Design of car audio and video synchronization algorithm based on neural network model

Bin Li[a,b,*], Haipeng Cao[a,b,1]
[a]China Automotive Technology and Research Center Co., Ltd.; [b]China Auto Information Technology (Tianjin) Co., Ltd., Tianjing, China.

ABSTRACT

The transmission process of automobile monitoring video network will be disturbed by complex network environment, such as network delay, jitter and other factors, which will lead to the phenomenon of multimedia asynchronization, and audio-video synchronization, as one of the key technologies, has attracted more and more attention. In this paper, an audio-video synchronization algorithm of automobile monitoring video based on neural network model is proposed. By improving the variable-length coding algorithm and its mapping rules, the audio information is dynamically grouped, and the dynamic double mapping relationship between prediction mode group and variable-length code group is established. The simulation results show that the algorithm can ensure the correctness and integrity of audio data, and realize the synchronous coding transmission of high-definition and ultra-high-definition video. The algorithm is applied to vehicle monitoring video, which ensures the time constraint relationship between audio data and video data, that is, realizes the synchronous playback of audio and video, thus effectively ensuring the driver's driving safety, avoiding traffic accidents and making it easier for the driver to drive the vehicle.

Keywords: Audio and video synchronization, Automobile monitoring, Neural network

1. INTRODUCTION

In recent twenty years, with the rapid development of electrical and electronic technology and software industry, traditional industries are also facing revolutionary changes. In the automotive industry, the application of automotive electronic components and related new technologies has brought unprecedented challenges and opportunities to the automotive industry[1]. The flexibility and maneuverability of the car make it an indispensable and important means of transportation in people's daily work and life [2]. Image and video information not only fills people's work and life from the media, but also has extremely important value in military and scientific research. Audio-video synchronization can be divided into two aspects, one is intra-media synchronization and the other is inter-media synchronization. The former is a necessary condition for the latter, that is to say, it is possible to realize the synchronization between audio and video media only if the synchronization between audio and video media is realized first, and it is meaningful to design the synchronization control between audio and video media. In the application of smart car, the quality of audio and video synchronization control directly affects the playback quality of smart car files, and users can directly feel the synchronization effect of smart car [3]. Therefore, the quality of audio and video synchronization plays a vital role in the quality of the whole application. It is a very practical research direction to apply audio and video synchronization to intelligent vehicles to solve the problems caused by blind observation areas, ensure driving safety, avoid traffic accidents and make drivers easier to drive vehicles [4].

*libin2018@catarc.ac.cn; [1]caohaipeng@catarc.ac.cn

Seventh International Conference on Mechatronics and Intelligent Robotics (ICMIR 2023), edited by Srikanta Patnaik,
Tao Shen, Proc. of SPIE Vol. 12779, 127791E · © 2023 SPIE · 0277-786X · Published under a
Creative Commons Attribution CC-BY 3.0 License · doi: 10.1117/12.2688736

The synchronization of audio data and video data includes two parts, one is internal synchronization of audio and video media, and the other is synchronization between audio and video media. Synchronization channel needs high real-time and priority, and the complexity of synchronization information and control information is different with different synchronization granularity [5]. As the granularity becomes finer and finer, there will be more synchronization information and more frequent control operations. Because the network environment is very complex and constantly changing, it is very common to have congestion, which also leads to the loss of synchronization information at any time [6]. In the process of motion estimation, Choudhury et al. established the corresponding matching relationship by modulating the parity of the optimal motion search point with 1/4 pixel accuracy, which reduced the impact on video quality, but at the same time caused distortion drift between frames [7]. This method can keep the code rate stable, but the video quality will still decrease due to error accumulation, which will lead to the incorrect extraction of audio information. In order to avoid a great impact on video quality, Sarkar et al. proposed an audio embedding algorithm based on inter-frame prediction mode, but the embedding data capacity of this algorithm is small, with only 2 bits per macroblock [8]. Civera et al. put forward a method of embedding audio coded data in variable size blocks, which embeds 2.56bit data in each macro block on average, which improves the embedding capacity and ensures the accuracy of audio data [9]. In this paper, an audio-video synchronization algorithm of automobile monitoring video based on neural network model is proposed, so as to ensure the driver's driving safety, avoid traffic accidents and make it easier for the driver to drive the vehicle.

2. METHODOLOGY

2.1 Audio and video synchronization technology

In order to make the video and sound collected by the monitoring system truly restore the monitored environment, in addition to improving the coding accuracy and optimizing the algorithm, due to the great difference in information redundancy between audio and video, the difference in coding rate and transmission bandwidth often makes them unsynchronized at the user terminal. The audio and video acquisition module converts the original audio and video analog signals into digital signals through sampling and quantization, and prepares for the next audio and video coding [10]. The amount of uncoded audio and video data is very large, so it is unrealistic to directly transmit such a large amount of information under the existing network environment. Because the compressed audio and video signals can greatly reduce the transmission bandwidth, but at the same time, it will lead to the decline of signal quality, the increase of algorithm complexity and a lot of control overhead, and the effectiveness of coding will decline. Therefore, it is necessary to deal with audio and video coding properly, and the efficiency and quality need to be considered comprehensively.

There are a lot of redundant information in voice and video data, especially video data. After compression, the data volume will drop greatly, which makes it possible to transmit real-time multimedia data through the network. In the audio-video synchronization technology, the information flow of the vehicle monitoring system is mainly transmitted in the communication network system with the help of multimedia information technology. In the actual transmission process, there are time delay jitter experienced by adjacent media units in a single media stream and time difference between related media units of audio and video, that is, offset. There are two kinds of offset: intra-media offset and inter-media offset, and there is a corresponding allowable range of offset. For audio or video, when the delay jitter is less than 0.01s, the audio and video playback is in a synchronous state, otherwise it is out of step. The whole audio-visual intelligent vehicle driving monitoring system consists of three modules: front-end media server, data management forwarding server and customer receiving end. The design scheme of the whole system is shown in Figure 1.

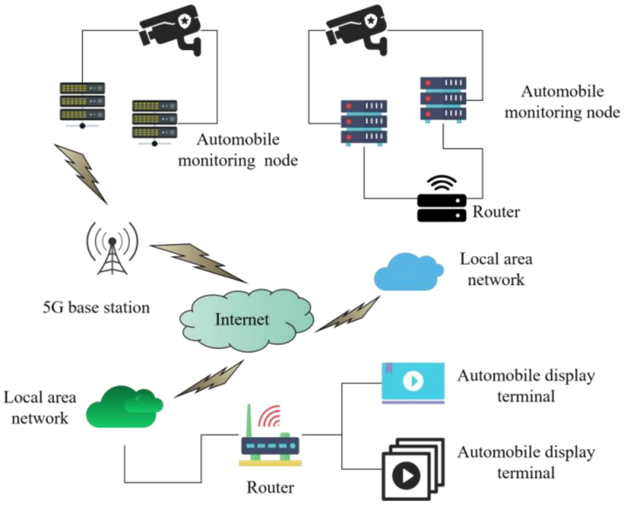
Automobile monitoring node
Automobile monitoring node
Router
5G base station
Local area network
Internet
Local area network
Automobile display terminal
Automobile display terminal
Router

Figure 1. Audio and video traffic monitoring system architecture

The front-end media server is responsible for the collection, compression and transmission of audio and video data streams. Its audio source can be a microphone, and the video data source can be a ball machine or a gun machine commonly used in the market, which is compatible with the traditional monitoring front-end equipment to the maximum extent. Due to the difference of communication network topology and hardware resource distribution, it may lead to delay, which will lead to the mismatch of audio and video streams in reception. It is necessary to optimize the synchronization in audio and video technology, and adjust the time relationship between units in hardware media through software algorithms to eliminate the time offset between various media. The time offset of related media units between different media streams is also called deviation. No matter how human beings progress, their subjective perception ability can't be compared with computer quantization accuracy, so as long as the deviation is controlled within a reasonable range, audio and video media streams can be considered synchronous.

2.2 Audio and video synchronization and feature fusion algorithm

The process of audio acquisition, coding and transmission is separated from that of video [11]. That is to say, two data streams, one is audio data stream and the other is video data stream, which are independent of each other and do not affect each other. According to the ratio of the output value corresponding to the correct node to the total output value under each sample, the probability that the samples are classified into the correct category is calculated, and the fitness function is positively correlated with this probability. Therefore, for the same classification result, the improved fitness function can give different classification costs to each classification result according to the probability that each sample is correctly classified. The feature extraction model of automobile surveillance video based on improved CNN algorithm is shown in Figure 2.

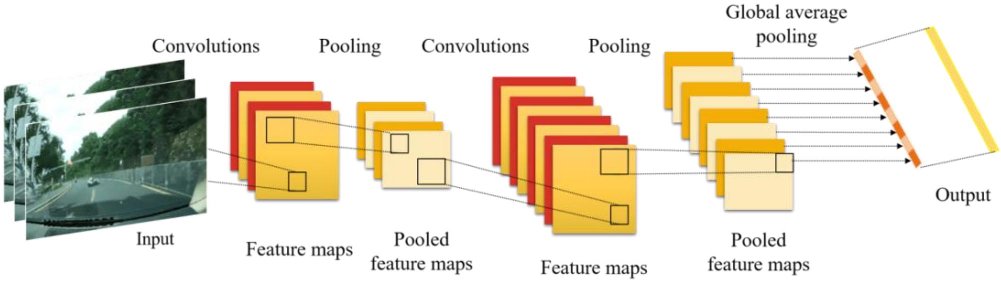

Figure 2. Feature extraction model of automobile monitoring video images

The voice media at the receiving end adopts the synchronization method based on playing time limit, that is, the voice media unit in the cache of the sink end is used to compensate the network delay jitter. The specific method is to set up several cache units and prefetch thousands of speech frames, so as to ensure the continuity of speech playback through the data in the cache when the network transmission is jittery. In order to ensure the continuity of automobile monitoring video playback, the target node must prefetch at least B_p multimedia objects, and their values can be calculated by the following formula:

$$B_p = \left[D_{\max} - D_{\min}\bar{\theta}\right] \tag{1}$$

The target node only needs to set the cache units of B_c multimedia objects at most, and its value can be calculated by the following formula:

$$B_c = \left[\frac{2 \times (D_{\max} - D_{\min})}{\theta}\right] \tag{2}$$

$D_{\min}$ represents the minimum delay time of transmitting a multimedia object on the network; $D_{\max}$ represents the maximum delay time of transmitting a multimedia object on the network; θ stands for the broadcast period of multimedia objects, that is, a synchronization time unit. The speech delay jitter control algorithm given in this paper is to set the receiving end speech playback buffer with the size of B_c speech frames according to the above formula, and start audio playback when there are B_p speech frames in the audio playback buffer.

Collect audio data and extract synchronization information at the same time. At this time, the amount of audio data is very large, which is not suitable for direct transmission. Because there is a lot of redundant information in audio data at this time, it provides the possibility of information compression. Next, it is sent to the audio coding module for coding. After coding, the amount of data is greatly reduced. At this time, it can be transmitted through the network. Send these data to the audio sending module for sending. The video data stored in the computer is a set of two-dimensional discrete matrices, and so is the local extreme point found by Gaussian pyramid. It is necessary to construct a Gaussian difference DOG space for fitting. In the process of processing, some edge points or points with relatively low contrast need to be removed, so as to improve the stability of feature point detection. Firstly, Taylor formula is used to expand the function $D(x, y, \sigma)$ in the scale space, and the extreme point is located at the origin through displacement:

$$D(X) = D + \frac{\partial D^T}{\partial X} X + \frac{1}{2} X^T \frac{\partial^2 D}{\partial X^2} X \tag{3}$$

By solving the derivative of $D(X)$ with respect to X and making the equation zero, the position of the extreme point $\hat{X}$ can be obtained:

$$\hat{X} = \frac{\partial^2 D^{-1} \partial D}{-\partial X^2 \partial X} \tag{4}$$

And then get:

$$D(X) = D + \frac{1}{2} \frac{\partial D^T}{\partial X} X \tag{5}$$

At this time, the offset relative to the interpolation center can be found. If the offset is greater than 0.5 in any dimension, it indicates that the interpolation center has shifted to another nearby point near the feature point, so it is necessary to select a new feature point position. Then the interpolation is repeated at the new feature point position until convergence.

3. RESULT ANALYSIS AND DISCUSSION

The main purpose of adopting audio and video synchronous streaming technology is to ensure the minimum delay, ensure the quality of audio and video and achieve the synchronization effect when the car video terminal plays audio and video data. In the synchronization technology of audio and video transmission, audio and video data can be transmitted in real time. The audio acquisition thread will be responsible for collecting audio data and writing the collected data into the shared memory. The requirements of the acquisition module are high, and it will determine whether the audio data is missing or delayed from the source. Because the audio data stream is continuous, any data loss and delay in the middle will cause unbearable phenomenon for users in the car video terminal. To solve the problem of car audio and video synchronization, we must first solve the synchronization within the media, that is, to ensure that the audio and video data can be transmitted to the car video terminal in a complete and real-time manner without mutual interference. In terms of integrity, it is necessary to ensure that data is not lost in the process of collection, compression, transmission and playback. In terms of real-time, audio and video data should be sent out within the maximum allowable delay, and decompressed and played in the car video terminal in time. After the video feature passes through the classifier, the feature vector is mapped into a classification score vector, that is, the probability distribution vector of the sample corresponding to the video feature in the whole motion category, and the index position corresponding to the maximum probability value is taken as the classification result of the model. The classification accuracy of different algorithms is shown in Figure 3.

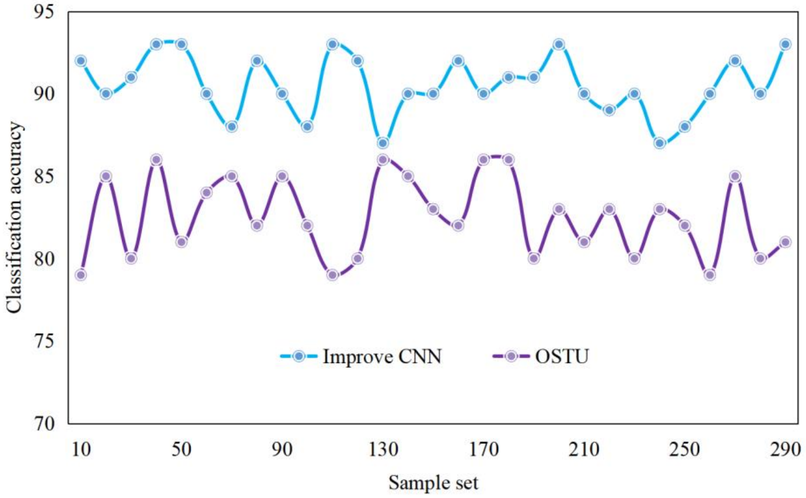

Figure 3. Classification accuracy of different algorithms

In the synchronization control algorithm, based on the normal audio playback time, audio is used as the master media and video is used as the slave media, and the synchronization control between media is realized by adjusting the video playback process. The main purpose of audio and video feature matching is to integrate the feature information obtained from different CNN, and then remove the redundant part, and use the integrated feature information for subsequent operation or analysis. The measurement method based on deep learning can measure the semantic similarity of high-level feature vectors, which makes it possible to solve the problems of intra-class differences and inter-class overlap of actions. On the platform of Matlab, the efficiency of different methods of vehicle monitoring video audio-video fusion model is tested, and the recognition efficiency is evaluated by running time. The statistics of calculation time experimental results of different feature dimensionality reduction are shown in Table 1.

Table 1. Dimension reduction time of audio-video fusion model of automobile monitoring video

Video scene	Training sample		Test sample	
	OSTU	Improve CNN	OSTU	Improve CNN
Daytime driving scene	7.68	6.23	7.24	5.65
Daytime parking scene	6.45	5.61	6.76	4.66
Night driving scene	9.68	6.55	7.88	5.99
Night parking scene	8.75	5.38	8.81	4.31

In the real-time video monitoring system, the transmission of audio data and video data must consider both real-time and transmission quality. The audio decoding module takes out the audio data from the audio buffer pool, and the synchronization control module judges whether the audio data is synchronized with the video data being played. If it is synchronized, it decodes and plays it, otherwise it adopts other processing methods. After sending out the information data, the system will carry out the signal acquisition and coding process. In this process, there may be asynchronous phenomenon between video and audio data. Once this situation occurs, it will increase the cache pressure of the car video terminal and reduce the portability of the car video terminal software. Therefore, it is necessary to adopt appropriate algorithms to avoid this situation as much as possible. In the case of a large number of unknown input data features, it is not appropriate to specify a unified static structure for different types of action sequence data with different lengths. A lot of attempts must be made to obtain a suitable network structure in order to obtain satisfactory analysis results. Figure 4 shows MAE of audio and video feature fusion of different algorithms. Figure 5 shows the time of audio and video feature fusion of different algorithms.

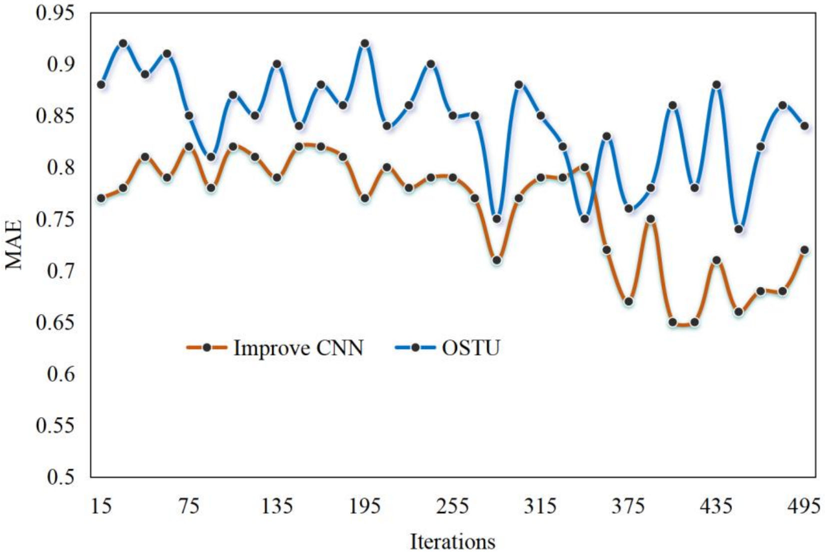

Figure 4. MAE of different algorithms

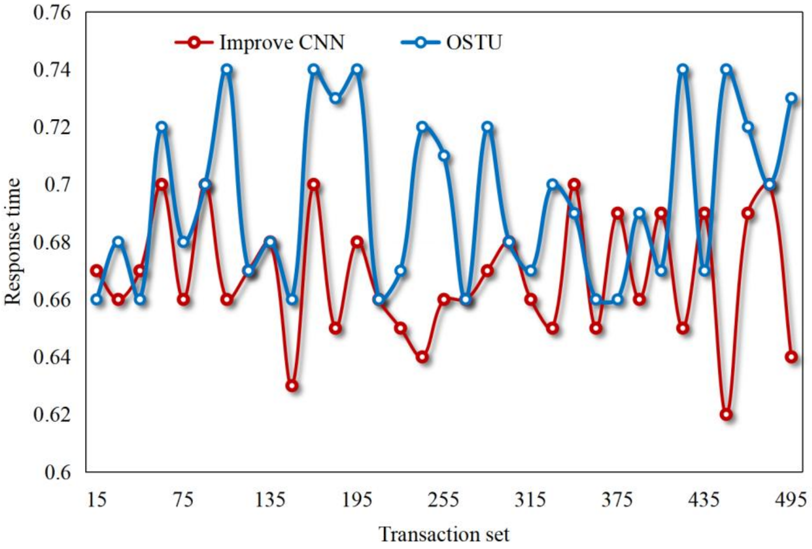

Figure 5. Response time of different algorithms

It can be seen that the error of the improved CNN on the test set is obviously improved compared with the comparative algorithm, and the overall reduction is about 20%, and the response time of the algorithm is obviously reduced. An adaptive operator is added to improve CNN's local search strategy, so that the local search range decreases with the iterative algorithm, and then the local search is more targeted. In the design process, it is necessary to make specific statistics on the size of audio data packets to ensure that audio can be played normally in normal decompression and playback, to ensure the highest transmission efficiency, and to minimize the packet loss rate.

4. CONCLUSION

In the traffic monitoring system, audio and video synchronization is a key problem. Affected by network delay and congestion, the audio and video information received at the driver's end often appears asynchronous, which will greatly reduce the user's experience. In this paper, an audio-video synchronization algorithm of automobile monitoring video based on neural network model is proposed, so as to ensure the driver's driving safety, avoid traffic accidents and make it easier for the driver to drive the vehicle. The results show that the error of the improved CNN on the test set is obviously improved compared with the comparative algorithm, and the overall reduction is about 20%, and the response time of the algorithm is obviously reduced. In the application of smart car, the quality of audio and video synchronization control directly affects the playback quality of smart car files, and users can directly feel the synchronization effect of smart car. Therefore, the quality of audio and video synchronization plays a vital role in the quality of the whole application. By analyzing the application of audio-video synchronous stream in multimedia information technology, the system design is combined with the needs of customers to ensure real-time monitoring of audio-video information under low broadband conditions. After compressing audio and video, the cost is saved and the operation efficiency is improved. From the sensory point of view, the decoded image after embedding data is not much different from the original decoded image, and there is no obvious distortion problem.

REFERENCES

[1] Hao, C., Ping, S., Cheng, Y., et al., Testing a Firefly-Inspired Synchronization Algorithm in a Complex Wireless Sensor Network. Sensors, 17(3), pp. 544 (2017).

[2] Lee, G., Ko, H., Pack, S., An Efficient Delta Synchronization Algorithm for Mobile Cloud Storage Applications. IEEE Transactions on Services Computing, 10(99), pp. 341-351 (2017).

[3] Ioana, A., Korodi, A., Improving OPC UA Publish-Subscribe Mechanism over UDP with Synchronization Algorithm and Multithreading Broker Application. Sensors, 20 (19), pp.5591 (2020).

[4] Ahmad, R., Zubair, S., Alquhayz, H., Ditta, A., Multimodal speaker diarization using a pre-trained audio-visual synchronization model. Sensors, 19(23), pp.5163 (2019).

[5] Xu, M., Borji, A., Zhu, C., et al., Introduction to the Issue on Perception-Driven 360° Video Processing. IEEE Journal of Selected Topics in Signal Processing, 14(1), pp. 2-4 (2020).

[6] Mansouri, N., Watelain, E., Jemaa, Y. B., et al., Video-processing-based system for automated pedestrian data collection and analysis when crossing the street. Journal of Electronic Imaging, 27(2), pp. 023016-023016 (2018).

[7] Choudhury, H. A., Sinha, N., Saikia, M., Correlation Based Rood Pattern Search , no. CBRPS) for Motion Estimation in Video Processing. Journal of Intelligent and Fuzzy Systems, 36(12), pp. 1-11 (2019).

[8] Sarkar, S., Bhairannawar, S. S., KB, R., FPGACam: A FPGA based efficient camera interfacing architecture for real time video processing. IET Circuits, Devices & Systems, 15(8), 814-829 (2021).

[9] Civera, M., Fragonara, L. Z., Surace, C., Using Video Processing for the Full-Field Identification of Backbone Curves in Case of Large Vibrations. Sensors, 19(10), pp. 2345 (2019).

[10] Ikawa, S., Takada, N., et al., Real-time color holographic video reconstruction using multiple-graphics processing unit cluster acceleration and three spatial light modulators. Chinese Optics Letters, 18(01), pp. 23-27 (2020).

[11] Baptista Rios, M., López-Sastre, R. J., Acevedo-Rodríguez, F. J., Martín-Martín, P., & Maldonado-Bascón, S., Unsupervised action proposals using support vector classifiers for online video processing. Sensors, 20(10), pp.2953 (2020).