Video Background Music Recognition and Automatic Recommendation Based on GMM Model

Wei Zhou, Ke Ma*

Minjiang University, Tsai Chi-Kun Acadamy of Music, China,4350108
E-mail: 2515@mju.edu.cn, lsl851219@163.com
*Corresponding author

Keywords: background music, recommendation, chinese music, recognition, and gaussian mixture model (GMM)

Received: April 19, 2023

Recognizing background music in videos is a widely utilized technology in the global music business. With the use of classification, the data about the audio signal's frequency response, orchestration, and temporal structure is represented. In the beginning, identification was a human process. This operation may now be carried out autonomously because of developments in technologies and signal-processing techniques. Due to the widespread utilization of social networks, many smartphones come with a video-shooting feature that people often employ to create user-generated entertainment and communicate it with others. Nonetheless, it might be difficult to choose background music that complements the subject. Those who want to include background music in their videos must actively search for the audio. Nevertheless, since it is a procedure that requires a lot of time and effort, the emphasis of this study is on the construction of a system that will assist people in more easily and quickly obtaining the proper background music for their interests. For automatic recognition of video background music and recommendation, we implemented a Gaussian mixture model (GMM). Using principal component analysis, audio characteristics were recovered for effective recognition. The outcomes were assessed using performance measures and contrasted with previously used methods. The findings indicate that the suggested GMM produces superior performance.

Povzetek: Članek obravnava prepoznavanje in avtomatsko priporočanje ozadne glasbe v videih. Uporablja se model Gaussian Mixture Model (GMM) za prepoznavanje značilnosti zvoka. Rezultati kažejo, da predlagani GMM prinaša boljše rezultate v primerjavi z obstoječimi metodami.

1 Introduction

Music is an expressive kind of art that reflects real-life human emotions by using sound as a medium of expression. Being a significant part of human spirituality, music has always played a significant role in everyday life. Music may help individuals focus, reduce stress from work and school, and be healthy for their physical and mental health. People may experience auditory pleasure and spiritual fulfillment through music, as well as aid with the release of negative emotions like melancholy, loneliness, and depression as well as an increase in vigor and desire. Digital media resources for audio and video have strong transmission channels and easy storage mediums thanks to the Internet's fast development of these technologies, and digital content supplies and online music streaming entertainment consumers have seen tremendous expansion. The digital music age has begun [1]. Users' interests in video background music increasingly develop their styles as people's willingness to collect and keep video background music grows, which poses significant challenges for the preferences suggestion function of music-listening software. The system's suggested songs for users are now more accurate thanks to several manual and easy mining techniques, but the architecture of such video music in the background itself has not seen a proportional improvement in terms of analysis and processing power. Considering the connotations of the video background music themselves, it's possible that only a skilled composer or music critic could properly get the essence of such video background music and be able to identify its genre while listening to it [2].

Ancient music; environment; sustainable development With the fast expansion of China's economy, traditional music has had to contend with an increasingly acute issue of existence; many of these genres are now at their breaking point or in danger of extinction. The challenge facing China's music research community is how to specifically focus on cultural heritage while the country's economy develops quickly. Traditional Chinese music must be protected, and this process must also include the protection of related ecosystems. There is essentially no crisis around the loss of musical resources or concerns over

the preservation of cultural traditions in this unique historical setting of identity [3]. The atmosphere and themes that are intended to be represented, the majority of which consist of passive perceptions, can only be approximately heard by ordinary users when these crucial bits of information are present. The isomorphic relationship between the background music for a video and human sentiments can only be understood by artists. Humans are limited in their ability to comprehend background music in videos because of this, and everyone is different. As digital technology has advanced, it has been applicable in many spheres of social and economic life. Finding the isomorphic link between video background music and the emotions and feelings of humans using computers has however become a way to address these issues [4]. The most important tools for assisting users in locating pertinent information in vast information spaces are recommender systems. From the user's point of view, the author looked at this popularity bias in the movie industry. The researchers have shown that customers who like uncommon goods are often underserved by modern recommendation systems. Recreate the research and carry it out in the musical field. Music recommendations differ from, say, movie recommendations in several ways, including the sheer volume of options [5].

Personal attributes have been extensively studied in the fields of Communication and Recommendation Systems. The consequences of both personal traits and adaptability on affirmation, apparent diversity, and beliefs have been examined in the Music Recommendation Systems domain. The concept of educating users about the workings of recommender systems has gained more attention in recent years. Small worlds' visual interpretations of social connections and similarities between users and others helped to make the suggestion process transparent [6]. Music is a domain where non-standard solutions to the classic recommendation problem can be helpful because of the special needs of music consumers. To illustrate these points, let's look at how music stacks up against online movies and products, two other popular content kinds promoted by recommender systems. Movies, for instance, usually demand a person's undivided attention for 1-3 hours, and one movie is normally watched at a time. On the reverse hand, music is something that can be enjoyed at any time of day and in almost any setting. Similarly, internet buying is generally a concentrated behavior that most individuals participate in for a shorter amount of time contrasted to music listening. Both are subject to personal preference, but when it comes to music, listeners typically let their emotions rather than their intellect guide their decisions [7]. Algorithms used in recommendation systems allow for precise recommendations to be made based on users' tastes and requirements. They have a profound impact on our everyday lives, guiding our video and music consumption and our online purchasing decisions, yet are rarely given the credit they deserve. Considering the many uses, the design of such systems involves a variety of choices and strategies. There were historically two major schools of thought when it came to making recommendations. Similarities in the items' properties are used to make recommendations in content-based filtering. A user is more likely to like B if A is liked and B is also liked. Collaborative filtering, on the other hand, makes advantage of users' shared preferences to generate suggestions based on their shared tastes in music, media, and products. Users A and B may share interests if they are similar [8].

With audio content analysis, fresh recordings may be processed and categorized mechanically. This book just touches on some of the most popular musical genres since there are so many and new ones are continually appearing. Smart infotainment systems and audio streaming services rely heavily on music genre identification for music retrieval, suggestion, and customization. One challenge of categorizing digital data into predetermined classifications is the automatic identification of music genres. Digitization of the analog stream is a must. When the signal is processed, musical features are extracted. Finding the right set of characteristics is essential for accurately characterizing the various types of music. The gathered information is used as input in the categorization process [9]. Popular uses of machine learning include genre classification and music recommendation systems. The author framed the task of identifying musical genres as a pattern recognition challenge. They have used Gaussian classifiers (GS), and Gaussian mixture models (GMM), to extract rhythmic texture information. This article provides a comprehensive explanation of the beginning process database's characteristics utilized for music classification, as well as the numerous classifiers available for usage with this database. Convolution neural networks have recently been used to categorize musical genres [10]. For efficient recognition and automatic recommendation, we employed the Gaussian mixture model (GMM).

2 Video background music recognition

To develop an advanced background music structure based on deep learning and the Internet of Things, it is initially essential to evaluate and explain the native extraction of features of the scale-invariant addition of new features algorithm, the designation greatness of the support vector machine (SVM), and the effectiveness of the deep learning-based convolutional neural network within multi-scale extracting the features. The Smart Home is then equipped with a sophisticated background music system. On this foundation, a feature extraction method that extracts the essential features of the scene photos and is reliant on the intermediate feature structure is suggested [11]. They first determine the rhythmic correlations between the provided video and the background music to

develop the background music that goes with it. Next, they suggest a Controllable Music Transformer, which allows for both local administrations of the aforementioned rhythmic aspects and global control of the user-specified music genre and instrument employed. Both objective and subjective assessments demonstrate that the created background music has exceptional musical quality while also being satisfactorily compatible with the input videos [12]. The Drop Pathway methodology, which randomly removes the Audio route during training, was suggested in the study as a useful regularization method. They use hierarchy audiovisual synchronizing to learn shared audiovisual properties, drawing inspiration from earlier neuroscience work. They provide cutting-edge findings on six datasets for video action categorization and detection, carry out in-depth ablation experiments, and demonstrate the generalizability of AVSlowFast, a self-supervised audiovisual transfer learning algorithm [13].

Neural networks, which become massive groupings of computer components intended to trigger brain activity, are used in AI-based music production. To check whether the neural network (NN) recognizes patterns the same as the human brain does after repeated exposure to novel stimuli, it may be subjected to musical stimulation. It will ultimately learn how to use them. According to experts, until AI initially copies a human-created data gathering, it will not be able to produce music. by offering a conceptual framework for handling multimedia data [14]. They propose the KT-Speech-Crawler method for automatically building datasets for speech recognition using YouTube crawlers. We provide several filtering and post-processing procedures that extract data for end-to-end neural voice recognition model training. Readings and spontaneous speech are captured in automatically gathered samples under a variety of circumstances, such as background noise, music, recordings from a distance microphone, and a range of accents and reverbs [15]. The prospect of generally applicable solutions offered by machine learning techniques is predicated on understanding from just labeled samples. Convolution recurrent neural networks, which have demonstrated success in related fields like recognition of handwritten texts, are the deep neural networks that they propose to apply in the paper [16].

They suggest a technique that utilizes time-domain source segregation for enhancing ASR when background music is present. They retrieve the voice signal from the speech-music mixture in the waveform's domain using Conv-TasNet, a separated network that has attained cutting-edge performance for multi-speaker source segregation. Moreover, they suggest employing both separation and ASR goals to jointly fine-tune a which was before Conv-TasNet front-end with a consideration ASR back-end. They tested our approach using ASR, combining audio files with soundtracks from a broad range of Japanese cartoons [17]. In the study, they present a synthesis system for interactive background music that is driven by visual information. In two stages—scene visual analysis and music in the background synthesis—they use a cascade technique to create the music. To begin, they use neural networks to assess the emotion of the source scene in search of a deep learning-based solution. Second, to improve the consistency of feeling between auditory and visual criteria and music continuity, actual background music is generated by maximizing objective functions that directs the choosing and movement of music clips [18].

Table 1: Summary of related works

Reference	Description	Limitation
[11]	To generate a smart background music player based on Internet of Things (IoT) and deep learning. To extract the underlying features of scene images, a feature extraction method based on the middle-level feature framework is suggested. The anti-interference, toughness, and recognition capabilities of the provided method are all superior.	compatibility tests and fault tolerance are not explained
[12]	They proposed a Controllable Music Transformer (CMT). To create background music that matches the provided video.	Combined training data from videos and music may be included in future studies.
[13]	They proposed AVSlowFast Networks. To learn joint audiovisual features, we employ hierarchical audiovisual synchronization.	The backbone may be included in further studies

[14]	They proposed AI BD and Discussed the use of music in therapy and its apparent benefits on behavior. The suggested model outperforms the latest research in every metric, including performance (97.2%), reliability (95.6%), and survivability (98.2%).	Artists who love music and its psychological benefits may pursue it. Instead of taking music education lessons, one might benefit others by volunteering.
[15]	They suggested KT-Speech-Crawler and developed an open-source solution that generates datasets automatically for end-to-end neural voice recognition system training. On the WSJ and TED datasets, we showed the obtained samples' utility.	The difficulty of soliciting enough user contributions for a dataset with a sufficient number of speakers, accents, surroundings, and recording conditions
[16]	The proposed convolutional recurrent neural networks (CRNN) were developed with the intent of accurately transcribing sheet music into a format that could be handled by a computer.	Measure the impact of an LM.
[17]	They propose integrating the Conv-TasNet system for speech-music separation in the temporal domain with an attention-based ASR system to separate a speech-music mashup into its individual voice signal in the waveform domain. Their strategy succeeds throughout the board, including in the presence of classical, jazz, and popular music.	The signal-to-noise ratio is slightly high.
[18]	Introduced interactive background music synthesis algorithm guided by visual content. Creating synthesized music for various scenarios can advance study in a variety of subjects.	Scenarios with various lighting, backdrops, and layouts that correspond to various intended uses and styles

3 Materials and method

These days, micro-videos are a progressive mode on the Internet. Micro-videos enable individuals to record and communicate their everyday routines because they comprise rich image features, audible noises, and linguistic labels and characterizations in comparison to messages or images. Hence, the automatic recommendation is very important. We presented the GMM for video background music recognition and automatic recommendation. Figure 1 shows the flow of the proposed method.

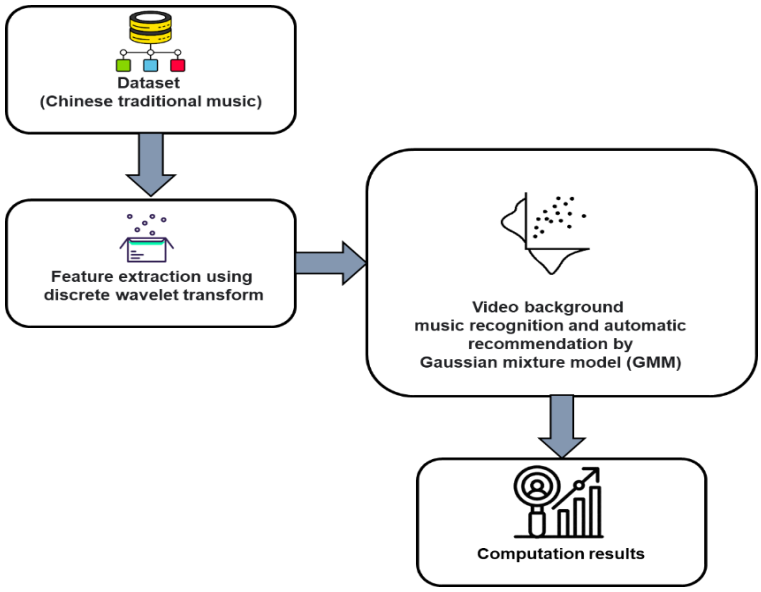

Figure 1: Flow of the proposed method

A. Dataset

ChMusic is a collection of traditional Chinese music used to build and assess musical device identification models. There are 11 different musical equipment used in this study. There are five traditional Chinese music specialists for each musical equipment, totaling 55 for the database. In this collection, every musical sample was exclusively performed by one device. Each piece of music is archived as an a.wav file. These recordings are designated in the format "x.y.wav," where "x" denotes the equipment number, which ranges from 1 to 11, and "y" denotes the music number, which ranges from 1 to 5. These recordings were created using a twin-channel recorder with a 44100 Hz sampling rate. These musical snippets range in length from 25 to 280 seconds [19].

B. Feature extraction using discrete wavelet transform

A non-stationary timeframe modeling approach suited for application with musical signals is wavelet transformation. It is a helpful tool for dividing and organizing non-stationary signals into their many harmonic components across a range of timeframes. A discrete time series, f(n), which in this research is the discrete signal of f(x), analyzed at 500Hz, can be mathematically decomposed into its sub-bands of wavelet correlations that comprise the function using the discrete wavelet transform. The parent wavelet $\varphi_{s,i}(x)$ compression, and interpretation may be used to calculate the wavelet decomposition, as indicated in (1), where $s, i \in r, c > 0$, and r is the wavelet space and S and C are the sizing and changing factors, correspondingly.

$$\varphi_{s,i}(a) = \frac{1}{\sqrt{s}}\varphi\left(\frac{t-\tau}{s}\right) \qquad (1)$$

The discrete signal f(n) is continuously filtered up to a set level n to calculate the deconstruction. The filtering is made up of a high pass filter to get the comprehensive factor (CF) and a low pass filter to get the estimation factor (EF). Because of the frequency object's reduction by 50 percent, the signal is quantized by half after every filter degree, or level n-1, to account for this.

As of daubechies 4 (db4) technique is best suited to handle music signals, it is employed as the parent wavelet in this research. The bandwidth range of the incoming signal fn is 0-500Hz. The signal must be divided into the lower bandwidth delta band at stage 8, with the relevant range for music signals being 0-60Hz, however as the important bandwidth range is in the alpha beat, the filtering signal will only be completely divided into the alpha band at stage 6 in CF. Levels 1, 2, and 3's complexity coefficients are regarded as noise since their frequencies did not fall inside the 0–60 Hz range of music. The deconstructed signal's wavelet value is still too high and unsuitable for direct feature identification using a computational model. As a result, feature separation is used to condense the characterization of a vast amount of data into the signal's depiction set of characteristic vectors. Both the time domain and the harmonic domain features of the characteristic may be retrieved. The empirical technique of the period character is the most straightforward and widely utilized characteristic to describe the big collection of data. It is possible to employ quantitative characteristics like average, midpoint, modal, standard deviation, highest, and lowest. In this research, the wavelet parameter discrete-time sequence standard deviation is calculated utilizing equation (2), where n denotes the discrete signal length and x is the specific n's signal intensity.

$$\sigma^2 = \frac{1}{N-1}\sum_{j=0}^{N-1}\left(a_j - \mu\right)^2 \qquad (2)$$

C. Video background music recognition and automatic recommendation by Gaussian mixture model (GMM)

Numerous music recognition and recommendation challenges have been effectively completed using gaussian mixture modeling. In this particular piece of research, GMM is brought to work to describe music in the form of a probabilistic distribution function (PDF) by utilizing a continuous mixture of Gaussian element PDFs. Equation 3 defines a Gaussian combination volume as the balanced average of M element values.

$$p(A|\lambda) = \sum_{i=1}^{M} x_i y_i(A) \qquad (3)$$

where $B_i(A)$ are the element intensities and ai are the mixture values, and x is a D-dimensional arbitrary variable. Every element concentration is a D-variate variant of the Gaussian distribution.

$$B_i(A) = \frac{1}{(2\pi)d/2\sqrt{|\Sigma_i|}} e^{(-0.5(A-M_i)^t \sum_i^{-1}(A-M_i))} \qquad (4)$$

Given a dxd correlation matrix i and a dx1 median vector M_i, the mixing weights x_i meet the condition that: $\Sigma_i x_i = 1$by being positive values. While a complete correlation matrix, or a correlation matrix with all of its members, is supported by the generic modeling form, only horizontal correlation structures are often utilized since they are less computationally difficult. The average vectors, correlation vectors, and mixture values from all element concentrations may all be used to form a single GMM. As a result, a GMM may be defined as = x_i , M_i , $\lambda = \{x_i, \mathrm{M}, \Sigma_i\}, \forall i = 1, \ldots\ldots, m.$

The Expectation-Maximization (EM) technique is used to calculate the variables of the GMM that best fit the dispersion of the training matrices given a collection of training matrices $A = \{A_1, \ldots., A_n, \ldots., A_n\}$ and a GMM design. The training of an EM-based GMM is shown in equation 5. It is displayed that the initialization of the GMM settings affects how the EM behaves. The k-means technique is widely used to generate the first GMM structure.

The following method may be utilized to automate music recognition using GMM. Every music in the dataset is subjected to the extraction of a collection of feature matrices. The GMM design λ_i is trained using these relevant features and then saved in the dataset. Automatic music recognition may be done when the algorithms for every piece of music are placed in the collection. The algorithm is shown a song clip and asked to identify it. From the music section, the same feature matrices $B = \{b_1, \ldots., b_2, \ldots., and\ b_N\}$ are retrieved. We compute the log-probability $logp(Y|\lambda_i) = \sum_{n=1}^{N} logp(Yn|\lambda_i)$ for all types i in the dataset to identify and suggest the music in the collection that matches these feature matrices superior.

4 Results and discussion

In this section, we evaluate the proposed GMM effectiveness. The assessment metrics are recognition accuracy, precision, recommendation quality, error rate, and computation time. The existing techniques used for comparison are Deep learning Internet of Things (DL-IoT) [20], Factorization Machines Approach (FMA) [21], and Hybrid Recommendation Algorithm Based on Collaborative Filtering (HRACF) [22].

A. Recognition Accuracy

Recognition accuracy in music or humming refers to how accurate or correct it is, even down to the slightest of details. Refers to how closely a measurement is to its agreed-upon value and describes how much the outcome of a measurement corresponds to the proper value or a standard. Accuracy ensures a higher recognition quality. Figure 2 indicates the recognition accuracy of the suggested method. Table 1 shows the result of the proposed method. It illustrates that the suggested method is more accurate than the existing method.

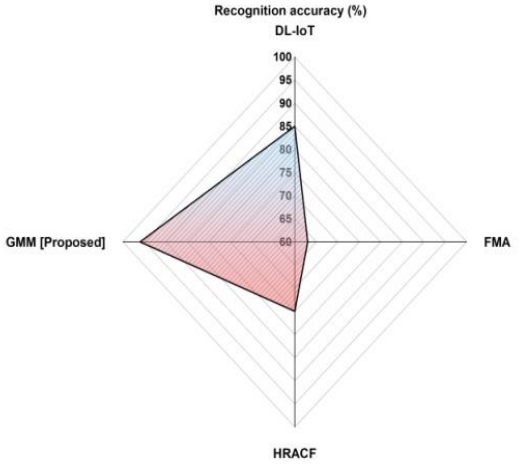

Figure 2: Recognition accuracy of the proposed and existing techniques.

Table 1: Results of recognition accuracy

Methods	Recognition accuracy (%)
DL-IoT	85
FMA	63
HRACF	75
GMM [Proposed]	96

B. Precision

Precision is the degree to which two or more measurements are similar to one another down to the very last digit. The percentage of relevant objects found among all items found is what is known as precision, which is an indicator of exactness. Figure 3 indicates the precision of the suggested method. Table 2 shows the result of the proposed method. It illustrates that the suggested method is more reliable than the existing method.

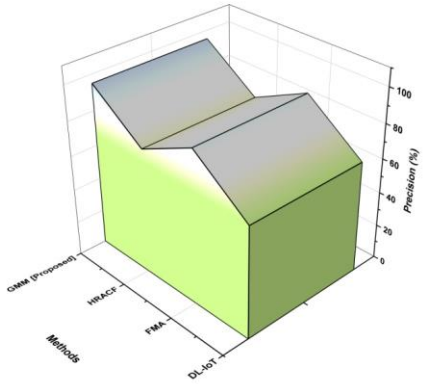

Figure 3: Precision of the proposed and existing techniques.

Table 2: Results of precision

Methods	Precision (%)
DL-IoT	65
FMA	89
HRACF	73
GMM [Proposed]	95

C. Recommendation quality

The quality of the music that is recommended is what determines the quality of the music that is offered by the model. It refers to the technique's ability to give users authentic soundtrack music in the background. Figure 4 indicates the recommended quality of the suggested method. Table 3 shows the result of the proposed method. It denotes that the suggested method is more efficient than the existing method.

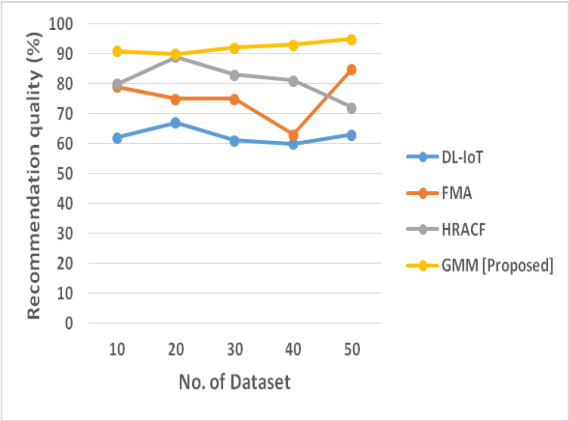

Figure 4: Recommendation quality of the proposed and existing techniques.

Table 3: Results of recommendation quality

No. of Dataset	Recommendation quality (%)			
	DL-IoT	FMA	HRACF	GMM [Proposed]
10	62	79	80	91
20	67	75	89	90
30	61	75	83	92
40	60	63	81	93
50	63	85	72	95

D. Error rate

The error rate is a measurement of how much a model deviate from the real mode in its predictions. The proportion of processing mistakes that a department makes is known as its error rate. Processing mistakes should be avoided at all costs because of the expense of fixing them. On the other hand, the error rate may be calculated by dividing the total amount of wrong predictions on the test set by all of the predictions on the test set. Figure 5 indicates the error rate of the suggested method. Table 4 shows the result of the proposed method. It illustrates that the suggested method is low error rate than the existing method.

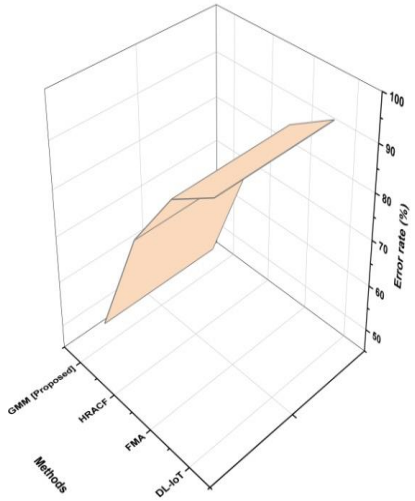

Figure 5: Error rate of the proposed and existing techniques.

Table 4: Results of the error rate

Methods	Error rate (%)
DL-IoT	95
FMA	89
HRACF	75
GMM [Proposed]	50

E. Computation time

The amount of time needed to complete a computing operation is called the computation time. When a calculation is represented as a series of rule applications, the computation time is inversely correlated with the number of rule applications. The execution time of each period during simulation is the computation cost. The amount of time needed for a computer to recognize a certain set of processes. Figure 6 indicates the computation time of the suggested method. Table 5 shows the result of the proposed method. It illustrates that the suggested method takes less time than the existing method.

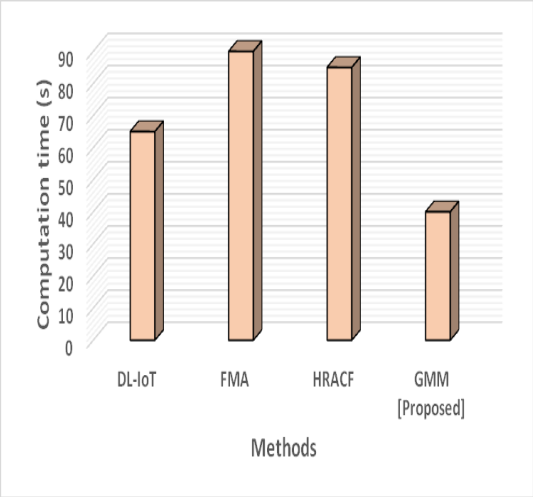

Figure 6: Computation time of the proposed and existing techniques.

Table 5: Results of computation time

Methods	Computation time (s)
DL-IoT	65
FMA	90
HRACF	85
GMM [Proposed]	40

4.1 Discussion

Music expresses emotions through sound. Digital music is expanding, but traditional music faces challenges. Video background music's connection to emotions is understood by skilled composers. Recommender systems help users find music. Genre identification and machine learning aid in music recommendation. Study [11] obtained in maps of the recognition effects based on several feature extraction techniques is 86.3%. UCF and HMDB have an initial learning rate of 0.005/0.01, and their decay schedule is a half-period cosine [13]. When compared to other existing models, the AI-BD approach is 97.8% more efficient and accurate. Analysis of performance, reliability, and survivability are all improved by 98.2%, 95.6%, and 97.2%, respectively [14]. The error rate of the symbol level was from 25.7% to 7.0% in the study [16]. The Values of performance metrics for our proposed method were obtained in terms of accuracy (96%), precision (95%), recommendation quality (95%), Error rate (50%), and Computation time (40S).

5 Conclusion

People's interests in video music in the background increasingly develop their unique styles as one's ability to acquire and retain video theme music grows, which poses significant challenges for the desired recommendation functionality of music-listening programs. It takes a lot of time to suggest music manually. We presented an effective Gaussian mixture model for automatic recommendation and recognition (GMM). For the study, traditional Chinese music is used. For feature extraction, the discrete wavelet transform is utilized. Recognition accuracy, precision, recommendation quality, error rate, and computation time are used to assess the efficacy of GMM. The suggested GMM outperforms existing approaches like Deep Learning-Internet of Things (DL-IoT), Factorization Machines Approach (FMA), and Hybrid Recommendation Algorithm Based on Collaborative Filtering (HRACF). The results show that the suggested GMM is efficient in automated recognition and recommendation. Based on this research, many implementations are feasible. For instance, the GMM model may be used to create a smartphone application that enables people to choose the right background music once they have finished filming a video. Future research will transfer video and musical feature attributes to a GMM to develop discriminative feature descriptions.

Acknowledgement

2019 Minjiang University social science project: A Research on The Application of Chinese Traditional Music in Background Music (project number: MYS19039)

References

[1] Xu, K., (2021). Recognition and classification model of music genres and Chinese traditional musical instruments based on deep neural networks. Scientific Programming, pp.1-8.

[2] Kai, H., (2021). Automatic Recommendation Algorithm for Video Background Music Based on Deep Learning. *Complexity*, *2021*, pp.1-11.

[3] Huien, L., (2019). The ecosystem of Chinese Traditional Music and Its Sustainable Development.

[4] Ulaganathan, A.S. and Ramanna, S., (2019). Granular methods in automatic music genre classification: a case study. *Journal of Intelligent Information Systems*, *52*, pp.85-105.

[5] Kowald, D., Schedl, M. and Lex, E., (2020). The unfairness of popularity bias in music recommendation: A reproducibility study. In *Advances in Information Retrieval: 42nd European Conference on IR Research, ECIR 2020, Lisbon, Portugal, April 14–17, 2020, Proceedings, Part II 42* (pp. 35-42). Springer International Publishing.

[6] Millecamp, M., Htun, N.N., Conati, C. and Verbert, K., (2019), March. To explain or not to explain: the effects of personal characteristics when explaining music recommendations. In *Proceedings of the 24th International Conference on Intelligent User Interfaces* (pp. 397-407).

[7] Andjelkovic, I., Parra, D. and O'Donovan, J., (2019). Moodplay: interactive music recommendation based on artists' mood similarity. *International Journal of Human-Computer Studies*, *121*, pp.142-159.

[8] Fessahaye, F., Perez, L., Zhan, T., Zhang, R., Fossier, C., Markarian, R., Chiu, C., Zhan, J., Gewali, L. and Oh, P., (2019), January. T-recsys: A novel music recommendation system using deep learning. In *2019 IEEE international conference on consumer electronics (ICCE)* (pp. 1-6). IEEE.

[9] Jakubec, M. and Chmulik, M., (2019). Automatic music genre recognition for in-car infotainment. *Transportation Research Procedia*, *40*, pp.1364-1371.

[10] Sheikh Fathollahi, M. and Razzazi, F., (2021). Music similarity measurement and recommendation system using convolutional neural networks. *International Journal of Multimedia Information Retrieval*, *10*, pp.43-53.

[11] Wen, X., (2021). Using deep learning approach and IoT architecture to build the intelligent music recommendation system. *Soft Computing*, *25*, pp.3087-3096.

[12] Di, S., Jiang, Z., Liu, S., Wang, Z., Zhu, L., He, Z., Liu, H. and Yan, S., (2021), October. Video background music generation with controllable music transformer. In *Proceedings of the 29th ACM International Conference on Multimedia* (pp. 2037-2045).

[13] Xiao, F., Lee, Y.J., Grauman, K., Malik, J. and Feichtenhofer, C., (2020). Audiovisual slow fast networks for video recognition. *arXiv preprint arXiv:2001.08740*.

[14] Sun, W. and Sundarasekar, R., Research on pattern recognition of different music types in the context of AI with the help of multimedia information processing. *ACM Transactions on Asian and Low-Resource Language Information Processing*.

[15] Lakomkin, E., Magg, S., Weber, C. and Wermter, S., (2019). KT-speech-crawler: Automatic dataset construction for speech recognition from YouTube videos. *arXiv preprint arXiv:1903.00216*.

[16] Calvo-Zaragoza, J., Toselli, A.H. and Vidal, E., (2019). Handwritten music recognition for mensural notation with convolutional recurrent neural networks. *Pattern Recognition Letters*, *128*, pp.115-121.

[17] Woo, J., Mimura, M., Yoshii, K. and Kawahara, T., (2020), December. End-to-end music-mixed speech recognition. In *2020 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)* (pp. 800-804). IEEE.

[18] Wang, Y., Liang, W., Li, W., Li, D. and Yu, L.F., (2020), October. Scene-aware background music synthesis. In *Proceedings of the 28th ACM International Conference on Multimedia* (pp. 1162-1170).

[19] Gong, X., Zhu, Y., Zhu, H. and Wei, H., (2021), September. Chmusic: a traditional Chinese music dataset for evaluation of instrument recognition. In Proceedings of the 4th International Conference on Big Data Technologies (pp. 184-189).

[20] Wen, X., (2021). Using deep learning approach and IoT architecture to build the intelligent music recommendation system. Soft Computing, 25, pp.3087-3096.

[21] Singh, J. and Sajid, M., (2021), January. Factorization Machine Based Music Recommendation Approach. *In 2021 6th International Conference on Inventive Computation Technologies* (ICICT) (pp. 618-622). IEEE.

[22] Wenzhen, W., (2019), August. Personalized music recommendation algorithm based on hybrid collaborative filtering technology. *In 2019 International Conference on Smart Grid and Electrical Automation* (ICSGEA) (pp. 280-283). IEEE.