Intelligent Directing System for Music Concert Scene Based on Visual and Auditory Information

CAIHONG WANG*, Communication University of China, China
CHUHAN QIU†, Communication University of China, China
WANXIN XU, Communication University of China, China
WEIJIE ZHU, Communication University of China, China
HANLU LUO, Communication University of China, China
LEI CHEN, Communication University of China, China
ZIYI YU, Communication University of China, China
LIN FU, Communication University of China, China
XUANJUN CHEN, Communication University of China, China

We propose an intelligent directing system based on visual and auditory information, which mainly focuses on the concert scene. Our system performs real-time decoding of signals recorded and transmitted during the live concert, while providing audio and video streaming data. In this way, the system will extract the physical and musical features of the audio stream. These auditory information will determine the timing of camera to be switched. As for video streaming data, we choose to conduct real-time detection of data from different cameras from different viewpoints, including the blurring degree and short scale, to determine whether there is a high "switching confidence" in the current video channel. In addition, our system also classifies the instruments received at each video channel, to match the main instrument identified from the audio information. Whenever the live performance reaches a point where a camera switching is required, combining the above factors, our system will obtain a 'switching score' for each camera. Then the system will select the camera with the highest score for switching operation.

CCS Concepts: • **Human-centered computing** → **Scenario-based design**.

Additional Key Words and Phrases: intelligent directing, deep learning, music information retrieval, computer vision

ACM Reference Format:
Caihong Wang, Chuhan Qiu, Wanxin Xu, Weijie Zhu, Hanlu Luo, Lei Chen, Ziyi Yu, Lin Fu, and Xuanjun Chen. 2023. Intelligent Directing System for Music Concert Scene Based on Visual and Auditory Information. In . ACM, New York, NY, USA, 11 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

Live video directing, which refers to the process of directing live events such as concerts, broadcast or productions in real-time, involves controlling various aspects of the production, such as camera angles, volume and lighting. Generally speaking, in most cases, it requires professional directing personnel to operate the relevant equipment and this will consume a significant amount of human and material resources. A mature and reliable automatic broadcast system

Manuscript submitted to ACM

could significantly reduce the associated costs. What's more, algorithms can better capture some subtle factors of the current scene and make some correct operations that are difficult for humans to detect.

There is only a small amount of research related to intelligent broadcasting especially for concert scenes. In addition, these researchers were more concerned with visual features such as body movements, facial expressions, lip movements, etc. And they paid little attention to auditory information [11] [1] [8]. However, in the concert scene, the audience's actual experience of video channel switching is largely dependent on auditory information, especially where the importance of musical information is significantly increased. [4] added voiceprint features of the speakers to identify their identities, but doesn't apply to concert scenes.

Music concert is a pretty special scene for live broadcast because both the tempo and the melody of music will influence the directing. That's why we consider so many relevant characteristics in this study.

The remainder of this paper is divided into several sections. In Section 2, the proposed system is introduced and each part of it is explained. Section 3 described the experimental setup, and its results is presented in Section 4. At last, we have a discussion of the experimental results in Section 5 and the conclusion is set to Section 6.

2 METHOD

Here we use the block diagram to present our intelligent directing system in Figure 1, which has been implemented in Python.

The recorded signal from live concert will be decoded into audio streaming data and video streaming data individually, where in the video streaming data contains the transmission signals of all cameras on site when the audio streaming data contains only the stereo signal mixed by the mixer. According to the audio streaming data, our system judges the current switching rate based on music emotion recognition, which could be implemented through K-Nearest-Neighbors (KNN) algorithm [13]. However, for our directing system, too many emotion classifications wouldn't improve the switching effect, but may lead to an increase in computational complexity. So we reduce the number of KNN classifications and named the label three gears for switching rate. The physical features required for KNN include short-time Fourier Transform chromagram, Constant-Q chromagram and Chroma Energy Normalized, whose details will be described in the following sections. Besides, some physical features are also used as the features of the Support Vector Machine (SVM) algorithm to determine the predominant instrument of the current audio streaming [12]. For the musical information, our system mainly extracts its downbeat position, which determines the switching timing of the entire system. Rhythm and beat detection have many related works [5] [14] [6], and we mainly focus on the downbeat detection part.

Besides, in previous existing engineering practice [4], a 25 seconds delay was reserved for system to perform calculations and predictions, which actually conducted a real-time 'pre-analysis' on the data stream. To ensure the real-time performance of our system, we also set up a delayed broadcast time. Consider the length of analysis materials and processing time required by the system, we finally set the delay time as 30 seconds.

As previously mentioned, video streaming data contains signal from multiple cameras in the field. Therefore, the video streaming process in Fig. 1 actually involves simultaneous processing of multiple channels. Similar to audio streaming, there is also a process of instrument classification in the video streaming which is based on YOLO. We build our own dataset of musical instrument images, including over 2000 images of different musical instruments in total. The system determines the classes of instruments in the video frames captured by each camera and judges which camera has the highest similarity to the predominant instrument extracted from the audio streaming. Then, we combine the blur detection and shot scale recognition results which will ultimately provide a 'switching confidence' of each camera.

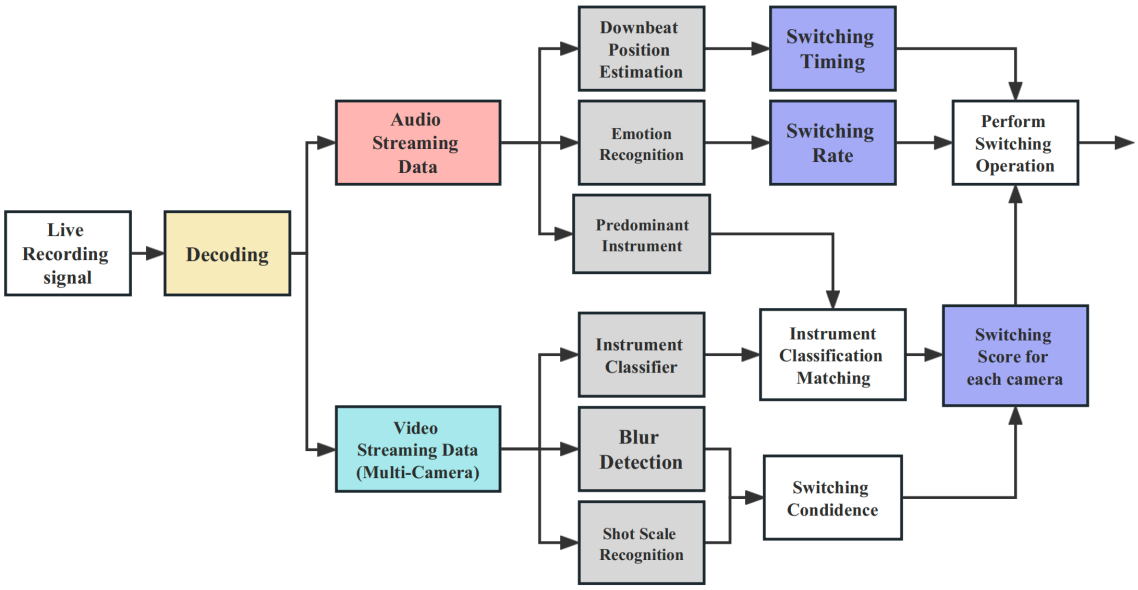

Fig. 1. Block diagram of intelligent directing system

Integrating the switching confidence and the matching results, the system comprehensively scores each camera and obtains a 'switching score' of them. Whenever the switching timing is about to arrive, the system will select the camera with the highest score for switching operations.

2.1 Switching Timing

There have already been many related studies on rhythm and beat detection. For traditional algorithms, most of them are based on extracting the fluctuation strength in the signal and estimating the beat period of it [5] [14]. By some extent, these traditional algorithms are similar to using Autocorrelation to calculate the fundamental frequency of the signal. However, with the increasing application of deep learning in audio signal processing, some research has begun to add neutral networks to rhythm and beat detection.

BeatNet is an online system for joint beat, downbeat, and meter tracking proposed by Heydari M. [6], which produces beat and downbeat activation using a CNN and RNN combination, and performs inference using two particle filtering stages: 'Beat Tracker Particle Filtering' and 'Downbeat Tracker Particle Filtering'. Besides, they proposed an information gate mechanism in the inference module to speed up the inference significantly, which makes BeatNet suitable for real-time applications. In this way, BeatNet can well meet the real-time processing need of our directing system.

In musical notation, a bar is the most basic regular rhythmic unit which reflects the fluctuation strength of music performance, and a bar usually starts with a downbeat. Through the estimation of downbeat activation by BeatNet, we use it to count the number of bars played. In previous directing cases, directors generally switched the cameras in a loop of several bars, such as four or eight. So we determine the switching timing by the position of downbeat, and adjust how many bars are used as a cycle to switch through the 'switching rate'.

Table 1. Relationship between rate level and frame/time interval

Rate Level	Frame Interval	Time Interval (s)
Fast	70 - 150	3 - 6
Middle	150 - 250	6 - 10
Slow	≥ 250	≥ 10

2.2 Switching Rate Adjustment

The switching rate adjustment of our directing system is inspired by music emotion recognition [13]. After our analysis of many directing cases, we finally came up to the conclusion that the speed of video channel switching is actually largely influenced by the intensity of music performance. Based on this premise, the task of switching rate adjustment can be converted to music emotion recognition. We analyze multiple instances of concert broadcast directing and summarized up to three levels of switching frequencies. Corresponding switching time interval has been shown in Table 1.

Based on the prediction result of KNN, the system will adjust the switching rate to the corresponding level in Table 1. We make the features that need to be extracted as few as possible in order to meet the real-time processing need of the system. So the features we ultimately chose include: short-time Fourier Transform chromagram, Constant-Q chromagram and CENS (Chroma Energy Normalization Statistics). Chroma refers to the respective pitch spelling attribute contained in the set of twelve-tone equal temperament and a pitch class is defined as the set of all pitches that share the same chroma [3].

To compute the chromagram, first aggregate all spectral information that related to a given pitch class into a single coefficient, then sum up all pitch coefficients that belong to the same chroma [9].

$$C(n,c) = \sum_{p\in[0:127]:p \bmod 12=c} Y_{LF}(n,p) \tag{1}$$

where Y_{LF} refers to the log-frequency spectrogram which has been discussed in [9], and $c \in [0,11]$ that corresponds to the twelve notes in each octave. We use Librosa, a widely used audio signal processing library of Python, to implement the extraction of these features.

2.2.1 Short-time Fourier Transform Chromagram. The STFT chromagram is the simplest chroma feature, which could be directly computed according to (1). But before that, we need to obtain the log-frequency spectrogram from the STFT X. The log-frequency spectrogram is defined by equation (2).

$$Y_{LF}(n,p) = \sum_{k\in P(p)} |X(n,k)|^2 \tag{2}$$

$P(p)$ refers to the frequency range of each MIDI pitch where $p \in [0,127]$. In this way, Y_{LF} actually assigns each spectral coefficient $X(n,k)$ to the pitch with center frequency that is closest to the centre frequency of MIDI pitches [9]. Besides, we use a hanning window with the size of 1024 samples, which corresponds to about 23ms at 48kHz, to compute the short-time Fourier transform.

2.2.2 Constant-Q Chromagram. Constant-Q transform, which is also known as CQT, uses a logarithmically spaced frequency axis which is similar to mel scale [2]. Similar to the STFT chromagram, The computation of constant-Q chromagram is based on constant-Q transform, which can be defined by equation (3).

$$X(k) = \frac{1}{N(k)} \sum_{n=0}^{N(k)-1} W(k,n)x(n)e^{-j2\pi Qn/N(k)} \tag{3}$$

wherein that

$$N(k) = \frac{Q}{f_k T} \tag{4}$$

T is the sampling time and Q refers to the ratio of frequency to bandwidth, which is defined as $\frac{f}{\delta f}$. f_k refers to the kth spectral component of constant-Q transform and is thus

$$f_k = (2^{1/24})^k f_{min} \tag{5}$$

After applying the short-time constant-Q transform to the signal, further calculation steps are the same as (1) and (2).

2.2.3 Chromagram Energy Normalization Statistics. CENS features are often used for audio matching and retrieval applications, which considers the short-time statistics over energy distributions within the chroma bands [10]. The calculation process of CENS includes following steps:

(1) Compute the normalization of each chroma vector which obtained;
(2) Quantify the amplitude based on suitably chosen thresholds in a logarithmic fashion;
(3) Apply smoothing and down-sampling operations as the subsequent steps.

2.3 Instrument Classification Matching

To determine the matching degree between multiple video streaming data and single audio streaming data, it is necessary to extract the predominant instrument from the mixed audio signal and multi-camera signals. In this section, we discuss the predominant instrument detection of audio and video respectively, and introduce a musical instrument image dataset for concert scenes which is constructed by ourselves.

2.3.1 Audio Streaming Identification. The extraction of predominant instrument in audio signal processing is also based on feature engineering, and for instrument recognition, the main features we focus should have the ability to reflect the timbre of sound, such as CENS, MFCCs, etc. Then, we use these features as the input to the machine learning model, which is used as a classifier to determine. [12] compared the performance of different machine learning algorithms in predominant instrument classification, and Support Vector Machine algorithm achieved the highest accuracy, though unable to distinguish between flute and organ properly.

Scikit-learn is an open-source Machine Learning Library which provides the implementation of SVM. We use Librosa to extract features from audio stream and input them into the SVM model. To further improve the effectiveness of the algorithm, in addition to MFCCs we add some chroma features to better capture the timbre characteristics in the audio stream, including the constant-Q chromagram and CENS mentioned earlier.

2.3.2 Video Streaming Identification. Common target detection models include R-CNN, SDD and YOLO. Among them, YOLO series algorithms are the most widely used in objective detection algorithm at present because of its fast detection speed and high accuracy. We use it for instrument detection. Wherein the input of YOLO is enhanced by Mosaic data, which can alleviate the rich data set and reduce the training time to some extent. There is currently no publicly available musical instrument dataset suitable for our needs. In order to meet the switching requirements of the system under the scene change, we manually collected 1675 pictures of musical instruments appearing in the video from the online concert video as a data set. The database contains 17 common musical instruments: bassoon, cello, clarinet, cello for double, flute, French horn, harp, marimba, oboe, percussion, piano, timpani, trombone, trumpet, tuba, viola and violin. We use Labeling to label the images in the dataset. Finally, the model of detecting the main musical instrument is obtained by training YOLO, which can locate the musical instrument in the picture, mark the position coordinates of the musical instrument, and output the label of the musical instrument classification.

2.4 Switching Confidence

During the switching process of broadcast directing, sometimes the on-site cameramen need to adjust the focus of the cameras, such as shifting the focus of camera from one object to another. However, during the adjustment process, there may be some degree of blurring in the camera image, which is absolutely necessary to avoid during the broadcast directing. Therefore, we propose a blur detection algorithm to determine whether the blurriness of each camera image has reached a threshold that cannot be switched. Besides, in order to increase the richness of the screen and enhance the excitement of the program, directors often use different scenes to connect different lenses. Therefore, we introduce a detection algorithm for recognizing the scene of the current camera image, together with the blur detection, as the 'switching confidence' in Figure 1.

2.4.1 Blur Detection. We mainly use OpenCV and Fast Fourier Transform (FFT) which are both implemented in NumPy for blur detection. After receiving a frame of video information, we first use FFT to covert the signal to the frequency domain, and remove the DC and low frequency components. Then we use inverse FFT to calculate the amplitude of the reconstructed frame and take the average value of it. This average value represents the blur value. Essentially, the blur value is a quantization of the frequency of the frame. The clearer the frame and the higher the frequency, and the larger the calculated value. Conversely, the smaller the calculated value. Therefore, by comparing the obtained blur value with a preset threshold value, the blur detection flag value for the frame can be obtained.

2.4.2 Shot Scale Recognition. The shot scale identification model totally has six classifications of shot scale: extreme-close-shot, close-shot, medium-shot, full-shot, long-shot and no-human-shot. The shot of figures' shoulder and above is identified as extreme-close-shot (ECS), waist and above shot is identified as close-shot (CS), knee and above shot is identified as medium-shot (MS) and the shot of whole body of the character is identified as full-shot (FS). Small range shot scale without clear character is identified as the long-shot (LS). Frames that are smaller than LS whose subject is not a person are identified as non-human-shot (OS).

We use ResNet-50 convolutional neural network. Part of the data of the existing video shot scale (VSS) was used for training. The data includes 24,000 horizontal screen video frames of various aspect ratios of sports programs, variety shows, films and TV dramas, ceremonies, news, documentaries and other programs. A random gradient descent (SGD) optimizer is used to randomly select a sample (x^i, y^i) from the training set each round to update the model parameters, where x is the input data and y is the label.

The gradient descent algorithm is used to iteratively solve the minimum loss function, so as to obtain the minimum loss function and the best model parameters. The parameter update can be written as (6).

$$\theta_{t+1} = \theta_t - \eta \nabla L(\theta_t; (x^i, y^i)) \tag{6}$$

where θ_t is the parameter vector at step t and θ_{t+1} is the parameter vector one step following.

2.5 Switching Score

Switching score is the indicator that ultimately determines which camera to perform the switching operation. By comprehensively considering the results of instrument matching and switching confidence, we can obtain the camera with the highest switching score among the currently switchable cameras. However, after our analysis of multiple cases of artificial broadcast directing, we think it is necessary to set several rules to limit the switching operation.

Fig. 2. indexes and probability of transition-probability-matrix

[7] used HMM model for modeling, integrating video information and text script information, and utilizing editing rules. Inspired from this, we propose that the switching operation between different shot scales conforms to a probability distribution model. We first assume that each shot scale is regarded as a state, and the switching between shots is regarded as a state transition, then we design a probability distributions matrix called 'transition-probability-matrix', to describe the entire switching objects and process, which is shown in Figure 2. This matrix is obtained by analyzing 5 professional broadcast concerts. We used a shots segmentation program to segment the shots of these concerts, and then collected data on types, numbers, and transitions of the shots. We set a threshold to calculate the similarity value of HSV (Hue/Saturation/Value) between the two frames before and after shot switching. The segmentation program performs the segment operation when the similarity value is less than the threshold. After conducting the above processing on five concerts, we totally get 2650 short shots. According to the 6 types of shot scales classifications mentioned earlier, the transfer between shots can be represented by a 6 * 6 matrix. We count the number of transitions between shots and the transition probability is given by equation (7).

$$p(i,j) = \frac{N(i,j)}{\sum_{m=1}^{6} N(i,m)} \tag{7}$$

where $p(i,j)$ refers to the transition probability from shot scale i to j. $N(i,j)$ is the number of transitions from shot scale i to j in the 2650 short shots counted by us.

In Figure 2, the indexes of two dimensions of the transition-probability-matrix are CS, ECS, FS, LS, MS, and OS, and each element in the matrix corresponds to a fixed transition probability.

Among these six shot types, only extreme-close-shot (ECS), close-shot (CS) and medium-shot(MS) contain an appropriate number of instruments. In other shot types, overly complex (e.g. LS) or lacking (e.g. OS) instrument categories will interfere with the accuracy of the Instrument Classification Matching module. In this way, the result of the Instrument Classification Matching module will only be considered when the transition-probability-matrix determines that the next switching shot is ECS, CS or MS. The calculation method for Switching Score is given by equation (8).

$$\zeta_S(i) = \begin{cases} \alpha(i)\Lambda(i) & \gamma_p \in \{ECS, CS, MS\} \\ \Lambda(i) & \gamma_p \in \{FS, LS, OS\} \end{cases} \tag{8}$$

wherein that

$$\alpha(i) = (v_A(i) \wedge v_V(i)) \cdot \rho_A(i)\rho_V(i) \tag{9}$$

$\zeta_S(i)$ refers to the switching score of i-th camera. γ_p represents the next shot to be switched which is predicted by the transition-probability-matrix. When $\gamma_p \in \{ECS, CS, MS\}$, $v_A(i)$ and $v_V(i)$ respectively refer to the instrument prediction of audio and video stream when $\rho(i)$ represents the corresponding confidence score. The AND operator $\wedge$ between them represents that only when the two predictions are the same, the corresponding camera will be considered. $\Lambda(i)$ refers to the switching confidence that calculated from the blur value and shot scale recognition.

3 EXPERIMENTAL SETUP

In order to test the performance of our intelligent concert directing system, we design a subjective evaluation experiment. The experimental materials are sourced from the programs 'Tchaikovsky: Symphony No.5', 'Swan Lake' and 'Vltava' of the 2022 New Year Concert at Communication University of China. Each program has 4 channels of cameras for broadcasting. The experimental materials are altogether 7 videos, which are divided into test set and experiment set. The duration of the test set is 1 minute and 34 seconds, which is used for subjects to familiarize themselves with the experiment's content. The experimental set consists of 6 videos, each of which is between one and a half to two minutes. Among them, 3 clips are generated automatically by our intelligent directing system and the remaining 3 clips are intercepted from the manual broadcasting output PGM signal. Disrupt the sequence of these 6 clips and merge them into a video for the subjects to watch. Before each clip is played, subjects will be prompted by playing the segment number in the video. The subjects are students from Communication University of China, aged 22 to 24, with a total of 16 participants, 8 male and 8 female.

To evaluate the quality of video generated by the directing system, we propose eight evaluation indexes which are: Stability, Smoothness, Scene switching, Camera motion, Suitability, Emotional expression, Thematic and Entirety.

Stability a basic index to evaluate the quality of video. The abnormal jitter of the camera will cause the blur of the video and the blur detection module of the system will eliminate such unusable shots to address this issue. The stability index can be used to evaluate the performance of blur detection module.

Fig. 3. Experimental materials

Smoothness which refers to how smoothly the video plays that good clip should not be frozen or delayed. Smoothness index can be used to evaluate system performance over time scales.

Scene switching evaluate the rationality of the connection and viewing comfort of different shots during switching. This index can effectively detect the performance of shot scale recognition module and the rationality of transition-probability-matrix.

Camera motion evaluate the changes in the image caused by the camera's push, pull and pan in the video.

Suitability detect the consistency of video switching style and music style, such as happy music paired with fast tempo switching, and lyrical music paired with slow tempo switching.

Emotional expression detect the correlation between video style and music emotion.

Thematic evaluate the degree of thematic distinctiveness of a video and whether it can highlight the theme.

Entirety evaluate the visual experience that the overall combination of video style, color, music and sound effects brings to the audience.

During the test, subjects watch the video at a distance of 1 meter from the same electronic devices of the same size and brightness, and rate each video based on their own feelings from the above eight aspects. The scoring criteria are divided into five levels: 1 is bad, 2 is normal, 3 is good, 4 is superior and 5 is perfect. The test video has a total length of 1 minutes and 43 seconds. Before the official start of the experiment, experimenter explains the meaning of evaluation indexes and scoring standards to the subjects. At the beginning of each experimental segment, "Segment 1" to "Segment 6" will appear on the screen in sequence to indicate the serial number of the current segment for the subjects. We keep the venue quiet throughout the formal experiment process to eliminate other factors from interfering.

4 RESULT AND DISCUSSION

We collect raw data by asking subjects to fill out a survey questionnaire, which includes 16 subjects' scores on 7 videos in 8 aspects. When organizing and analyzing data, we exclude data with significant uncertainty and variability, which absolutely has no reference value. Then we group and organize all the data. Firstly, separate the intelligent system directing data and the manual directing data.Then, arrange them according to different programs and evaluation indicators, calculate the average. The resulting data is shown in Table 2. Group the data according to evaluation indicators: Stability, Smoothness, Scene switching and Camera motion as a group; Suitability, Emotional expression, Thematic

and Entirety as another group. We use Matlab to visualize these two sets of data and obtain a histogram of the average value, as shown respectively in Figure 4. The bar chart can visually display the comparison and differences between data, such as the comparison of the average scores of different programs with the same indicator, or the comparison of the average scores of the same program with different switching methods for the same indicator and program.

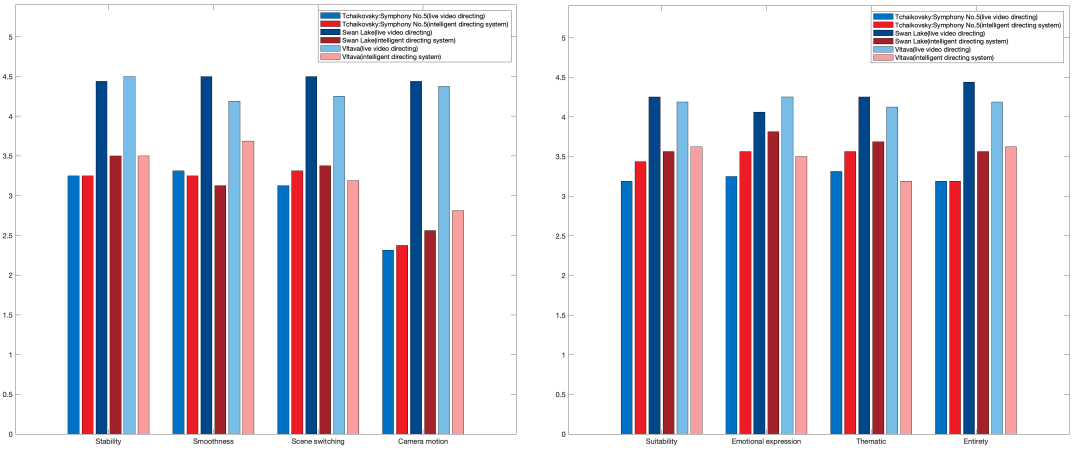

Fig. 4. Results of the group of Stability, Smoothness, Scene switching and Camera motion, Suitability, Emotional expression, Thematic and Entirety

Observing the results of figures and tables, we can find that the difference in scores between intelligent and manual directing is basically within 1 point, indicating that our intelligent directing system can achieve a switching effect similar to that of manual broadcasting. The indicator with the largest difference in scores is camera motion, which is due to the lack of judgment on video camera motion in our model, which is a need for improvement in future models. In addition, there is a significant gap in scene switching index, indicating that there is still a gap between our intelligent guidance and manual guidance. Analyzing the switching confidence, it can be seen that there are currently few concerts used to calculate the transition-probability-matrix, which has certain specificity. In order to better match the smoothness of human viewing, it is necessary to calculate more concert switching methods in order to form a more accurate switching probability matrix.

Table 2. Average score and difference between intelligent and manual switching of 'Tchaikovsky: Symphony No.5', 'Swan Lake', 'Vltava'

Evaluation indexes	Tchaikovsky: Symphony No.5			Swan Lake			Vltava		
	Intelligent	Manual	Diff	Intelligent	Manual	Diff	Intelligent	Manual	Diff
Stability	3.25	3.25	0	3.5	4.4375	0.9375	3.25	3.25	0
Smoothness	3.25	3.3125	0.0625	3.125	4.5	1.375	3.25	3.3125	0.0625
Scene switching	3.3125	3.125	-0.1875	3.375	4.5	1.125	3.3125	3.125	0.1875
Camera motion	2.375	2.3125	-0.0625	2.5625	4.4375	1.875	2.375	2.3125	0.0625
Suitability	3.4375	3.1875	-0.25	3.5625	4.25	0.6875	3.4375	3.1875	-0.25
Emotion expression	3.5625	3.25	-0.3125	3.8125	4.0625	0.25	3.5625	3.25	0.3125
Thematic	3.5625	3.3125	-0.25	3.6875	4.25	0.5625	3.5625	3.3125	-0.25
Entirety	3.1875	3.1875	0	3.5625	4.4375	0.875	3.1875	3.1875	0

5 CONCLUSION

In this paper, we propose an intelligent directing system that can achieve automatic switching of cameras. We consider various factors related to camera switching during the broadcast directing, including switching timing, switching rate and switching camera selection. We use downbeat detection of audio streaming data to predict the switching timing, and obtain the switching rate based on feature extraction and KNN algorithm. For the selection of switching cameras, we propose two large modules: instrument matching and switching confidence. Among them, the shot scale recognition in switching confidence will directly determine whether instrument matching is involved in the final camera selection. In the subjective evaluation experiment, we propose eight evaluation dimensions with a total score of five, to compare our intelligent system with professional manual directing and the score difference between them is basically within one point. The final result shows that our system has great potential, but there are still some modules and algorithms of it that need improvement in the future.

ACKNOWLEDGMENTS

This work was supported by the National Key R&D Program of China (No. 2021YFF0900700), and partly supported by the Fundamental Research Funds for the Central Universities (No.2018CUCTJ085 and No. 3132018XNG1848).

REFERENCES

[1] Ido Arev, Hyun Soo Park, Yaser Sheikh, Jessica K. Hodgins, and Ariel Shamir. 2014. Automatic editing of footage from multiple social cameras. *ACM Transactions on Graphics* (July 2014). https://doi.org/10.1145/2601097.2601198

[2] Judith C. Brown. 1991. Calculation of a constant Q spectral transform. *Journal of the Acoustical Society of America* (Jan. 1991). https://doi.org/10.1121/1.400476

[3] Daniel P.W. Ellis. 2007. Chroma feature analysis and synthesis. http://www.ee.columbia.edu/ dpwe/resources/matlab/chroma-ansyn/.

[4] Chen Ge. 2021. Intelligent Broadcasting Helps Innovate New Media Programs at the 2021 Spring Festival Gala - Analysis of the Application of Artificial Intelligence Switching Technology. *Modern Television Technology* (March 2021), 35–40. https://doi.org/10.3969/j.issn.1671-8658.2021.03.005

[5] Peter Grosche and Meinard Müller. 2011. Extracting Predominant Local Pulse Information From Music Recordings. *IEEE Transactions on Audio, Speech, and Language Processing* 19, 6 (Aug. 2011), 1688–1701. https://doi.org/10.1109/TASL.2010.2096216

[6] Mojtaba Heydari, Frank Cwitkowitz, and Zhiyao Duan. 2021. BeatNet: CRNN and Particle Filtering for Online Joint Beat Downbeat and Meter Tracking. *CERN European Organization for Nuclear Research - Zenodo* (Aug. 2021). https://doi.org/10.5281/zenodo.7036495

[7] Mackenzie Leake, Abe Davis, Anh Truong, and Maneesh Agrawala. 2017. Computational Video Editing for Dialogue-driven Scenes. *ACM Transactions on Graphics* (July 2017). https://doi.org/10.1145/3072959.3073653

[8] K. L. Bhanu Moorthy, Moneish Kumar, Ramanathan Subramanian, and Vineet Gandhi. 2020. GAZED– Gaze-Guided Cinematic Editing of Wide-Angle Monocular Video Recordings. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, 1–11. https://doi.org/10.1145/3313831.3376544

[9] Meinard Müller. 2015. Fundamentals of Music Processing. *Springer International Publishing eBooks* (Jan. 2015). https://doi.org/10.1007/978-3-319-21945-5

[10] Meinard Müller and Sebastian Ewert. 2011. Chroma Toolbox: Matlab Implementations for Extracting Variants of Chroma-Based Audio Features. In *International Society for Music Information Retrieval Conference*. https://doi.org/10.5281/zenodo.1416032

[11] Yingwei Pan, Yue Chen, Qian Bao, Ning Zhang, Ting Yao, Jingen Liu, and Tao Mei. 2021. Smart Director: An Event-Driven Directing System for Live Broadcasting. *ACM Trans. Multimedia Comput. Commun. Appl.* 17, 4 (Nov. 2021), 18 pages. https://doi.org/10.1145/3448981

[12] Karthikeya Racharla, Vineet Kumar, Chaudhari Bhushan Jayant, Ankit Khairkar, and Paturu Harish. 2020. Predominant Musical Instrument Classification based on Spectral Features. In *2020 7th International Conference on Signal Processing and Integrated Networks (SPIN)*. IEEE, 617–622. https://doi.org/10.1109/SPIN48934.2020.9071125

[13] Aida Ualibekova and Pakizar Shamoi. 2022. Music Emotion Recognition Using K-Nearest Neighbors Algorithm. In *2022 International Conference on Smart Information Systems and Technologies (SIST)*. IEEE, 1–6. https://doi.org/10.1109/SIST54437.2022.9945814

[14] Jose R. Zapata, Matthew E. P. Davies, and Emilia Gómez. 2014. Multi-Feature Beat Tracking. *IEEE/ACM transactions on audio, speech, and language processing* 22, 4 (April 2014), 816–825. https://doi.org/10.1109/TASLP.2014.2305252

Received 7 April 2023; revised May 2023; accepted 9 May 2023