Advances in Computer, Signals and Systems (2022)
Clausius Scientific Press, Canada

DOI: 10.23977/acss.2022.060505
ISSN 2371-8838 Vol. 6 Num. 5

Research on Design and Implementation of Intelligent Guide System

Fang He

Communication University of China, Beijing, 100024, China

Keywords: Artificial intelligence, Guide, Character recognition

Abstract: At present, the application of artificial intelligence technology is developing continuously, and the application fields are more and more. The application of intelligent switching technology based on image recognition also appears in TV program production. How to shorten the production cycle of recorded and broadcast programs and enrich the program effect of live programs has become a key issue in program production today, and intelligent guidance system has emerged as the times require. It is of great research significance and application value that how to use different algorithm technologies to solve the different requirements of program production in different scenes and give switching suggestions that meet the aesthetic requirements of human beings.

1. Introduction

In the production process of a TV program, the director needs to schedule the whole scene and select and switch the camera channel. They are required to choose one of the many camera signals closely related to the program content planning, and the relatively wonderful signals are presented to the audience. Using traditional production methods, it is necessary for the director to guide the shooting of the program in the whole process, and at the same time, to quickly select from multiple signals with diverse perspectives, which has a high demand on the personal ability of the director. In particular, live programs usually last for a long time, and the director needs to maintain high concentration throughout the whole process, bear great work pressure, and also bring challenges to the wonderful presentation of programs.

At present, intelligent guided broadcasting products are scarce in the market, and the products that have come into being are basically customized products. For example, the guided broadcasting switching function included in the integrated media intelligent live broadcasting production built by the Central Radio and TV Station, as well as intelligent guided broadcasting products in some specific scenarios, such as alpine skiing scenes and military parade scenes at the Winter Olympics. The intelligent guidance system designed in this paper includes but is not limited to these specific scenes. A general guidance tool will be introduced to enable it to identify shot groups in various scenes and adapt to shooting shots of various qualities. Regardless of the level of photographers, this system can classify and process these shots and give optimal switching suggestions to help users produce programs.

2. Research status at home and abroad

At present, some large domestic TV stations have started to try to integrate intelligent technology into the guidance system and put it into practical application. For example, China Central Radio and Television Corporation used intelligent switching technology in the 2021 Spring Festival Gala New Media Multi machine Shooting Program [1]. The identification of information mainly includes four aspects: voice recognition, lip reading, face recognition and human key point detection. In order to give the time for equipment and software to extract features and give the position of switching points, there is a time difference of about 25 seconds between live program recording and actual broadcast. During the 25 seconds, the intelligent algorithm will use design logic to give switching strategies.

In foreign countries, there is no complete intelligent guidance system. MIT has released an artificial intelligence system called PicelPlayer, which uses deep learning to identify the sound of musical instruments and understand the corresponding relationship between sound and musical instruments in the screen. Swapnil Vitthal Tate et al. proposed a face tracking framework, which can use Haar features for face detection and Gabor feature extraction for recognition, and can perceive the changes in the pose and direction of the person in the video.

The design purpose of this system is to mine the relationship between the change of scene and the change of performance content through data analysis, in order to achieve the following purposes: 1. real-time recognition of performance content. 2. Based on the analysis of the existing audio and video data, propose a shot switch proposal to reduce the pressure of the guide work.

3. Key technology

3.1. Overview of technical route

The system covers three layers: algorithm layer, technology layer and application layer. The application layer is mainly oriented to users, providing users with visual operation interface for live broadcast guidance; The technology layer and algorithm layer provide technical support for the application layer. The core technology of the technology layer is the recognition and classification of the real-time state of the scene, which facilitates the technicians to update and maintain the products in the later period, including the optimization training of the algorithm model and the update iteration of the subsequent products.

3.2. Technical module analysis

The technology layer can be divided into three modules, namely, video and audio analysis module, shot analysis module and guidance strategy module. First, the shot analysis module recognizes and classifies the input video signal, while the video and audio module detects the emotion and rhythm of the input audio signal, and performs basic object and face recognition on the video signal, providing effective information for subsequent shot switching. Finally, the video and audio signals are sent to the guidance strategy module, and the video screen and video switching output is carried out using the algorithm model successfully trained in advance. The three modules are described in detail below:

(1) Video signal analysis module

This module will use image processing related algorithms such as target detection to analyze the input multi-channel video signals, so as to extract the characteristics of each signal corresponding to the subject, and classify the video signals. It mainly uses lip reading technology to judge speakers [2], and then compares and confirms the identity through the face database established in advance [3]. The position of the human body is detected by using the depth neural network Single Shot MultiBox Detector algorithm [4], and then the key points of the human body are predicted by using the depth

neural network based on the thermograph to determine the human body movements [5].

(2) Audio signal analysis module

Similar to the analysis of video signal, the background music of this module will use the main instrument recognition algorithm and music emotion recognition algorithm to analyze the input audio stream, so as to extract the feature information of each small piece of music and label the audio signal. Audio emotion analysis mainly includes K-Nearest Neighbor (KNN), Support Vector Machine (SVM) and Bayesian algorithms. Some machine learning methods such as artificial neural network, regression analysis and self-organizing mapping are also widely used in this field. Among these algorithms, KNN and SVM are more common, while SVM is better than KNN in processing high-dimensional data, which can effectively avoid over fitting, and has a better effect in music emotion classification.

The voice of the performer is judged by identifying the audio stream or voiceprint recognition. Acoustic features are a set of acoustic description parameters extracted from sound signals by computer through algorithms, including template matching, Gaussian mixture model, joint factor analysis, depth neural network method, etc. However, the deep neural network method is a data driven method. If there are a large number of audio records and audio files to train in the early stage, it can achieve the desired good effect.

(3) Lens analysis module

The lens is the smallest unit of the camera to capture visual information of a specific action or time at a time. For the same image content, different types of lenses will bring different feelings to the audience, affecting the information transmitted by the visual elements in the screen. In this module, it is necessary to classify the video signal from the signal source, so as to cooperate with other modules to output the optimal shot. This system will combine the image analysis method and depth learning algorithm to first determine whether the shot is a fixed shot. If it is a fixed shot, it will analyze whether the shot is a close shot, middle shot or a long shot; If it is a moving lens, it is also necessary to analyze the lens movement and judge the push, pull and swing of the lens while analyzing the scene. In addition, it may be necessary to judge the transformed shot, and the shot can be segmented if conditions permit. At present, the commonly used shot analysis algorithms include naive Bayesian classifier, convolutional network model TSN and I3D, and deep network model SGNet for shot type classification. The system will partially improve the above algorithms and apply them.

(4) Guide strategy module

This module will use machine learning algorithm to combine the results of the first two modules, learn the relationship between the rhythm emotion of the performance and the time point, speed and switching mode of screen switching, establish a set of performance guidance switching strategies, make screen switching select the cut in point and cut out switching according to the performance content, rhythm, performance atmosphere, etc., and ensure that the re editing of the performance content can be completed through switching control, It also ensures the high quality of editing, which can make the audience feel the double aesthetic feeling of audio-visual, and ultimately achieve a high degree of unity of sound and picture.

4. Research on key issues

4.1. Understanding of sound and camera content

The recognition algorithm of this system will be based on the existing recognition algorithm and improved. For audio content, the current recognition algorithms are mainly the theme recognition algorithm and the depth neural network. The former can extract the main melody track of music, and the latter can judge the identity of the speaker. We will use millions of audio records and audio files as the training samples of the recognition algorithm, or input multiple types of features into the classifier for recognition, and finally give voice judgment. For music emotion, there are many excellent

algorithms that can be used in the task of intelligent guidance at present, some of which are common but inefficient, and some algorithms are better at processing high-dimensional data. The author intends to compare the results of different algorithms and choose the one that best fits the task of intelligent guidance to apply in this system. For shot content recognition, plan the AI algorithm is used to intelligently analyze the screen information of each stand, and intelligently select the stand signal that can be displayed after combining with the sound.

4.2. Guide switching logic

The effective recognition of the emotion and rhythm of sound by AI can make the camera language richer, the emotion more full, and the switching point of the camera more accurate. The switching rhythm of the video screen should change with the rhythm of the performance scene. The switching rhythm is slow when it is slow, and can be used to express emotions by overlapping. The switching rhythm should also be accelerated when it is tense.

When rebroadcasting performance programs, the coordination of performance pictures and sound content is also very important. Each performance will have a certain theme level change in the overall performance. In the process, the guide should be consistent in sound and painting. For this reason, the author will use the main voice channel recognition technology to identify the main speakers in the current performance, and use the recognition technology to quickly find the current actor's position from multiple video signals, so as to assist the guide in accurate screen switching.

5. Conclusion

The intelligent guidance system can give guidance opinions in a short time, even realize one click guidance, and support real-time operation preview. Intelligent guidance can play an important role in both recorded and live programs. At present, some domestic large-scale program sites also try to display and store the intelligent guidance switching signal as a sub cut, so as to assist the on-site guidance personnel in lens switching.

The intelligent guiding system designed this time can not only realize the basic functions of common guiding stations, including previewing live broadcast pictures, supporting multi picture switching, and switching required video signals at any time. It is also planned to use artificial intelligence to achieve audio and video analysis and picture recognition by collecting a large number of video sequences of multi-channel broadcast scenes, using audio and video feature extraction, character behavior recognition and other algorithms for training, Character recognition, automatic shot segmentation and switching and other functions are used to preprocess the live broadcast images and provide users with certain shot switching suggestions to ensure the smoothness and accuracy of the subsequent live broadcast effects.

The future intelligent guidance system can also integrate 5G, cloud computing, artificial intelligence and other new generation information technologies to provide more convenient, lighter and lower cost services for audio-visual applications in various vertical fields, and create a new audio-visual service format with higher audio-visual quality, better immersive experience and more personalized needs.

References

[1] Chen Ge. Intelligent Guide Helps New Media Program Innovation in 2021 Spring Festival Gala--Analysis of the Application of Artificial Intelligence Switching Technology [J]. Modern TV Technology, 2021 (03): 35-40

[2] Swapnil Vitthal Tathe and Abhilasha Sandipan Narote and Sandipan Pralhad Narote. Face Recognition and Tracking in Videos [J]. Advances in Science, Technology and Engineering Systems, 2017, 2(3): 1238-1244.

[3] Xu Y, Yan W, Yang G, et al. CenterFace: JointFace Detection and Alignment Using Face as Point [J]. Scientific

Programming, 2020, 2020: 1-8.
[4] Liu W, Anguelov D, Erhan D, et al. SSD: Single ShotMultiBox Detector [J]. 2015.
[5] Su Z, Ye M, Zhang G, et al. Cascade FeatureAggregation for Human Pose Estimaion [J]. 2019.