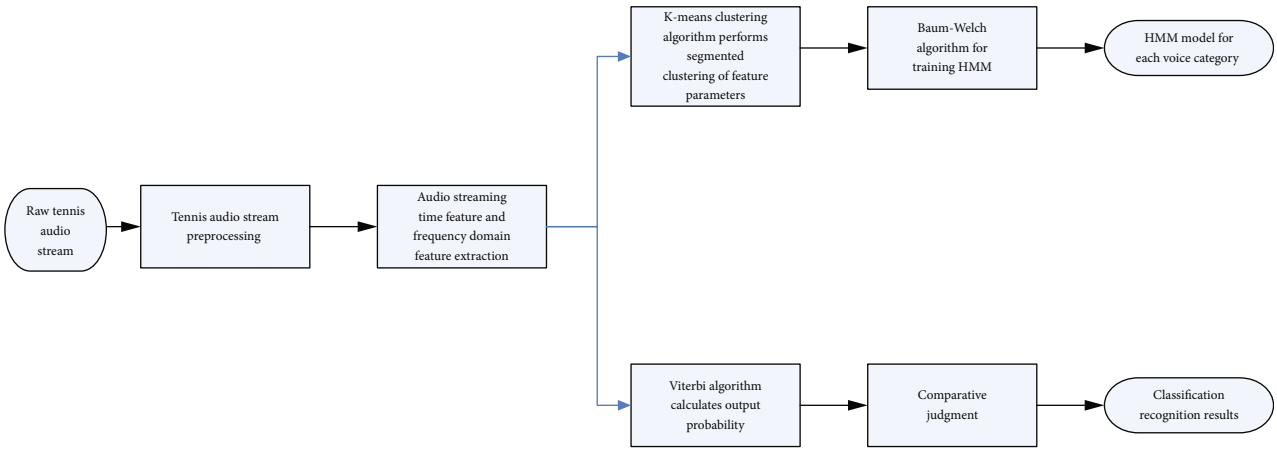

Figure 3: The flow diagram of the classification of I give you the ball audio stream based on HMM.

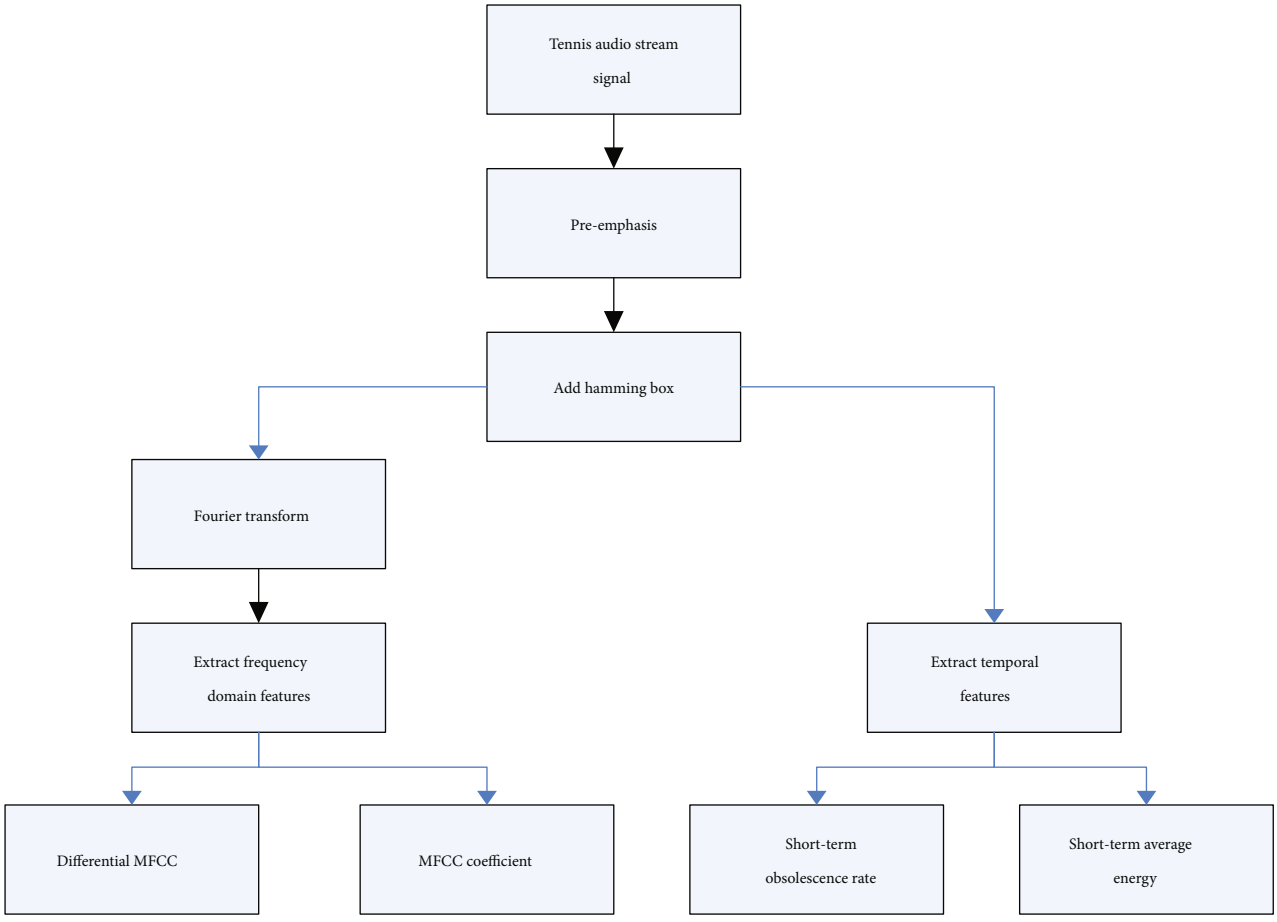

Figure 4: Block diagram of the extraction process of time domain feature and frequency domain feature parameter of tennis audio stream.

4.2. Overall Framework of Tennis Audio Stream Analysis

(1) First, preprocess the original audio stream, and extract the time-domain and frequency-domain features of the audio

(2) Second, analyze the extracted feature parameters using K-means clustering algorithm. And model the analysis results

(3) Finally, the Baum-Welch algorithm is used to train the HMM parameters. After the model is trained,

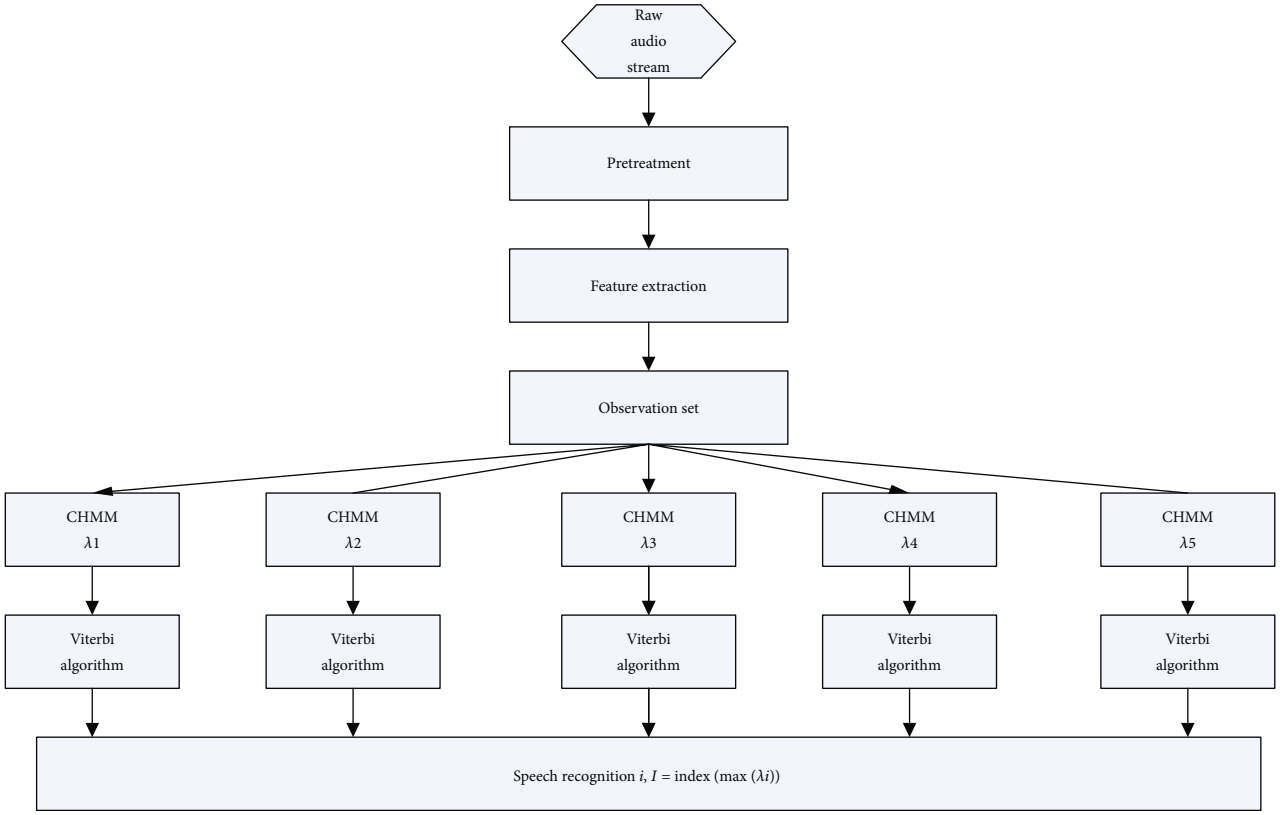

Figure 5: Flow chart of audio recognition and classification method based on CHMM.

load the tennis audio to be classified, and use the Viterbi algorithm to classify and recognize the audio to obtain the final result. The specific audio classification process block diagram is shown in Figure 3

4.3. Tennis Audio Feature Analysis and Extraction. The extraction process of the time-domain feature parameters of the tennis audio stream in this paper is shown in Figure 4.

4.4. Parameters of Training Audio Model. The core idea of parameter training is based on a specific sequence of observations $A = A_1, A_2, \cdots, A_t$ and the initial model parameters $\lambda = (E, F, Q, N, B)$; the model parameters are repeatedly adjusted to form a new model λ, so that $C(O/\bar{\lambda}) > P(O/\lambda)$ will converge until the $C(O/\bar{\lambda})$ is converged. In the process, try to maximize the probability $C(O/\bar{\lambda})$, and finally get the best $\lambda = (Q, N, B)$. It can be seen from the reevaluation formula shown in Section 3.1.2 that the appropriate model structure and the corresponding initialization parameters π_i, a_{ij}, and b_{jk} should be selected before the model parameter training [9].

The continuous hidden Markov (CHMM) observation sequence is generated by simulating the Gaussian probability density function. Moreover, many linear combinations of Gaussian probability density functions are often used in simulations, and each Gaussian probability density function has its own mean and covariance. Since the HMM model can be represented by triple $\lambda = (Q, N, B)$, the CHMM can be represented by a similar structure $\lambda = (E, F, W_{jt}, \mu_{jt}, \delta_{jt}^2)$, where W_{jt} is the weight of the l-th mixed Gaussian element in state j, μ_{jt} is the mean value of the l-th mixed Gaussian element in state j, and δ_{jt}^2 is the covariance of the l-th mixed Gaussian element in state j [22].

The most important step in CHMM model training is the selection of initial values. Choosing the correct initial value means that the number of iterations required to reach the convergence state is the least, and the calculation efficiency is significantly improved accordingly.

4.5. Audio Classification and Recognition. After training the CHMM model of five types of tennis audio using the method introduced in the above section, we use the Viterbi algorithm to classify and recognize the five type of tennis audio that have been trained. The specific classification steps are [23, 24] as follows:

(1) In preprocessing the input tennis audio stream, firstly, the tennis audio stream is divided into a sequence of audio segments of length ls, and then a Hamming window is added to each audio segment to obtain a total of nFrame audio frames and each audio frame extraction feature parameters:

$$W_i = \gamma_{i1}, \gamma_{i2}, \cdots, \gamma_{i26}, \quad 1 \le i \le \text{nFrame}, \tag{10}$$

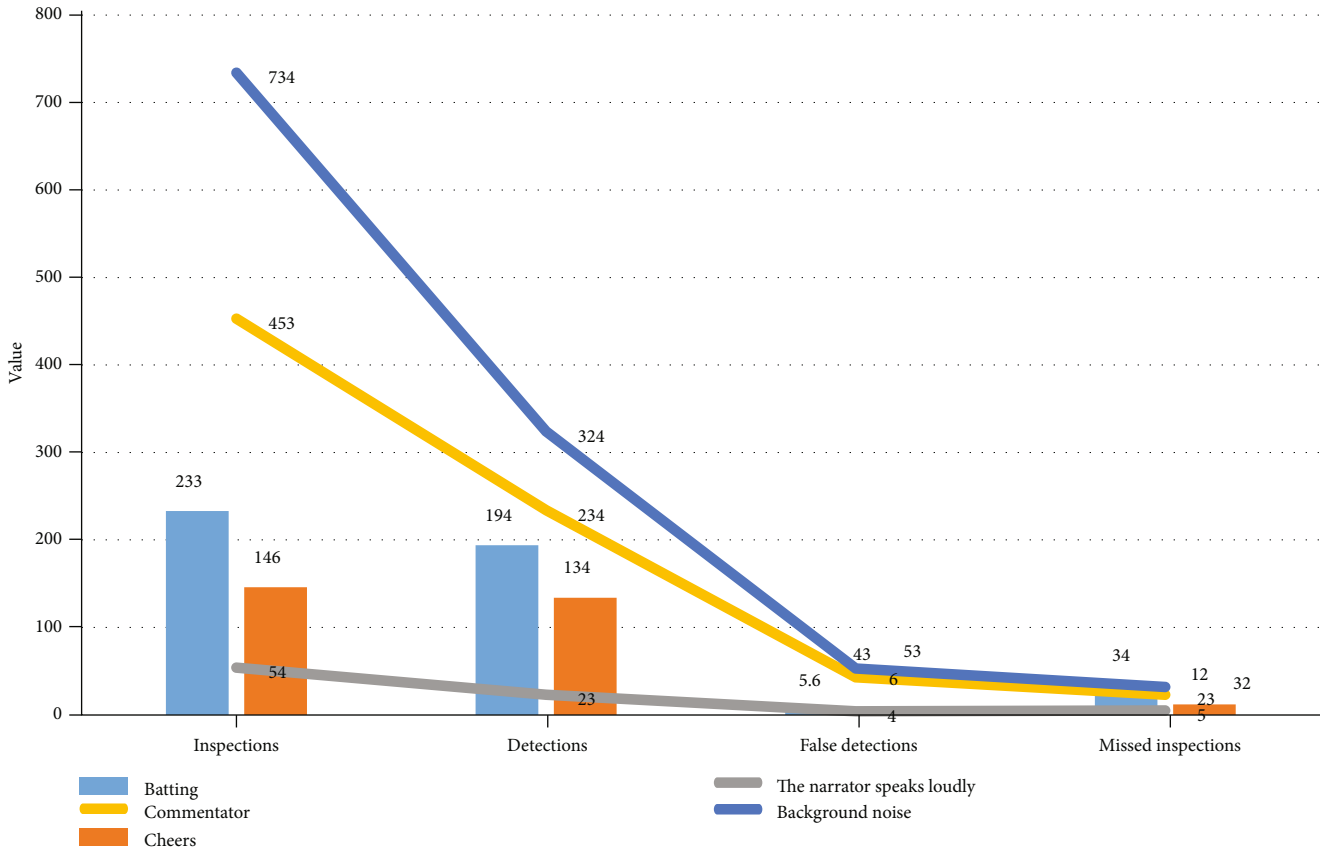

Figure 6: Experimental results of classification, recognition, and detection of audio segments in tennis matches.

where F_i represents the feature vector of the i-th frame and γ_{ij} represents the j-th feature value of the i-th frame.

(2) Obtain the feature vector $W = W_1, W_2, \cdots, W_{\text{nFrame}}$ of the audio segment from the feature parameters of the audio frame

(3) Calculate the output probability of five types of models: the sound of hitting a tennis ball, the excited cheers of the audience, the impassioned voice of the narrator, the gentle narration of the narrator, and the background noise contained in the tennis court. And these five types of models correspond to the observation sequence $W = W_1, W_2, \cdots, W_{\text{nFrame}}$ in step (2)

(4) Select the largest output probability $P(O/A)$ from the output probability group. The CHMM model corresponding to its subscript i is the category of the audio segment

(5) All audio segments are processed in steps (1)–(4), and finally, the classification and recognition of the tennis audio stream are completed. The whole process is shown in Figure 5

In order to verify the effectiveness of the algorithm, this article selects two tennis match audios for experimental analysis.

(1) The first segment contains 256 shots. Each of the five audio categories selects 20 samples as training data, and the audio test segment is composed of the remaining samples. The experimental results are shown in Figure 6

From the experimental results in Figure 6, it can be seen that the detection and recognition efficiency of background noise, impassioned commentary by the narrator, gentle commentary by the narrator, and cheers are relatively high. The recall rate of batting sound is only 82.3%, which is relatively low, and the number of false detection is relatively high. Analyzing the reasons, it is known that the sound of hitting a tennis ball is easily affected by the background noise of the playing field. For example, the referee's yelling when the player is unsuccessful in serving the ball or the sound of some hits in the game is relatively small and the sound is not obvious; these will increase the number of false detection and reduce the detection efficiency.

(2) This article selects a video of the US Open with 72 shots and 56 exciting events as the experimental material of the algorithm. The experimental results are shown in Figure 7

It can be seen from Figure 7 that the detection rate of the bottom line hit event has reached 92% and 88%, which are

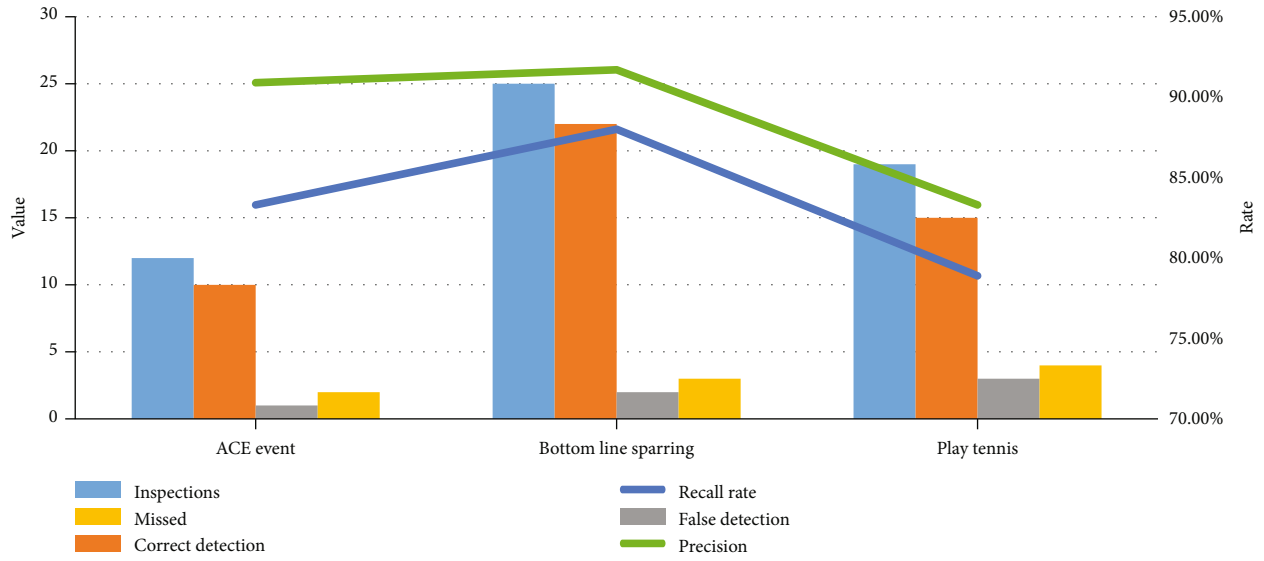

Figure 7: Detection results of exciting events in tennis video.

relatively high, while the detection efficiency of 84% and 80% of the ACE ball event and the tennis event is not very good. The main reasons are as follows:

(1) The duration of the successful ACE ball through the second serve is relatively long, and it is easy to miss the test

(2) It is easy to confuse the batting sound and the referee's shouting when the player fails to serve the ball, which reduces the detection rate of the batting sound

(3) The logo-based slow motion detection method needs to be further strengthened to reduce the missed detection rate and false detection rate of slow motion

In summary, the multimodal fusion-based tennis video event detection method constructed in this paper still has some areas to be improved and improved. However, compared with methods that rely solely on video features, audio features, and text features to detect exciting events, the detection efficiency is much improved. Therefore, the overall detection effect is satisfactory.

5. Conclusions

Applying the audio information of the lens to the lens class can better assist in the detection of some wonderful events in tennis. This article extracts audio signals from the game shots in the shot classification table and identifies the cheers and batting sounds in the audio signal of each game shot. Support vector machine (SVM) is a new machine learning method developed based on statistical learning theory. Through experiments, we found that the radial basis function has better classification performance and calculation speed in speech classification. In order to overcome the limitation that SVM can only handle two classification problems, this paper uses the method of combining SVM and decision tree-SVM decision method to deal with audio multiclassification problems. And according to this method, a SVM classifier is constructed to identify three audio categories: Silence/Non-Silence, Cheer/Non-Cheer, and Rally/Non-Rally. These three SVM classifiers are integrated to realize the recognition of the audio segment type of the tennis match. The experimental results show that it is feasible to use the SVM decision tree multilevel classifier to identify the audio segment type in the tennis match. The disadvantage of HMM is that it only depends on each state and its corresponding observation object: the sequence labeling problem is not only related to a single word but also related to the length of the observation sequence, the context of the word, and so on. The objective function and the prediction objective function do not match: what HMM learns is the joint distribution $P(Y, X)$ of the state and the observation sequence, and in the prediction problem, what we need is the conditional probability $P(Y \mid X)$.

Data Availability

Data sharing is not applicable to this article as no datasets were generated or analyzed during the current study.

Disclosure

We confirm that the content of the manuscript has not been published or submitted for publication elsewhere.

Conflicts of Interest

There are no potential competing interests in our paper.

References

[1] G. A. Rovithakis, M. Maniadakis, and M. Zervakis, "A hybrid neural network/genetic algorithm approach to optimizing feature extraction for signal classification," *IEEE Transactions on Systems, Man and Cybernetics, Part B (Cybernetics)*, vol. 34, no. 1, pp. 695–703, 2004.

[2] K. Zupanc and Z. Bosnic, "Automated essay evaluation with semantic analysis," *Knowledge-Based Systems*, vol. 120, pp. 118–132, 2017.

[3] R. Jenke, A. Peer, and M. Buss, "Feature extraction and selection for emotion recognition from EEG," *IEEE Transactions on Affective Computing*, vol. 5, no. 3, pp. 327–339, 2017.

[4] S. Ding, S. Qu, Y. Xi, and S. Wan, "A long video caption generation algorithm for big video data retrieval," *Future Generation Computer Systems*, vol. 93, pp. 583–595, 2019.

[5] N. N. Hurrah, S. A. Parah, N. A. Loan, J. A. Sheikh, M. Elhoseny, and K. Muhammad, "Dual watermarking framework for privacy protection and content authentication of multimedia," *Future Generation Computer Systems*, vol. 94, pp. 654–673, 2019.

[6] N. Krishnaraj, M. Elhoseny, E. L. Lydia, K. Shankar, and O. ALDabbas, "An efficient radix trie-based semantic visual indexing model for large-scale image retrieval in cloud environment," *Software: Practice and Experience*, vol. 51, no. 3, pp. 489–502, 2021.

[7] J. Korpela, R. Miyaji, T. Maekawa, K. Nozaki, and H. Tamagawa, "Toothbrushing performance evaluation using smartphone audio based on hybrid HMM-recognition/SVM-regression model," *Journal of Information Processing*, vol. 24, no. 2, pp. 302–313, 2016.

[8] D. Lijuan, X. Tao, W. Chunpeng, M. Wei, M. Jun, and Q. Honggang, "Integration of spatial and temporal characteristics of the video image visual significant degree detection method," 2016.

[9] D. Pan and G. Weijun, "Sports video classification method based on hidden Markov model," *Journal of Natural Science of Xiangtan University*, vol. 39, no. 1, pp. 73–77, 2017.

[10] S. Xinyi, R. Wang, and Z. Hongxiang, "An improved K-means clustering algorithm," *Computer and Digital Engineering*, vol. 46, no. 4, pp. 682–685, 2018.

[11] Z. Jin and C. Zemao, "Anomaly detection algorithm based on improved K-means clustering," *Computer Science*, vol. 43, no. 8, pp. 258–261, 2016.

[12] Z. Rongjuan, C. Xie, and H. Fenghua, "Player detection and tracking in tennis video," *Journal of Yanbian University (Natural Science Edition)*, vol. 45, no. 2, pp. 161–165, 2019.

[13] Q. Datong, Z. Sen, Q. Zhenggang, and C. Shujiang, "Driving condition construction method based on K-means clustering algorithm," *Journal of Jilin University (Engineering and Technology Edition)*, vol. 46, no. 2, pp. 383–389, 2016.

[14] C. Xu, "Semantic analysis of tennis audio based on hidden Markov model," *Information Technology*, vol. 8, pp. 103–106, 2019.

[15] V. N. Phu, V. T. N. Tran, V. T. N. Chau, N. D. Dat, and K. L. D. Duy, "A decision tree using ID3 algorithm for English semantic analysis," *International Journal of Speech Technology*, vol. 20, no. 3, pp. 593–613, 2017.

[16] J. John and C. K. Raju, "Design and comparative analysis of mobile computing software framework," in *2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT)*, Coimbatore, India, 2018.

[17] J. E. Bibault, P. Giraud, and A. Burgun, "Big data and machine learning in radiation oncology: state of the art and future prospects," *Cancer Letters*, vol. 382, no. 1, pp. 110–117, 2016.

[18] A. Buczak and E. Guven, "A survey of data mining and machine learning methods for cyber security intrusion detection," *IEEE Communications Surveys& Tutorials*, vol. 18, no. 2, pp. 1153–1176, 2017.

[19] C. Helma, T. Cramer, S. Kramer, and L. De Raedt, "Data mining and machine learning techniques for the identification of mutagenicity inducing substructures and structure activity relationships of noncongeneric compounds," *Journal of Chemical Information and Computer Sciences*, vol. 44, no. 4, pp. 1402–1411, 2018.

[20] X. Yuan, Z. Ge, L. Ye, and Z. Song, "Supervised neighborhood preserving embedding for feature extraction and its application for soft sensor modeling," *Journal of Chemometrics*, vol. 30, no. 8, pp. 430–441, 2016.

[21] S. Rabie, S. Aridhi, E. M. Nguifo, and M. Maddouri, "Feature extraction in protein sequences classification: a new stability measure," *Medical Physics*, vol. 37, no. 6, pp. 683–689, 2018.

[22] H. Zhao, Z. Wang, and F. Nie, "Orthogonal least squares regression for feature extraction," *Neurocomputing*, vol. 216, no. DEC.5, pp. 200–207, 2016.

[23] V. A. Nugroho, D. P. Adi, A. T. Wibowo, M. Y. T. Sulistyono, and A. B. Gumelar, "Klasifikasi jenis pemeliharaan dan perawatan container crane menggunakan algoritma machine learning," *MATICS*, vol. 13, no. 1, pp. 21–27, 2021.

[24] L. Pitak, K. Laloon, S. Wongpichet, P. Sirisomboon, and J. Posom, "Machine learning-based prediction of selected parameters of commercial biomass pellets using line scan near infrared-hyperspectral image," *Processes*, vol. 9, no. 2, pp. 316–320, 2021.