Wireless Networks (2024) 30:2839–2854
https://doi.org/10.1007/s11276-024-03693-6

ORIGINAL PAPER

Longitudinal tear detection method for conveyor belt based on multi-mode fusion

Yimin Wang[1] · Yuhong Du[1] · Changyun Miao[2] · Di Miao[3] · Yao Zheng[1] · Dengjie Yang[2]

Accepted: 2 February 2024 / Published online: 13 March 2024

Abstract

The longitudinal tear of conveyor belts is the most common accident occurring at the workplace. Given the limitations on accuracy and stability of current single-modal approaches to detecting the longitudinal tear of conveyor belts, a solution is proposed in this paper through Audio-Visual Fusion. According to this method, a linear CCD camera is used to capture the images of the conveyor belt and a microphone array for the acquisition of sound signals from the operating belt conveyor. Then, the visual data is inputted into an improved Shufflenet_V2 network for classification, while the preprocessed sound signals are subjected to feature extraction and classification using a CNN-LSTM network. Finally, decision fusion is performed in line with Dempster-Shafer theory for image and sound classification. Experimental results show that the method proposed in this paper achieves an accuracy of 97% in tear detection, which is 1.2% and 2.8% higher compared to using images or sound alone, respectively. Apparently, the method proposed in this paper is effective in enhancing the performance of the existing detection methods.

Keywords Belt conveyor · Longitudinal tear detection · Shufflenet_V2 · CNN-LSTM · Decision fusion

1 Introduction

Belt conveyors are the most used transportation equipment in various places such as mines, power plants, and ports [1]. Therefore, the safe and stable operation of belt conveyors is essential in these sectors. As the core component and most expensive part of a belt conveyor [2], conveyor belt is most prone to failure. According to industrial statistics, a coal company suffered 42 conveyor belt-related incidents over a ten-year period, of which approximately 80% involved longitudinal tears [3]. Therefore, it is necessary to prevent safety incidents through a research on the effective methods of longitudinal conveyor belt tear detection.

The traditional methods used to detect longitudinal tearing include mechanical detection [4, 5], ultrasonic detection [6], and electromagnetic detection [7, 8]. However, these methods are disadvantaged by low accuracy and high structural complexity. Currently, the most popular choice is image-based detection. In as early as 2014, Yang et al. [9] proposed the use of a linear CCD camera to capture the images of conveyor belt for tear feature recognition, which achieves online monitoring. In 2016, Li et al. [10] improved the accuracy of conveyor belt tear detection by applying the SSR algorithm. In 2017, Qiao et al. [11] proposed a binocular vision detection method based on the fusion of infrared and visible light. In 2018, Li et al. [12]. introduced laser to assist the detection of conveyor belt tearing through the continuity of laser, thereby enhancing the adaptability of the detection system to harsh working environments. In 2020, Yang et al. [13] combined ordinary images with infrared images for the recognition of conveyor belt features, which also improved the adaptability of this system to external environments and thus its precision. In 2021, Zhang et al. [14] used an improved YoloV4 model to train and test the images of conveyor belt, enhancing the real-time

✉ Yuhong Du
1920051101@tiangong.edu.cn

✉ Changyun Miao
miaochangyun@tiangong.edu.cn

[1] School of Mechanical Engineering, Tiangong University, Tianjin 300387, People’s Republic of China

[2] School of Electronics and Information Engineering, Tiangong University, Tianjin 300387, People’s Republic of China

[3] School of Electronic Engineering, Tianjin University of Technology and Education, Tianjin 300000, People’s Republic of China

performance and accuracy of the detection system. However, due to the harsh working environment in which conveyor belts operate, the images of belt surface are prone to external disturbances, which has adverse effects on the accuracy of image detection methods. While sound signals can be used to judge the status of the system in operation. In 2022, Miao et al. [15] adopted the MFCC method to extract features from sound signals for the detection of tearing. However, the detection of conveyor belt tears through sound signal recognition is suitable only for ongoing tear, which leads to certain limitations.

The motivation behind this paper is to address the existing issues associated with the detection of longitudinal tears in belt conveyors. Traditional methods of detection, ranging from mechanical techniques to those based on imagery, have been demonstrated to be inadequate, given their low accuracy or vulnerability to external disturbances in the challenging environments where these belts operate. Moreover, while sound signals offer a comprehensive insight into the system's operational state, they exhibit limitations in detecting longitudinal tears. Recent advancements in audio-visual fusion techniques in other domains have showcased their potential in enhancing detection accuracy. Considering these factors, and drawing inspiration from promising outcomes of combined audio and video signal detection in different areas, this study aims to harness the strengths of both imagery and sound from the operational environment of conveyor belts. The goal is to overcome existing challenges and significantly boost the accuracy and adaptability of methods detecting longitudinal tears.

Based on the existing approach to audio and video signal detection, and allowing for the limitations of the existing methods, this paper proposes the use of both images and sounds of conveyor belt captured from the worksite as the source of information to detect the tear of the conveyor belt, which can improve the accuracy of detection and the adaptability to the working environment. The main contributions of this paper are as follows:

1. A method of conveyor belt tear detection based on audio-visual fusion is proposed. This method achieves an accurate of detection that is 1.2% higher than using images alone, and 2.8% higher than using sound alone.
2. The audio signals captured from where the conveyor belt works are processed using a feature extraction algorithm that combines Gammatone Frequency Cepstral Coefficients (GFCC) and Linear Frequency Cepstral Coefficients (LFCC), thereby improving the accuracy of audio signal recognition.
3. According to the accuracy of image and sound classification, a decision fusion rule suitable for conveyor belt tear detection is developed for classifying the operating state of conveyor belts.

2 Related works

This section provides a succinct introduction to the background and related works on the detection methodology grounded in image and sound fusion. Currently, there has been some significant progress achieved in the methods of detection that combine audio and video for language emotion analysis and online video analysis. In 2017, Poria conducted the inaugural comprehensive literature review on various domains of affective computing [16]. After delineating the influences of diverse single factors on analysis outcomes, he summarized the prevailing methods for information fusion across different modalities. In 2018, Mohammad and colleagues crafted an audio-visual speech recognition system [17], employing a bimodal information combination framework grounded in feature fusion strategies and utilizing artificial neural network mathematical models. In 2019, a model proposed by Deepak Kumar Jain and his team was underpinned by a singular deep convolutional neural network [18], comprising convolutional layers and deep residual blocks. Initially, labels for all facial images were set for training. Subsequently, images were passed through the proposed DNN model. Their contribution was the categorization of each image into one of six facial emotion categories. In 2021, Zhang et al. [19] used a method that combined sounds and images to analyze the features in sound videos, with better results achieved than conducting MFCC feature analysis alone. In 2022, Francis et al. [20] used a method combining audio and video to analyze the audio and video data of urban surveillance, which enabled the detection of ongoing violent incidents, thus enhancing the safety of residents.

With the evolution of detection methods integrating image and sound in linguistic sentiment analysis, scholars have introduced this technique into the realm of fault detection, positioning it as the avant-garde mode of fault and damage detection. In 2021, Liu et al. [21] adopted a method combining audio and infrared video signals to detect the faults in conveyor belt idlers. Given the complex working environment where belt conveyors work, they effectively detected faults to achieve a better outcome of detection than the method based on a single source of information. In 2022, DE DONATO et al. [22] engineered a railway maintenance deep learning sound and image defect detection system using smart audio and video sensors. Deployed in railway upkeep, this system conferred notable benefits upon railway departments, especially rendering commendable results in predictive maintenance.

3 Material and methods

Due to the harsh environment in which conveyor belts operate, they are required to operate as normal when left unattended. The conveyor belt longitudinal tear detection method based on image and sound fusion primarily consists of image classification, sound classification, decision fusion, and state output, as illustrated in Fig. 1. In this paper, the operation of conveyor belt is divided into two categories: normal and tearing. If no tear occurs, the conveyor belt operates as normal; once tear is detected, an alarm is raised, and the operation of conveyor belt is stopped.

The method of tearing detection for conveyor belts, which is based on audio-visual fusion, relies on a linear CCD camera to capture the images of conveyor belt and on a microphone to collect audio signals during the operation. Then, an improved Shufflenet_V2 network is used to classify the images, and a CNN-LSTM network is used to classify the audio signals processed by LFCC and GFCC in combination. Finally, the classification of images and sounds is inputted into the decision module for output. Various measures such as alarms and shutdowns are taken according to the exact results.

3.1 Image classification

Since the conveyor belt tear detection system is required to produce an excellent real-time performance, the selection of an image classification model must be based on the consideration given to not only the accuracy of the model but also the time taken for image detection. In this paper, the Shufflenet_V2 network, which performs well in accuracy, is chosen for the final image classification network. The process flow of the image classification method is depicted in Fig. 2. Initially, images are captured via a linear CCD camera, followed by image data preprocessing using wavelet denoising and image enhancement algorithms. Subsequently, an improved Shufflenet_v2 network is utilized for image classification.

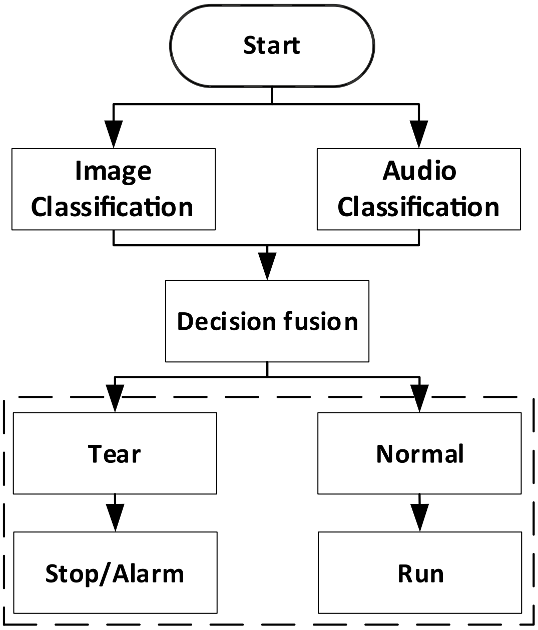


Fig. 1 Flow chart of this method

3.1.1 Improved Shufflenet_V2 network

The core idea of the Shufflenet_v2 network [23] is to solve the problem of information not flowing between different

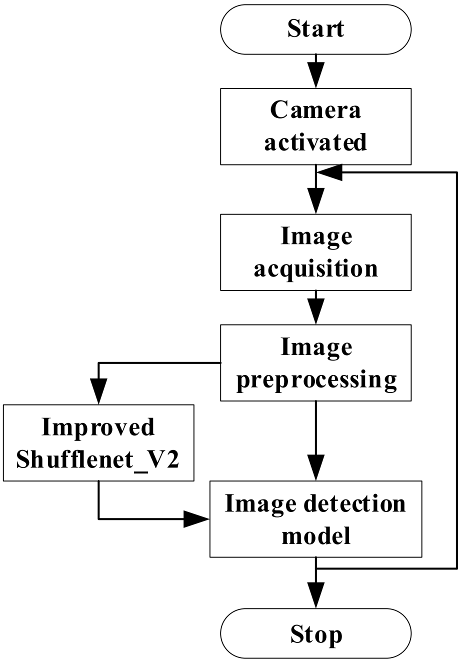


Fig. 2 Flow chart of image classification

Table 1 Shufflenet_V2 network structure

Layer	Stride	Size	FLOPs(M)
Input	–	224*224	–
Conv 3×3	2	112*112	9.0
Max pool	2	56*56	0.3
Stage1	2	28*28	27.8
Stage2	2	14*14	56.4
Stage3	2	7*7	32.7
Conv 1×1	1	7*7	23.4
PC	–	–	0.01

channels through channel shuffling. The network is designed based on the Stage module, as shown in Table 1.

In this paper, the Shufflenet_v2 network is optimized. By introducing the CBAM [24] attention mechanism into the original Shufflenet_v2 network, the accuracy of the model is improved. The CBAM attention module consists of two parts: the Channel Attention Module (CAM) based on the channel attention mechanism and the Spatial Attention Module (SAM) based on the spatial attention mechanism. Through Global Average Pooling (GAP) and Generalized Max Pooling (GMP), CAM can be used to compress the spatial information of the feature values, for example, the original c*h*w dimension information is compressed to c*1*1dimension. Then, the MLP module is applied to compress the channel number into 1/r of the original. With the expansion back to the original number of channels, the results are added after two activation element by ReLU. Finally, the sigmoid activation function is used to output the results. SAM performs max pooling and average pooling operations on the result of CAM to obtain two 1*H*W feature maps, while Concat is used to splice the two obtained feature maps for obtaining a 2*H*W feature map. Also, the result is returned to a 1*H*W feature map through a 7*7 depth-separable convolution. Finally, sigmoid operation is conducted on the result, as shown in Fig. 3.

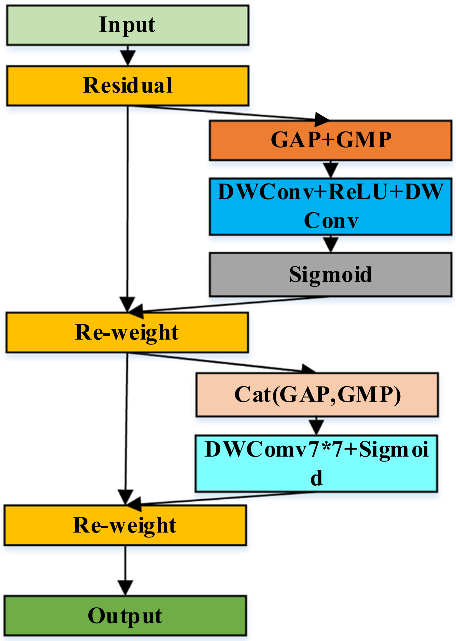


Fig. 3 The structure diagram of CBAM

The CBAM attention module is introduced after each Stage module to analyze complex visual information rapidly and efficiently, thus improving the accuracy of classification for the network. The improved structure is shown in Fig. 4.

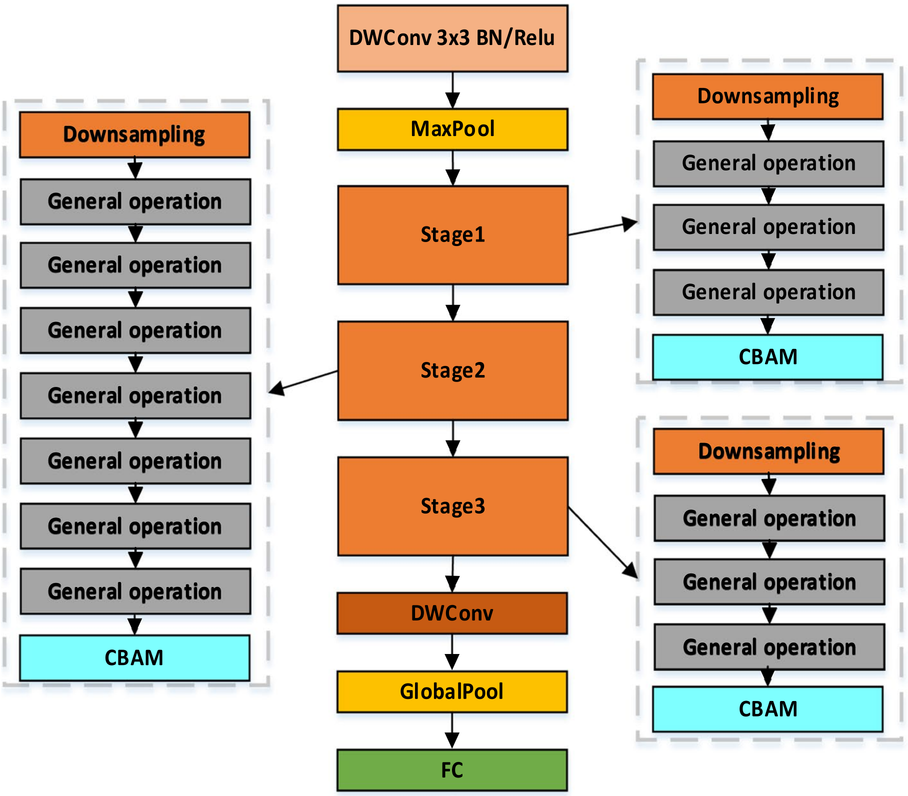


Fig. 4 The improved structure

3.1.2 Transfer learning

Due to the difficulty in capturing the images of conveyor belt tears, the number of images that can be used for training is relatively small. To prevent the overfitting caused by an insufficient number of datasets, the public dataset ImageNet is first used for model training to construct a pre-trained model of the improved Shufflenet_v2 network. Then, transfer learning is performed using the datasets of the conveyor belt tear image.

3.2 Sound classification

With the advancement of artificial intelligence, deep learning has now been widely practiced in the field of sound classification, with some remarkable results achieved. In this paper, a method of sound classification based on the CNN-LSTM model is used to classify the audio signals acquired from the conveyor belt. The workflow for sound classification is shown in Fig. 5.

The proposed method initiates by capturing audio signals from the conveyor belt worksite using an array of microphones. Subsequently, the gathered audio signals are denoised through a wavelet denoising algorithm. The denoised audio signals are then processed using an audio preprocessing algorithm; thereafter, features are extracted using a combination of LFCC + GFCC methods. Finally, the features of extracted audio signals are inputted into the CNN-LSTM network for machine learning, thus achieving the detection of conveyor belt tears.

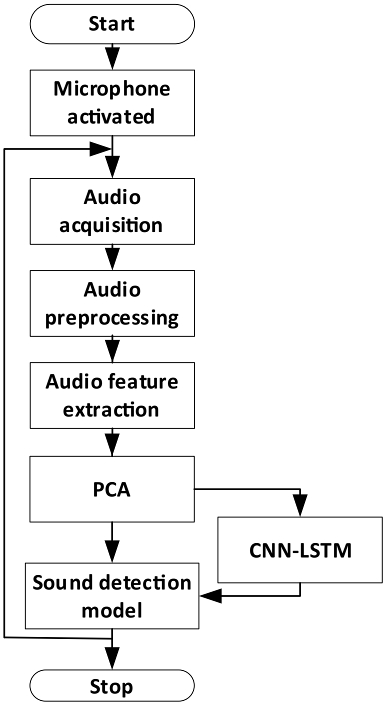


Fig. 5 Flow chart of sound classification

3.2.1 Preprocessing algorithm

The audio data preprocessing algorithm consists of pre-emphasis, framing, windowing, and Fast Fourier Transform (FFT), as shown in Fig. 6.

Firstly, a high-pass filter is used to enhance the strength of signal carrying the audio data in the high-frequency part. Then, the continuous audio signal is decomposed into the short data of 25 ms each. Furthermore, the Hamming window is used to window the data, the purpose of which is to make the segmented audio signals better meet the periodic requirements during FFT operation. The Hamming window function formula is expressed as follows:

$$W(n) = (1 - \mathrm{a}) - \mathrm{a} \times \cos\left[\frac{2\pi n}{\mathrm{N} - 1}\right], 0 \le n \le \mathrm{N} - 1, \quad (1)$$

Finally, FFT is performed on the data to obtain the frequency domain data, and the preprocessed audio signals are integrated by frame to obtain the spectrum of the audio signals.

$$X(i) = \sum_{i=0}^{\mathrm{N}-1} x(n)e^{-j2\pi i/\mathrm{N}}, 0 \le i \le \mathrm{N}, \quad (2)$$

3.2.2 GFCC + LFCC feature extraction algorithm

Feature extraction plays the most important role in audio signal recognition. In this paper, a GFCC + LFCC algorithm is applied to extract features from the audio signals of conveyor belt. Given the complexity in the sound of operating conveyor belt, which contains various noises, it is difficult to recognize the tearing sounds among them. Therefore, the robust GFCC is adopted as the main feature recognition algorithm, while the LFCC, which is highly efficient in detecting full-band sound signals, is used as the auxiliary algorithm for feature extraction.

(1) GFCC

The GFCC feature extraction algorithm is developed for the auditory characteristics of humans. In this algorithm, Gammatone is taken as a filter that can be used to simulate

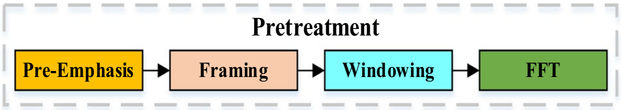


Fig. 6 The audio data preprocessing algorithm

the characteristics of frequency decomposition for a standard cochlea [25]. It is a linear filter described by using an impulse response, where the impulse response is the product of a gamma distribution and a sine tone. The time-domain expression is expressed as follows:

$$H(t) = \begin{cases} At^{n-1}\exp(-2\pi It)\cos(2\pi f_0 t + \varphi), & t \geq 0 \\ 0, & t < 0 \end{cases}, \tag{3}$$

where A represents the amplitude; n indicates the order of the Gammatone; I denotes the time decay coefficient; f_0 refers to the central frequency; φ means the initial phase.

(2) LFCC

The LFCC feature extraction algorithm is similar to GFCC as both of them rely on filters to process the preprocessed signals. The difference lies in the LFCC feature extraction algorithm using a linear triangular filter group, which is expressed as follows:

$$LFCC_i = \sum_{i=1}^{N} X_i \cos\left(\frac{m(i-0.5)\pi}{N}\right), m = 0 \ldots M, \tag{4}$$

where X_i represents the log energy output of each filter group, N indicates the order of the filter, and M denotes the dimension of the feature.

(3) GFCC + LFCC mixed feature extraction

The GFCC + LFCC mixed feature extraction performed in this study is intended to build a new feature matrix by concatenating the feature matrices extracted by GFCC and LFCC. Considering the dynamics of audio signals, the difference in calculation results of GFCC and LFCC is also taken as an auxiliary feature for concatenation. In this way, a feature matrix is obtained that shows strong robustness and sufficient frequency adaptability. The result obtained after concatenation is:

$$\begin{aligned} GFCCLFCC = [&(G_1, G_2 \ldots G_m), (\Delta G_1, \Delta G_2 \ldots \Delta G_n), \\ &(L_1, L_2 \ldots L_i), (\Delta L_1, \Delta L_2 \ldots \Delta L_j)] \end{aligned} \tag{5}$$

where L represents LFCC features, G represents GFCC features, ΔL represents the first-order difference of LFCC features, and ΔG represents the first-order difference of GFCC features.

(4) PCA dimensionality reduction

The feature matrix constructed by using the GFCC + LFCC mixed feature extraction algorithm proposed in this paper consists of GFCC, the first-order difference of GFCC, LFCC, and the first-order difference matrix of LFCC. If each of them has 40-dimensional features, then the final feature matrix is obtained as a 160-dimensional feature matrix. However, its high computational complexity is adverse to the implementation of the real-time detection method used in this paper, which makes it necessary to reduce the dimension of the final feature matrix. In this paper, the Principal Component Analysis (PCA) [26] algorithm is applied to analyze the feature matrix. Figure 7 illustrates the proportion of the first 40-dimensional feature values for LFCC, D-LFCC, GFCC, and D-GFCC, where the line marked at 0.98 indicates the position where the feature contribution reaches 98%. According to the analysis of the PCA algorithm, when the first 28 dimensions are taken by the LFCC feature, GFCC takes the first 16 dimensions, the first-order difference of LFCC takes the first 16 dimensions, and the first-order difference of GFCC takes the first 12 dimensions, with 98% of the effective information carried by the recombined feature matrix.

3.2.3 Building CNN-LSTM network

LSTM and CNN are the focus of research on speech recognition. According to the latest study of speech classification, a combination of different networks can lead to an excellent detection performance. In this paper, a CNN-LSTM network is built for classifying the sound of conveyor belt tear. The CNN-LSTM network configuration is depicted in Fig. 8. From the illustration, it is apparent that the model input is directly connected to a CNN network, which effectively mitigates variations in the frequency domain. The CNN network comprises six convolutional layers followed by three max-pooling layers. Subsequently, features with enhanced adaptability are fed into a two-layer LSTM network to glean more effective detection results. The process culminates with an output through a fully connected (FC) layer. Throughout the convolution (Conv) operation in the network, depthwise separable convolution (DWConv) is performed to reduce the computational workload on the network effectively.

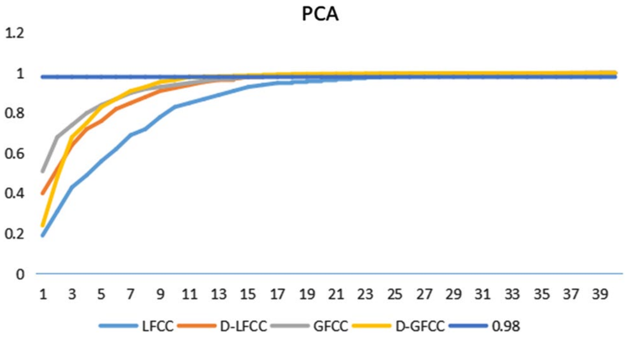


Fig. 7 The PCA analysis results

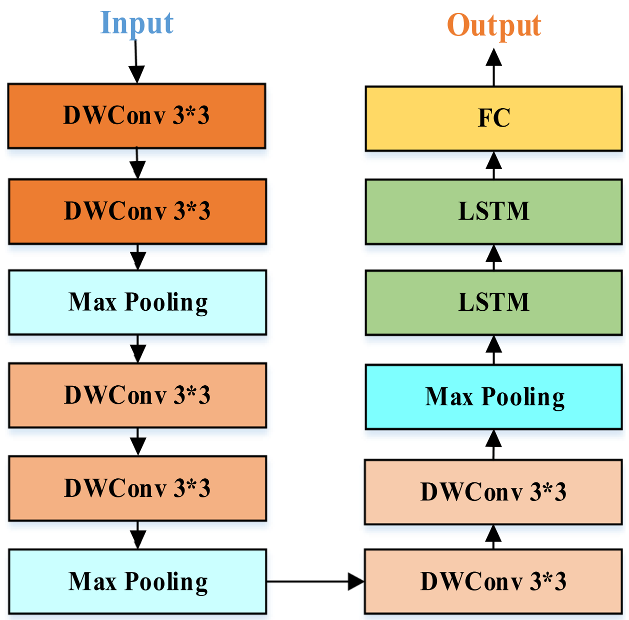


Fig. 8 The structure diagram of CNN-LSTM model

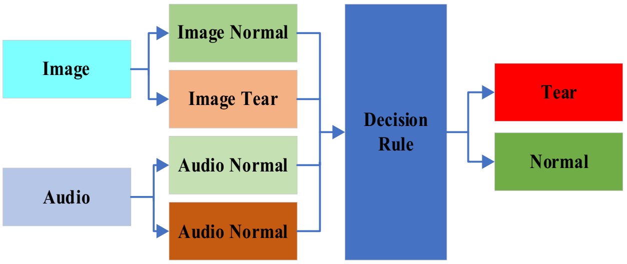


Fig. 9 The decision-making strategy through the Dempster–Shafer theory

3.3 Decision fusion strategy

In view of the complex working environment for the conveyor, both audio and video signals are used in this paper as the source of information for the fusion detection of conveyor belt tears. Considering the high requirements of the detection system for real-time performance, it is proposed in this paper to develop the decision-making strategy through the Dempster-Shafer theory. The decision-making strategy is illustrated in Fig. 9. The operating status of the conveyor belt is divided into two categories: normal and tear. The decision-making result of the running status of the conveyor belt is generated through the combination of image classification and audio signal classification. The results of image classification include normal images and torn images, and those of audio classification include normal audio and torn audio.

Table 2 Probability of classification

	Image	Audio
Normal	C_{IN}	C_{AN}
Tear	C_{IT}	C_{AT}

Table 3 New classifications and probabilities

	Image	Audio
Normal	$P_{I*}C_{IN}$	$P_{A*}C_{AN}$
Tear	$P_{I*}C_{IT}$	$P_{A*}C_{AT}$
Unknown	1-P_I	1-P_A

Firstly, image data are individually trained using the improved Shufflenet_v2 network proposed in this paper to obtain an accuracy rate P_I for classifying the state of conveyor belt operation. Then, the audio classification method proposed in this paper is applied to test the classification of audio data, with an accuracy rate P_A achieved for the sound classification method. The image classification method and audio classification method proposed in this paper are used respectively to classify real-time image data and audio data collected from where the conveyor belt works, thus obtaining the respective confidence level C_{IN} for normal images, the confidence level C_{IT} for torn images, the confidence level C_{AN} for normal sound, and the confidence level C_{AT} for torn sound. Table 2 lists the results of original classification and probabilities.

Due to the difference in accuracy of detection by the image classification model and the audio classification model, the accuracy of detection should be considered as a weighting parameter for calculating the final confidence level. To address the problem that the sum of probabilities obtained from different classifications is not equal to 1 after the introduction of weighting parameters, an "unreliable" classification is performed to accommodate the probability of inaccurate model predictions. This result is expressed in a new classification and probabilities as shown in Table 3:

Following the calculation process based on Dempster-Shafer theory, the normalization constant K must be calculated. According to the formula,

$$\mathrm{K} = \sum_{B\cap C\neq\emptyset} m_P(N)\cdot m_I(T), \tag{6}$$

it can be known that $\mathrm{K}=P_I{*}C_{TN}{*}P_A{*}C_{AN}+P_I{*}C_{IT}{*}P_A{*}C_{AT}+(1\text{-}P_I){*}(1\text{-}P_A)$.

Therefore, the Dempster-Shafer theory is applied to calculate the confidence levels P_N for normal state, torn state P_T, and unreliable state P_U, respectively.

$$P_N = P_I * C_{IN} * P_A * C_{AN}/\mathrm{K}, \tag{7}$$

$$P_T = P_I * C_{IT} * P_A * C_{AT}/\mathrm{K}, \tag{8}$$

$$P_U = (1{-}P_I){*}(1{-}P_A)/K, \tag{9}$$

Finally, the classification and probabilities based on Dempster-Shafer theory are obtained as shown in Table 4.

Given the harsh working environment where the conveyor belt operates, there might be a single source of information available to detect a torn state. Therefore, the threshold values T_P and T_A are set for the confidence levels of tearing conditions detected by using image data and audio signals, respectively. When one of the information sources exhibits strong confidence in conveyor belt tearing, the final state should be outputted as a tearing state regardless of whether the other source of information is used to detect a tearing state. This approach is effective in addressing the conflicts in Dempster-Shafer theory. The decision fusion strategy proposed in this paper is shown in Fig. 10.

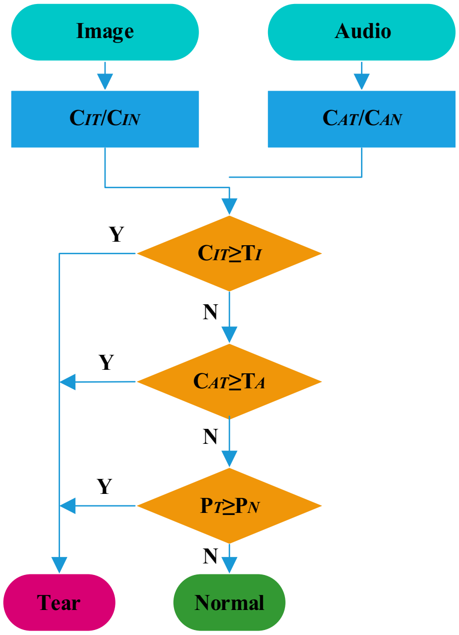


Fig. 10 The flow chart of decision fusion strategy

4 Results

4.1 Data

The dataset used in this study can be divided into two types: image data and audio data. The team designed an audio-visual data acquisition and storage device for data collection. The device consists of an industrial microphone, a control processing unit, an industrial switch, and a storage hard drive. A photograph of the physical device is shown in Fig. 11.

Due to the challenging nature of obtaining images of belt conveyor tears in operational settings, the dataset for this study is divided into two parts: one part consists of data collected under the working conditions of belt conveyors, while the other part is acquired from a laboratory-based system. The audio-visual data collection and storage device was installed in both industrial field environments and laboratory settings to gather sound and image data.

In operational conditions, image data is collected by using an industrial linear CCD camera. Considering the severe dust pollution at the place where the conveyor belt works, a linear light source and a protective cleaning device were installed to assist the collection of high-quality images. The linear light source is comprised of multiple LEDs, which are distributed according to the shape of the conveyor belt. The protective cleaning device consists of a high-definition protective cover and multiple water and jet devices, as shown in Fig. 12.

Acquiring image data in a laboratory environment is relatively simpler compared to operational conditions; the same industrial linear CCD camera is used to capture images of the conveyor belt. The environment of the conveyor belt in the laboratory setting is shown in Fig. 13.

Table 4 The classification and probabilities based on Dempster–Shafer theory

	Image	Audio	Decision
Normal	$P_I{*}C_{IN}$	$P_A{*}C_{AN}$	P_N
Tear	$P_I{*}C_{IT}$	$P_A{*}C_{AT}$	P_T
Unknown	1-P_I	1-P_A	P_U

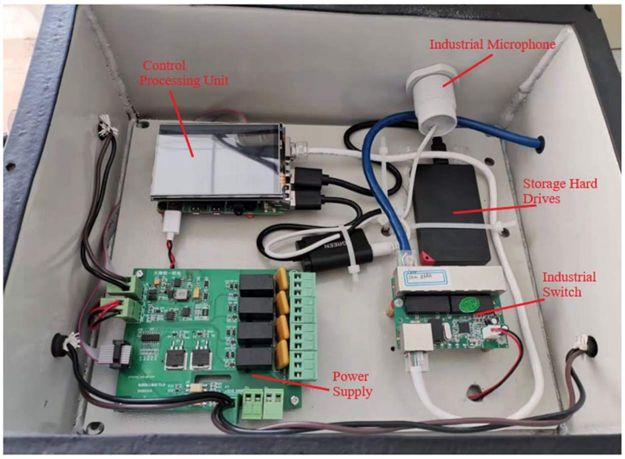


Fig. 11 The audio-visual data acquisition and storage device

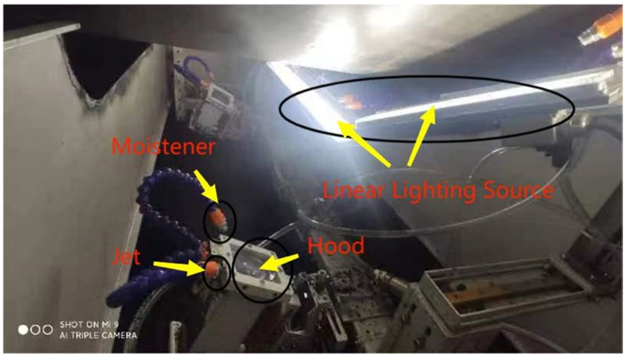


Fig. 12 The installation diagram of camera and linear light source

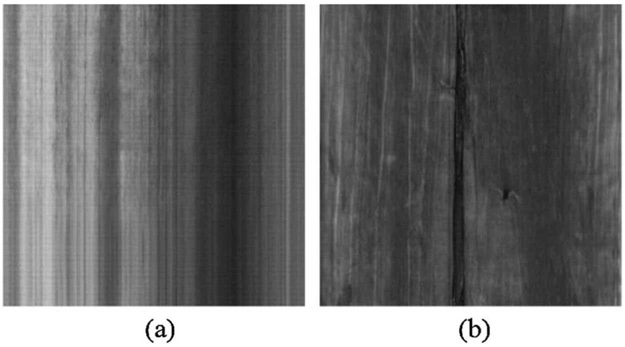


Fig. 14 Images collected: **a** images of normal; **b** images of Tear

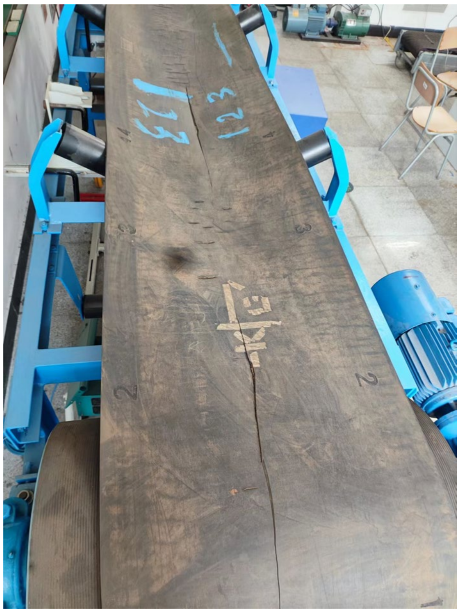


Fig. 13 The conveyor belt in the laboratory environment

The high-quality images obtained for this study are divided into the images of torn belts and those of normal ones. In this study, a total of 10,000 images of the conveyor belt were acquired, which included 4000 images of tears, with 1000 obtained under operational conditions and 3000 obtained in the laboratory; furthermore, 6000 images of the conveyor belt in normal condition were collected, with 2000 of those captured under operational conditions and 4000 in the laboratory setting. A training and testing dataset is created for the transfer learning by the improved Shufflenet_v2 model. Figure 14 displays the images of the conveyor belts collected during the study.

Audio data is collected by using an industrial microphone array. As tears often occur where the material is dropped, the industrial microphone array is installed under the conveyor belt at that point. With a total of 400 s of audio data pertaining to conveyor belt tears were acquired, with 100 s captured under operational conditions and 300 s recorded in the laboratory. Additionally, 600 s of audio data representing normal operational sounds were collected, with 200 s from operational conditions and 400 s from laboratory settings, a dataset is created for model training. Figure 15a represents the time-domain waveform of normal sound, while Fig. 15b illustrates the time-domain waveform of tearing sound.

4.2 Experimental environment

In this study, a cyclic testing system is used for the belt conveyors designed for testing. The system consists of two 25 m main conveyors (L1, L2) and two 4 m transfer conveyors (L12, L21), as shown in Fig. 16a. The linear CCD camera, linear light source, and microphone array are installed between the upper and lower belts of the main conveyor L1, as illustrated in Fig. 16b. The PC configuration intended for model training and testing in this study is as follows: Windows10 system, i9-12900 K CPU, Nvidia RTX3090 graphics card, 32 GB memory, 2 TB solid-state drive, and Python.

4.3 Experimental results

Firstly, the image dataset is used to perform transfer learning in the improved Shufflenet_v2 model. Before the model is trained, it is necessary to set the training parameters in a reasonable way to yield better results. The training parameters are listed in Table 5. The Stochastic Gradient Descent (SGD) optimizer is used to update the network parameters of the model for obtaining the minimum loss function. The accuracy and loss of model training are shown in Fig. 17.The accuracy of the improved Shufflenet_v2 model for the classification of conveyor belt tear images reaches 94.3%.

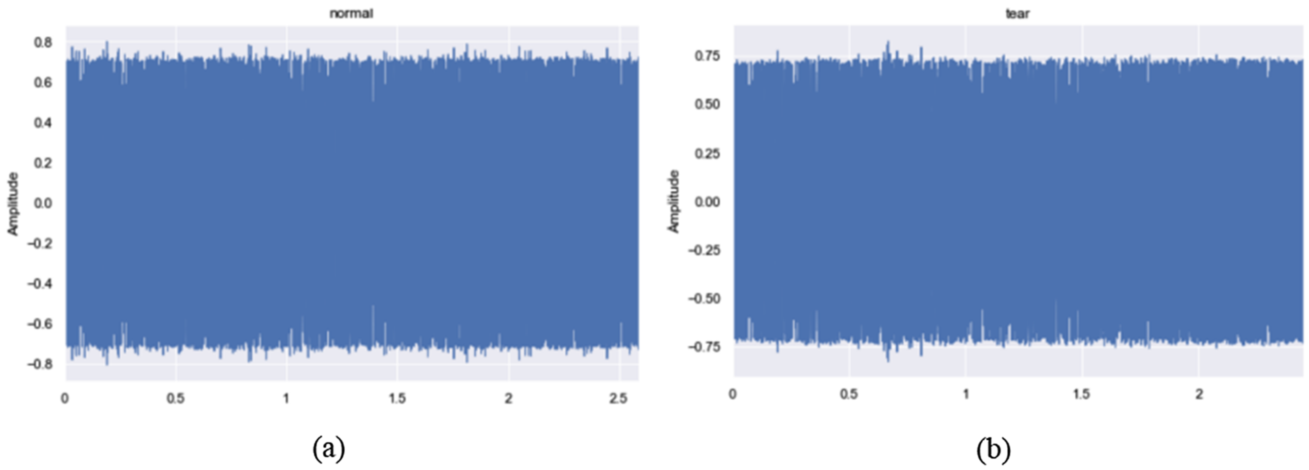


Fig. 15 The sound data collected in this experiment: **a** Normal; **b** Tear

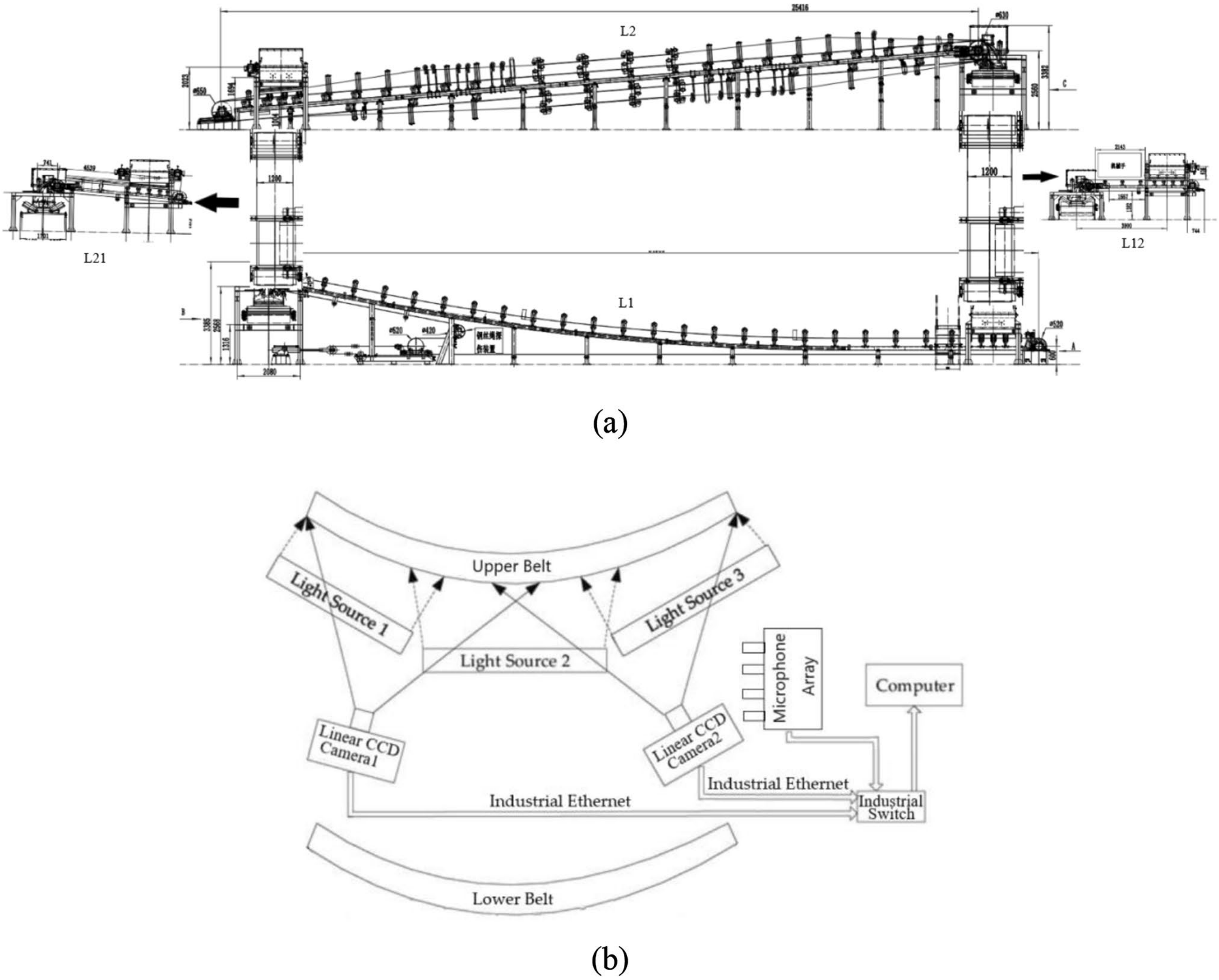


Fig. 16 The belt conveyor system used in this experiment: **a** the cyclic conveyor belt system; **b** the installation diagram of this article

Table 5 Training parameters

Parameters	Value
Input size	256 × 256
Batch	32
Epoch	300
Optimizer	SGD
Initial learning rate	1/64
Momentum	0.9
Weight decay	0.00004

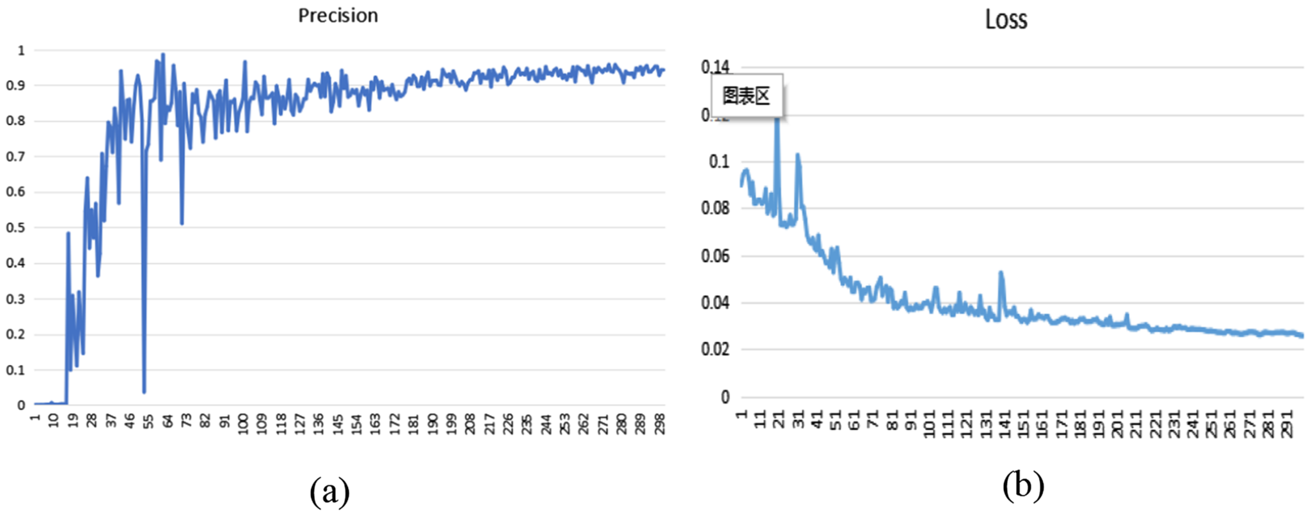


Fig. 17 The results of the improved Shufflenet_v2 model training: **a** precision; **b** loss

Table 6 The results of comparing

Model	Precision (%)	Recall (%)	F1 (%)	MFLOPs
Densenet	86.6	85.2	85.9	142
Xception	92.8	91.6	92.2	145
MobilenetV2	93.6	93.4	93.5	145
This paper	94.6	93.8	94.2	146

Table 7 The results of the sound detection method used in this paper are compared with other methods

	MFCC (%)	IMFCC (%)	LFCC (%)	GFCC (%)	This paper (%)
KNN	88.6	80.6	86.8	87.1	90.6
SVM	93.6	90.4	92.1	91.8	93.4
LSTM	92.1	88.1	90.2	90.1	93.1
CNN-LSTM (this paper)	93.5	90.1	92.4	92.2	94.2

5 Discussion

The method proposed in this study is compared with three other mainstream neural networks, namely Densenet, Xception, and MobilenetV2 networks. The results obtained through this comparison are shown in Table 6.

For the sound dataset, the sound classification method proposed in this paper is applied. Comparative experiments are developed to verify the advantages of this method, with accuracy as the main evaluation index for the comparison of results. Firstly, different methods of feature extraction are compared. In this paper, the LFCC + GFCC feature extraction method is compared with traditional MFCC, IMFCC, LFCC, and GFCC. At the same time, different classification methods are compared. In this paper, a CNN-LSTM model is built for audio signal classification, and then compared with other classification methods including KNN, SVM, and LSTM, the results as shown in Table 7.

After being labeled, the image data and audio data are integrated to generate voiced video data to verify the advantages of the detection method proposed in this paper. Based on the image recognition accuracy of 94.6% and the audio classification accuracy of 94.2%, the weight parameters w_A, w_P of the decision-making strategy developed in this paper are obtained through calculation. Then, 1000 images are extracted from the image dataset, including 400 images of torn belt and 600 images of normal ones, to produce a 40-s video. The final voiced video is produced by inserting the tearing sound and normal running sound into the corresponding positions of the video file, with the tearing sound corresponding to the tearing image and the normal sound corresponding to the normal image. Longitudinal tearing of conveyor belts, once occurring, can lead to substantial economic losses. The later the tear is detected, the greater the financial impact, making the real-time aspect of detection a critical metric for assessment methodologies. Table 8 presents a comparative analysis of the proposed method with existing prominent fault detection techniques, including the conveyor belt longitudinal tear detection methods of Yang et al. [9], Qiao et al. [10], Li and Miao [11], Zhang et al. [14], and the sound and infrared image detection results utilized by Liu et al. [21]. Liu's method requires the conversion of the visible light image dataset into an infrared image dataset for training and testing purposes.

It can be seen that the accuracy achieved by using the methods proposed by Yang and Li is relatively low, and that the method put forward by Zhang achieves a relatively high accuracy of 93.5%. In comparison, the accuracy achieved by the method proposed in this paper is 96%, which is 2.5% higher than the method proposed by Zhang. Therefore, the overall performance in audio-visual fusion is better than when the detection methods requiring the use of either image or sound alone are used. Furthermore, the methods utilized by Yang and Li are based on traditional image detection algorithms, mainly relying on CPU and memory for computation, which results in slower detection rates. In contrast, the approaches adopted by Qiao, Zhang, Liu, as well as the method presented in this paper, employ GPU acceleration, enabling rapid detection. The implementation of the method presented in this paper is based on a hardware environment consisting of an i9-12900 K CPU, Nvidia RTX3090 graphics card, and 32 GB of RAM. Under these hardware conditions, the average detection time of our method is 32 ms, which is capable of effectively processing audio and video data at a frame rate of 30 FPS, thus meeting the requirements for real-time detection.The confusion matrix representing the test results of the proposed method is depicted in Fig. 18.

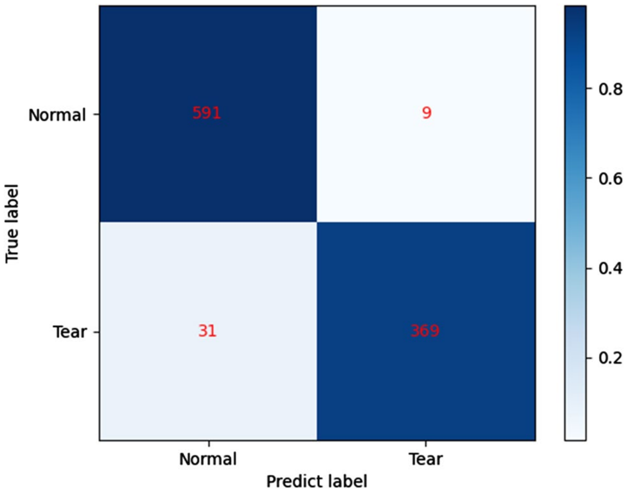


Fig. 18 The confusion matrix of this method

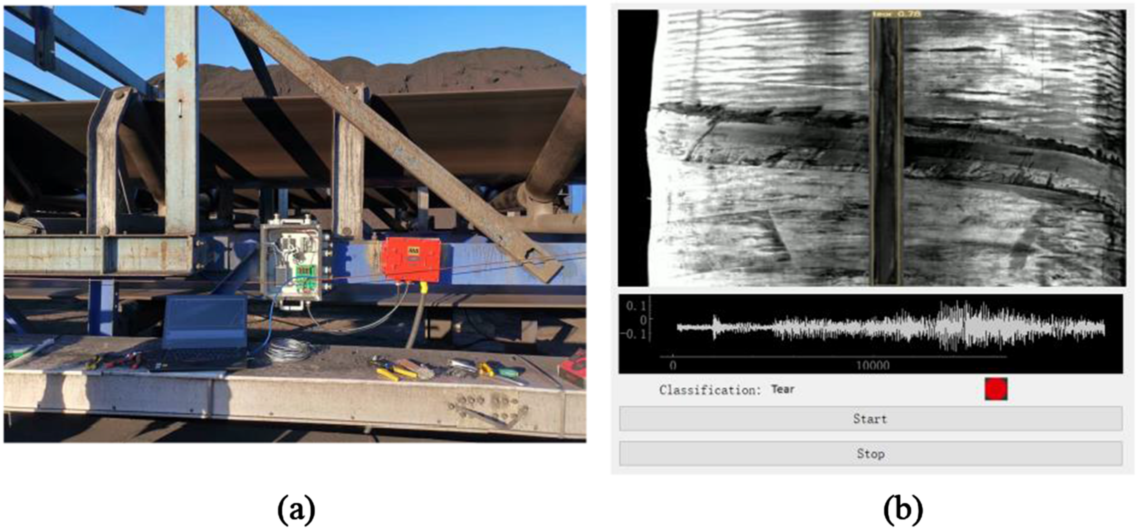
The confusion matrix indicates that there were 31 instances of missed detections (where a tear state was incorrectly identified as normal) and 9 instances of false detections (where a normal state was incorrectly identified as a tear). The rate of missed detections is notably higher than that of false detections. Analysis of the detection errors reveals that the missed detections mainly occur because the presence of a tear in the image does not necessarily coincide with the distinctive sound of tearing. Additionally, uneven lighting conditions in the operational environment can result in missed detections of tears in images.

The team has developed a longitudinal tear detection system for conveyor belts and has installed and tested it on conveyor machinery in mining sites, achieving satisfactory results. Figure 19a depicts a photograph of the tear detection system installed in a production environment, and Fig. 19b shows a screenshot of the detection system software interface.

The dataset presented in this study consists of data collected from laboratory settings as well as several mine shaft operational environments. Based on the analysis of the experimental process and results, we conclude that

Table 8 The results of this method compared with other classification

	Tear		Normal		Accuracy (%)	Time (ms)
	Tear	Normal	Tear	Normal		
Yang et al. [9]	301	99	118	482	78.3	328
Li & Miao [11]	323	77	79	521	84.4	226
Zhang et al. [14]	368	32	33	567	93.5	30.9
Qiao et al. [10]	370	30	32	568	90.8	29.7
Liu et al. [21]	376	24	16	584	93.6	33.5
This method	376	24	16	584	96	31.9



Classification: Tear

Fig. 19 The pictures of a belt conveyor working in a coal mine: **a** The photograph of the tear detection system; **b** the detection system software interface

the limitations and challenges encountered in the actual application process are as follows:

(1) Regarding the sound data, the environmental noise in real operational conditions affects the analysis. This study only addresses the environmental noise present in the existing data collection; however, during the system's actual operation, there may be incompatibilities with environmental noise from different types of belt conveyors and scenarios, which could impact the recognition accuracy.

(2) For image data, different operational conditions present varied actual interference sources. The quality of the image is a key factor in ensuring the accuracy of image detection. This article employs image enhancement algorithms to denoise and enhance images under operational conditions. However, the adaptability of the existing image denoising and enhancement algorithms to different operational conditions will be a challenge for the system's accurate operation.

In this study, linear LED light sources are used to effectively increase image brightness; high-definition protective covers, water spraying, and air jet devices are utilized to protect and clean the camera. However, due to the complexity of actual operational conditions, ensuring image quality over a long period will also be a challenge that needs to be addressed by the method proposed in this paper.

6 Conclusions

To address the poor accuracy and stability of single-modal longitudinal tear detection for conveyor belts, a longitudinal tear detection method is proposed for conveyor belts in this paper on the basis of audio-visual fusion to monitor the operating status of the conveyor belt. The method utilizes an improved Shufflenet_v2 model for the detection of conveyor belt tear images, and constructs a CNN-LSTM network for sound detection after extracting sound features using LFCC + GFCC. Decision fusion is then performed based on the obtained results from the conveyor belt tear image detection and sound detection to determine the final status of the conveyor belt operation. The results show that the performance of this method is superior to that of the detection methods in which either image or audio data is used.

7 Future work

In future research, we will first focus on further optimizing the methods for image and sound data acquisition under actual operating conditions to improve the quality of collected data. Subsequently, we will conduct research on data denoising and enhancement algorithms tailored to the characteristics of images and sound signals under operating conditions. Lastly, we will further investigate methods for integrating image and sound signals, conducting comparative validation between feature-level fusion and decision-level fusion, to identify more suitable methods for integration.

Acknowledgements This work was supported by Science and technology think tank youth talent plan (Grant No. 20220615ZZ07110010); National Natural Science Foundation of China-Shanxi coal-based low-carbon joint fund (Grant No. U1810121) and the Natural Science Foundation of Shanxi (Grant: No. 201801D121180).

Declarations

Conflict of interest The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

1. Andrejiova, M., Grincova, A., & Marasova, D. (2016). Measurement and simulation of impact wear damage to industrial conveyor belts. *Wear, 368*, 400–407.
2. He, D., Pang, Y., & Lodewijks, G. (2017). Green operations of belt conveyors by means of speed control. *Applied Energy, 188*, 330–341.
3. Cao, H. (2015). Study and analysis on tear belt and break belt of belt conveyor in coal mine. *Coal Science and Technology, 43*(S2), 130–134.
4. Peng, X. (2013). A novel image-based method for conveyor belt rip detection. In *IEEE International Conference on Signal Processing*.
5. Zakharov, A., Geike, B., Grigoryev, A., & Zakharova, A. (2020). Analysis of devices to detect longitudinal tear on conveyor belts. In *E3S Web of Conferences; EDP Sciences: Kemerovo, Russia*, volume 174, p. 03006.
6. Dobrota, D. (2015). Vulcanization of rubber conveyor belts with metallic insertion using ultrasounds. In Katalinic, B. (Ed.) 25th Daaam International Symposium on Intelligent Manufacturing and Automation, 2014, pp. 1160–1166.
7. Kozłowski, T., Błażej, R., Jurdziak, L., & Kirjanów-Błażej, A. (2019). Magnetic methods in monitoring changes of the technical condition of splices in steel cord conveyor belts. *Engineering Failure Analysis, 104*, 462–470.
8. Kozłowski, T., Wodecki, J., Zimroz, R., Błażej, R., & Hardygóra, M. (2020). A diagnostics of conveyor belt splices. *Applied Sciences, 10*, 6259.
9. Yang, Y., Miao, C., Li, X., & Mei, X. (2014). On-line conveyor belts inspection based on machine vision. *Optik—International Journal for Light and Electron Optics, 125*, 5803–5807.
10. Qiao, T., Li, X., Pang, Y., Lu, Y., Wang, F., & Jin, B. (2017). Research on conditional characteristics vision real-time detection system for conveyor belt longitudinal tear. *IET Science, Measurement & Technology, 11*, 11955–11960.
11. Li, J., & Miao, C. (2016). The conveyor belt longitudinal tear online detection based on improved SSR algorithm. *Optik, 127*(19), 8002–8010.
12. Xianguo, L., Lifang, S., Zixu, M., Can, Z., & Hangqi, J. (2018). Laser-based on-line machine vision detection for longitudinal rip of conveyor belt. *Optik (Stuttg)., 168*, 360–369. https://doi.org/10.1016/j.ijleo.2018.04.053
13. Yang, Y. L., Qiao, T. Z., Pang, T. Z., & Yan, S. (2020). Infrared spectrum analysis method for detection and early warning of longitudinal tear of mine conveyor belt. *Measurement, 165*, 107856.
14. Zhang, M., Shi, H., Zhang, Y., Yu, Y., & Zhou, M. (2021). Deep learning-based damage detection of mining conveyor belt. *Measurement, 175*, 1–9.
15. Miao, D., Wang, Y., & Li, S. (2022). Sound-based improved DenseNet conveyor belt longitudinal tear detection. *IEEE Access, 10*, 123801–123808. https://doi.org/10.1109/ACCESS.2022.3224430
16. Poria, S., Peng, H., Hussain, A., Howard, N., & Cambria, E. (2017). Ensemble application of convolutional neural networks and multiple kernel learning for multimodal sentiment analysis. *Neurocomputing, 261*, 217–230. https://doi.org/10.1016/j.neucom.2016.09.117
17. Rahmani, M. H., Almasganj, F., & Seyyedsalehi, S. A. (2018). Audio-visual feature fusion via deep neural networks for automatic speech recognition. *Digital Signal Processing, 82*, 54–63.
18. Shrivastava, K., Kumar, S., & Jain, D. K. (2019). An effective approach for emotion detection in multimedia text data using sequence based convolutional neural network. *Multimedia Tools And Applications, 78*(20), 29607–29639.
19. Zhang, J., Wen, X., Cho, A., & Whang, M. (2021). An empathy evaluation system using spectrogram image features of audio. *Sensors, 21*, 7111. https://doi.org/10.3390/s21217111
20. Reinolds, F., Neto, C., & Machado, J. (2022). Deep learning for activity recognition using audio and video. *Electronics, 11*, 782. https://doi.org/10.3390/electronics11050782
21. Liu, Y., Miao, C., & Li, X. (2021). Research on the fault analysis method of belt conveyor idlers based on sound and thermal infrared image features. *Measurement, 186*, 110177.
22. de Donato, L., Flammini, F., & Marrone, S. (2022). A survey on audio-video based defect detection through deep learning in railway maintenance. *IEEE Access, 10*, 65376–65400. https://doi.org/10.1109/ACCESS.2022.3183102
23. Ma, N. N., Zhang, X. Y., Zheng, H. T., & Sun, J. (2018). ShuffleNet V2: Practical guidelines for efficient CNN architecture design. arXiv:1807. 11164v1 [cs.CV].
24. Sanghyun, W., Jongchan, P., Joon-Young, L., & Kweon, I. S. CBAM: Convolutional block attention module. arXiv:1807.06521v2 [cs.CV].
25. Qi, J., Wang, D., Jing, Y., & Liu, R. S. (2013). Auditory features based on Gammatone filters for robust speech recognition. In *IEEE International Symposium on Circuits and Systems*, pp. 305–308.
26. Gupta, V., Saxena, N. K., Kanungo, A., et al. (2022). PCA as an effective tool for the detection of R-peaks in an ECG signal processing. *International Journal of System Assurance Engineering and Management*. https://doi.org/10.1007/s13198-022-01650-0
27. Zou, L., Xia, L., & Ding, Z. (2019). Reinforcement learning to optimize long-term user engagement in recommender systems: ACM. https://doi.org/10.1145/3292500.3330668[P].
28. Wang, Y. M., Miao, C. Y., Liu, Y., & Meng, D. J. (2022). Research on a sound-based method for belt conveyor longitudinal tear detection. *Measurement, 190*, 110787.
29. Chen, M., & Hernández, A. (2022). Towards an explainable model for sepsis detection based on sensitivity analysis. *IRBM, 43*(1), 75–86.
30. Pouard, P., & Collaange, V. (2007). Neuromonitoring par la spectroscopie dans le proche infrarouge en chirurgie cardiaque pédiatrique: Neuromonitoring by near infrared spectroscopy in paediatric cardiac surgery. *IRPM, 28*, 1959–2318.
31. Gupta, V., Mittal, M., & Mittal, V. (2022). A novel FrWT based arrhythmia detection in ECG signal using YWARA and PCA. *Wireless Personal Communications, 124*, 1229–1246.

32. Gupta, V., Mittal, M., & Mittal, V. (2021). FrWT-PPCA-based R-peak detection for improved management of healthcare system. *IETE Journal of Research, 69*(8), 5064–5078.
33. Gupta, A., Gupta, V., Mittal, M., & Mittal, V. (2022). An efficient AR modelling-based electrocardiogram signal analysis for health informatics. *International Journal of Medical Engineering and Informatics, 14*(1), 74.
34. Gupta, V., Mittal, M., Mittal, V., et al. (2022). Detection of R-peaks using fractional Fourier transform and principal component analysis. *Journal of Ambient Intelligence and Humanized Computing*. https://doi.org/10.1007/s12652-021-03484-3
35. Gupta, V., Mittal, M., & Mittal, V. (2021). Spectrogram as an emerging tool in ECG signal processing. *Wireless Personal Communications, 114*(4), 0929–6212.
36. Gupta, V., Mittal, M., & Mittal, V. (2022). A simplistic and novel technique for ECG signal pre-processing. *IETE Journal of Research*. https://doi.org/10.1080/03772063.2022.2135622
37. Ebad, S. A. (2022). Lessons learned from offline assessment of security-critical systems: The case of microsoft's active directory. *International Journal of System Assurance Engineering and Management*. https://doi.org/10.1007/s13198-021-01236-2
38. Amanbek, N., Mamayeva, L. A., & Rakhimzhanova, G. M. (2021). Results of a comprehensive assessment of the quality of services to the population with the use of statistical methods. *International Journal of Systems Assurance Engineering and Management*. https://doi.org/10.1007/s13198-021-01278
39. Alketbi, A., Nasir, Q., & Abu, T. (2020). Novel blockchain reference model for government services: Dubai government case study. *International Journal of System Assurance Engineering and Management, 11*(6), 1170–1191.
40. Gupta, S., Gupta, P., & Parida, A. (2017). Modeling lean maintenance metric using incidence matrix approach. *International Journal of System Assurance Engineering and Management, 8*(4), 799–816.
41. Ye, W., Wang, H., & Zhong, Y. (2022). Optimization of network security protection situation based on data clustering. *International Journal of System Assurance Engineering and Management*. https://doi.org/10.1007/s13198-021-01529-6
42. Xu, Q., Wu, D., Jiang, C., et al. (2022). A composite quantile regression long short-term memory network with group lasso for wind turbine anomaly detection. *Journal of Ambient Intelligence and Humanized Computing, 14*(3), 2261–2274. https://doi.org/10.1007/s12652-022-04484-7
43. Son, Y., Zhang, X., Yoon, Y., et al. (2022). LSTM–GAN based cloud movement prediction in satellite images for PV forecast. *Journal of Ambient Intelligence and Humanized Computing, 14*(9), 12373–12386. https://doi.org/10.1007/s12652-022-04333-7
44. Gundu, V., & Simon, S. P. (2021). PSO–LSTM for short term forecast of heterogeneous time series electricity price signals. *Journal of Ambient Intelligence and Humanized Computing, 12*(2), 2375–2385. https://doi.org/10.1007/s12652-020-02353-9
45. Reznikov, I., Chuprakov, D., & Bekerov, I. (2023). Analytical model of 2D leakoff in waterflood-induced fractures. *Journal of Rock Mechanics and Geotechnical Engineering, 15*(7), 1713–1733.
46. Zeng, L., Zhang, H., Han, Q., et al. (2021). An LSTM-based driving operation suggestion method for riding comfort-oriented critical zone. *Journal of Ambient Intelligence and Humanized Computing*. https://doi.org/10.1007/s12652-021-03327-1
47. Ubaid, A. M., & Dweiri, F. T. (2020). Business process management (BPM): Terminologies and methodologies unified. *International Journal of System Assurance Engineering and Management*, *11*, 1046–1064.

Yimin Wang received the B.S. degree from Tiangong University, Tianjin, China, in 2011 and the M.S. degree from Tiangong University, Tianjin, China, in 2014. He is currently pursuing the Ph.D. degree at school of Mechanical Engineering, Tiangong University, Tianjin, China. His main research fields are fault detection and digital image processing.

Yuhong Du a distinguished faculty member at Tianjin Polytechnic University. She holds a Ph.D. and mentors doctoral students. Her research expertise lies in the fields of mechanical engineering and pattern recognition, where she has contributed significantly through her extensive academic work.

Changyun Miao a renowned professor at Tianjin Polytechnic University. With a doctorate degree and a role as a doctoral advisor, his research direction focuses on information technology and telecommunications engineering. His work has made considerable impacts on the advancement of his field.

Di Miao received the B.S. degree from Beijing Jiaotong University, Beijing, China, in 2009 and the Ph.D. from Beijing Jiaotong University, Beijing, China, in 2016. She is currently a professor in Tianjin University of Technology and Education, Tianjin, China. Her main research fields are digital image processing and network.

Yao Zheng is a doctoral candidate at Tiangong University.

Dengjie Yang is also pursuing a Ph.D. at Tiangong University.