Longitudinal tear detection method of conveyor belt based on audio-visual fusion

Che, Jian; Qiao, Tiezhu; Yang, Yi; Zhang, Haitao; Pang, Yusong

DOI
10.1016/j.measurement.2021.109152

Publication date
2021

Document Version
Accepted author manuscript

Published in
Measurement: Journal of the International Measurement Confederation

Citation (APA)
Che, J., Qiao, T., Yang, Y., Zhang, H., & Pang, Y. (2021). Longitudinal tear detection method of conveyor belt based on audio-visual fusion. *Measurement: Journal of the International Measurement Confederation*, *176*, Article 109152. https://doi.org/10.1016/j.measurement.2021.109152

Longitudinal Tear Detection Method of Conveyor Belt Based on Audio-visual Fusion

Jian Che[a,b], Tiezhu Qiao [a,b,*], Yi Yang[a,b] , Haitao Zhang [a,b], Yusong Pang [c],

[a] Key Laboratory of Advanced Transducers and Intelligent Control System, Ministry of Education, Taiyuan University of Technology, Taiyuan 030024, China

[b] College of Physics and Optoelectronics, Taiyuan University of Technology, Taiyuan 030024, China

[c] Section of Transport Engineering and Logistic, Faculty of Mechanical, Maritime and Materials Engineering, Delft University of Technology, 2628 CD Delft, Netherlands

Abstract: Conveyor belt tear detection is a very important part of coal mine safety production. In this paper, a new method of detecting conveyor belt damage named audio-visual fusion (AVF) detection method is proposed. The AVF method uses both a visible light CCD and a microphone array to collect images and sounds of the conveyor belt in different running states. By processing and analyzing the collected images and sounds, the image and sound features of normal, tear and scratch can be extracted respectively. Then the extracted features of images and sounds are fused and classified by machine learning algorithm. The results show that the accuracy of AVF method for conveyor belt scratch is 93.66%, and the accuracy of longitudinal tear is higher than 96.23%. Compared with existing methods AVF method overcomes the limitation of visual detection condition, and is more accurate and reliable for conveyor belt tear detection.

Keywords: Audio-visual feature extraction; Feature fusion; Longitudinal tear detection method; Machine learning

1.Introduction

Belt conveyor is an indispensable transportation equipment in coal mine production. Conveyor belt is the most accident-prone part of the whole belt conveyor [1],[2]. During the operation of the conveyor belt, gangue, anchored bolt and other metal hard objects at the transfer point may scratch and tear the conveyor belt[3],[4]. Learn from the research methods of turning tool wear[5],[6] we discover conveyor belt scratches are a precursor to longitudinal tear and longitudinal tear is longer and more common than horizontal tearing of conveyor belts[7]. Owing to the long transport distance of coal, if a longitudinal tearing accident occurs, the entire belt may be replaced, and normal production and transportation of the coal mine cannot be restored for a long period of time, which causes huge economic losses. If workers do not find the accident in time, it will lead to blockage of coal mine roadway, and would even cause casualties. Therefore, fast and accurate damage detection methods for conveyor belts are essential

* Corresponding author
E-mail address: qiaotiezhu@tyut.edu.cn (Tiezhu Qiao).

for coal mine safety production.

At present, it has become a trend to use computer vision technology to detect conveyor belt longitudinal tear. For example, Qiao and Li[8] used a laser and a surface light source to build a visual recognition system for longitudinal tear of the conveyor belt. Yang et al. [9]proposed a fast image segmentation algorithm based on line array CCD camera. Wang et al. [10] proposed a non-contact conveyor belt tear detection method based on image processing and pattern recognition. Wang and Sun[11] proposes a conveyor belt longitudinal tear detection method based on Haar-AdaBoost and Cascade algorithm under uneven light. However, the visible light CCD (Charge Coupled Device) may be affected by the environment in the mine due to the dark, watery and dusty working environment in the mine. Therefore, the detection accuracy will be greatly affected. Qiao et al. [12] proposed an binocular vision detection method based on the fusion of infrared and visible light. Yang and Qiao[13] proposed a longitudinal tear warning method based on infrared image ROI selection and image binarization. Yang and Qiao[14] proposed Infrared spectrum analysis method for detection and early warning of longitudinal tear. The infrared method is used to detect tearing, the principle is that when the conveyor belt is damaged by friction with hard impurities, thermal radiation is generated through the conveyor belt and detected by infrared CCD, However, if the tear process of the conveyor belt is slow and the temperature cannot rise rapidly, the thermal radiation cannot be detected by infrared CCD through the conveyor belt in time and there will be missed detection. Therefore, the validity and accuracy of test results will be affected as before. In order to adapt to the complex environment under the coal mine and improve the accuracy of tear detection, a new detection method which enables to adapt to the dark and dusty environment as well as to meet timely and accurate demand is needed.

At present, computer vision and sound detection technology are developing rapidly, the two are widely used in medical[15],[16], transportation[17],[18], security inspection[19],[20] and other fields such as emotion recognition[21] and acoustic scene classification[22]. Computer vision technology can replace the human eye for recognition, tracking and measurement, but the quality of computer vision image is very sensitive to obstacles, occlusion and light conditions. On the one hand, audio detection is equivalent to human ears. The sound signal is not affected by the light can still transmit information effectively when encountering obstacles. Sound signal has the advantages of convenient collection, low cost and space saving. The use of sound detection can be very good auxiliary computer vision and applied to the detection system. Audio-visual detection is not only non-contact but also more accurate and stable. When the conveyor belt rubs against hard impurities, it will produce a sharp sound, which can be well distinguished from the normal operation of the conveyor belt.

Image and sound are the most obvious features of longitudinal tearing or scratching of the conveyor belt. In summary, this paper proposes a longitudinal tear detection method based on audio-visual fusion (AVF) conveyor belt. The AVF method can not only adapt to the complex environment under the coal mine, but also detect the scratches of the conveyor belt more accurately.

Sound image information fusion and feature extraction are two key steps in AVF method. In terms of information fusion, according to the theory of information fusion[23], Data fusion can be done on the feature layer, data layer and decision layer. Since the image and sound are heterogeneous in nature, and their data have no correlation with each other, it is almost impossible to carry out information fusion on the data layer. When the conveyor belt is torn longitudinally, there is a certain correlation between the generation of tear and the sound produced. If they are fused at the decision layer, their correlation will be separated, which will lead to the decrease of detection accuracy. Therefore, the feature layer is selected for the data fusion of sound and image.

In terms of feature extraction, we extracted MFCC(Mel-Frequency Ceptral Coefficients), Spectral centroid, Short-time energy, ZCR(Zero Crossing Rate) and Spectral roll-off as sound features[24]. For image features, we mainly want to obtain the image contour information when the conveyor belt is torn longitudinally, so we extract the HOG (Histogram of Oriented Gradient) feature. After extracting the features of sound and image, network fuses two features. The fused features were sent to the machine learning model for training, and the trained model was used to identify the new tear. Compared with the previous methods. The AVF method adds sound features to the traditional computer vision detection of longitudinal tear of the conveyor belt. It not only overcomes the difficulty of collecting data with visible light CCD in the complex environment of coal mines, but also effectively solves the problem of missed detection by infrared detection methods. We analyzed the image and sound features, and for the first time proposed the fusion of sound features and image features in the feature layer, which eliminate the redundancy and contradiction between image and sound information and complement each other, and improve the real-time and reliability of the longitudinal tear detection of the conveyor belt.

The organization of this paper is as follows. Section 2 mainly introduces AVF methods, which includes sound feature extraction, image feature extraction, image and sound feature fusion and machine learning model. Section 3 gives the experimental results and analysis to verify the AVF method, we compared the accuracy of using image features alone and using audio-visual fusion features, and compared the AVF method with some visual detection methods that have been proposed, followed by the conclusions and possible improvements are discussed in Section 4.

2. AVF method

This section mainly introduces the method of AVF, which consists of four parts including image feature extraction, sound feature extraction, feature fusion and machine learning methods. Firstly, the image of tear and is scratch captured by visible light CCD and its features are extracted. Secondly, we use a microphone array to collect tear and scratch sounds and extract sound features. Thirdly, the fusion of image features and sound features at the feature layer, the fused audio-visual features were sent to the machine learning model for training. Finally, the existing tear or scratch can be distinguished by audio-visual fusion detection model of conveyor belt longitudinal tearing. The AVF method flowchart is shown in Fig.1. Each section is described as following section.

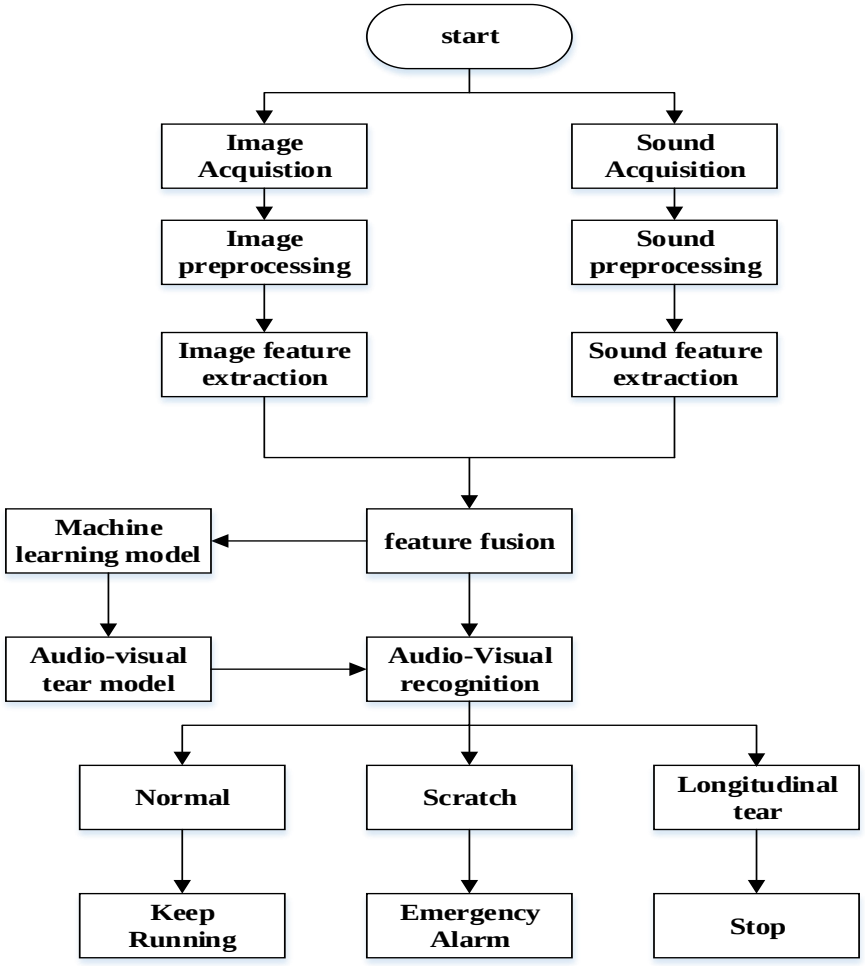


Fig.1.Flow chart of AVF method

2.1. Data acquisition

In order to train a good conveyor belt longitudinal tear audio-visual detection model,

it is very important to collect audio-visual data. We use visible light CCD and microphone to collect the image and sound data of normal operation, scratch and longitudinal tear of the conveyor belt at the same time. The specific acquisition methods of image and sound are as follows

2.1.1. Image acquisition

Since most of the longitudinal tear occurs at the transfer point of the conveyor belt[25] , we install the visible light CCD under the belt near the transfer point. Due to the dark and dusty working environment in the mine, it is hard to capture clear tear images using a visible light CCD alone, which greatly affects the accuracy of longitudinal tear detection. In order to overcome the impact of the complex environment of the coal mine on the captured images, auxiliary light sources and dust removal equipment need to be installed. We install the dust removal equipment on the front of the visible light CCD, and install the auxiliary light source below the visible light CCD as shown in Fig.2. The visible light CCD collects images of the under different conditions. The collected images are divided into three categories: the normal images, the scratched image and the longitudinal tear image are shown in Fig.3.

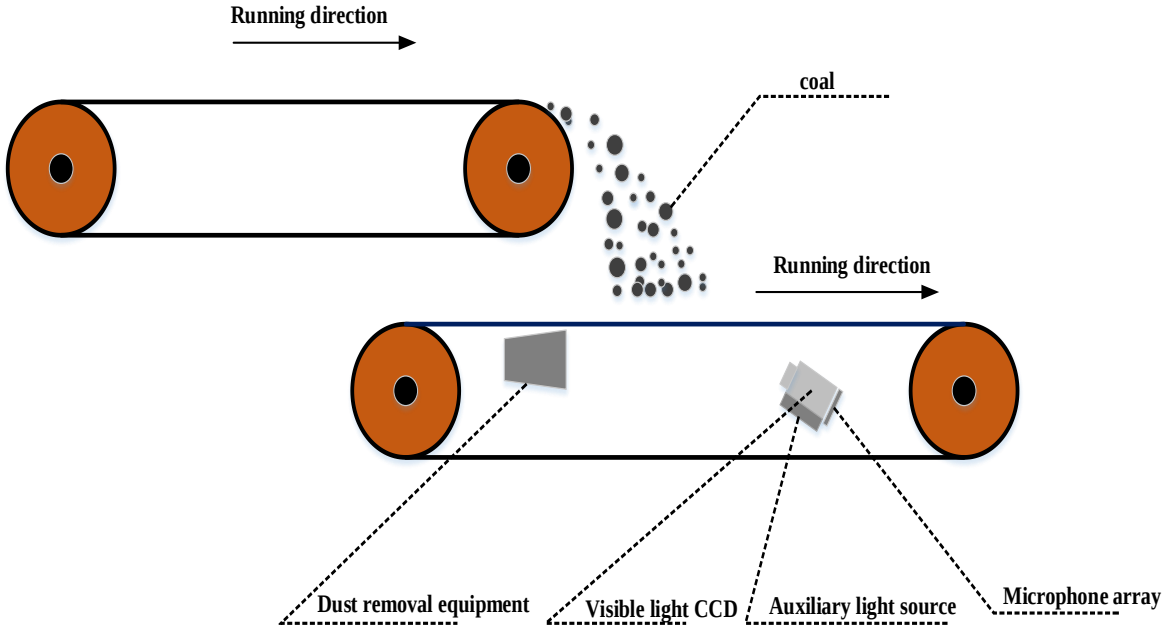
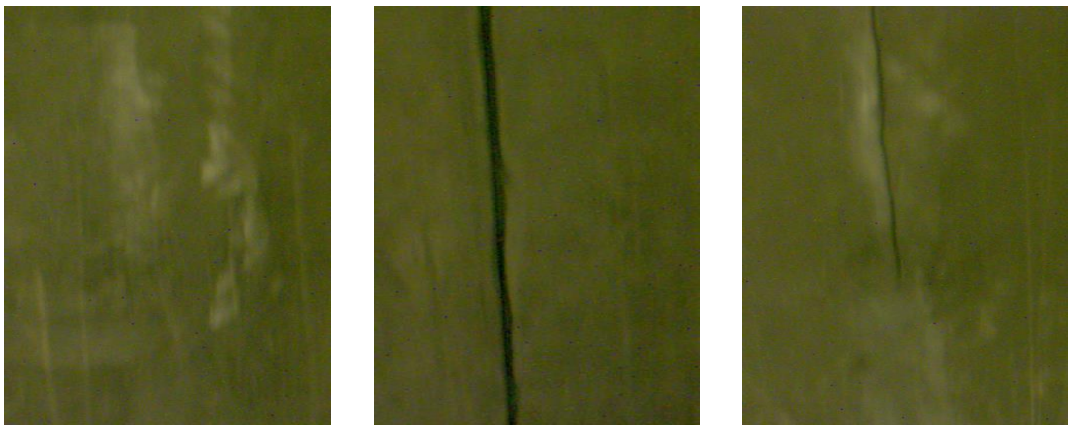


Fig. 2. Dust removal equipment of visible light CCD and auxiliary light source



(a) (b) (c)

Fig. 3. The collected images:(a) normal image. (b) tear image (c) scratch image

2.1.2 Sound acquisition

Microphone array is installed at the bottom of visible light CCD to collect the sound of the conveyor belt during normal transportation, the sound of scratches and the sound of tear. Setting the sampling frequency to 44.1khz. In order to eliminate the influence of aliasing, high-order harmonic distortion, and other factors on the quality of the sound signal caused by sound collection equipment and environmental sound, we need to preprocess the collected original sound. Pre-processing the sound signal under different conditions of the conveyor belt includes two parts: pre-emphasis and framing and windowing, pre-emphasis compensates for the high-frequency part of the sound, which improves the signal-to-noise ratio of the high-frequency part of the sound signal. Windowing divides the continuous sound signal into independent frame signals and smooth the frame signals, which is convenient for calculation in the frequency domain.

2.2 Data preprocessing and feature extraction

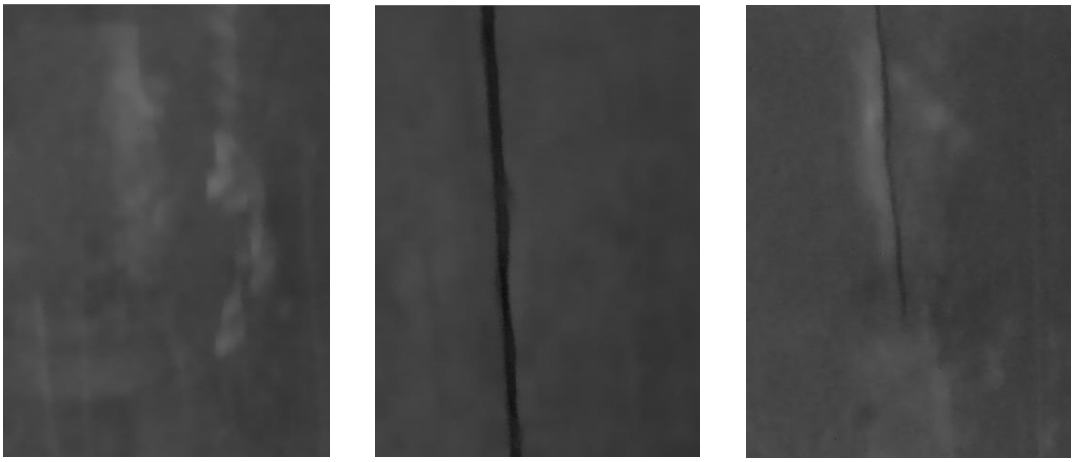
We preprocess the collected images and extract image features from the preprocessed images. For tear and scratch images, we found through experiments that the contour features of tears and scratches have good performance, so the contour features of the edges of tears and scratches are extracted. By comparing the sound of longitudinal tear, the sound of scratch and the sound of normal operation of the conveyor belt, it is found that there are obvious differences among them. Therefore, sound features can be used as an important feature in longitudinal tear detection of conveyor belt. There are two parts included in sound feature extraction: sound pretreatment and feature extraction. In this paper, MFCC, ZCR, Spectral-centroid, Short-term energy and Spectral roll-off are extracted as the feature of sound.

2.2.1Image preprocessing

In the collected visible light image, the damage area is obvious. However, in the background, except for the damaged area, in the actual working environment of coal mine, there will be some scratches and stains in the pictures we take. In order to better extract the damage features, we use median filtering to filter the image [26]. Assuming that the conveyor belt image captured by the Visible light CCD is $g(s,t)$ then the image processing formula with the pixel value $f(x,y)$ after the median filtering is as follows:

$$f(x,y) = \underset{(s,t)=S_{xy}}{median}\{g(s,t)\} \quad (1)$$

In order to extract features in the next step, grayscale the image after median filtering. As shown in Fig.4.



(a) (b) (c)

Fig. 4. The processed images:(a) normal image; (b) tear image; (c) scratch image.

2.2.2. Image HOG feature extraction

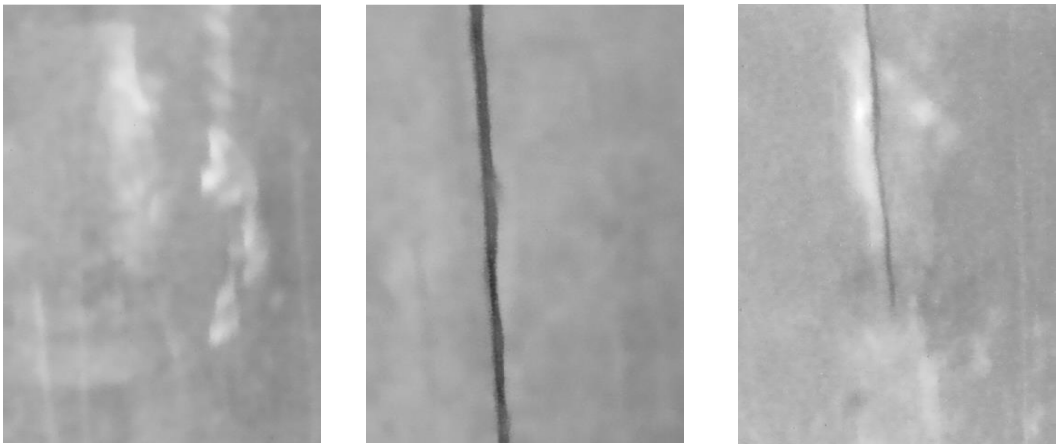
Dalal and Triggs[27] proposed an image feature description algorithm HOG based on gradient direction. HOG algorithm calculates and statistics the gradient and direction of image pixels, and calculates gradient histogram of local images to construct features, which can describe the edge of the detection object well. Compared with other image feature extraction algorithms, HOG operates on the local grid cells of the image, so it can maintain good invariance to image geometric and optical deformations, and can tolerate different forms between tears feature. The steps for the HOG algorithm are described as follows.

(1) Gamma correction

In order to overcome the impact of the dark and dusty mine on local shadows and image brightness changes, highlight the damaged parts of the picture, we first used gamma correction to process the image. Gamma correction can improve the image contrast effect of the darker or brighter part of the image. The formula of Gamma correction is as follows:

$$f(x) = x^{\gamma} \tag{2}$$

Where x is the pixel value of the image, and γ is the Gamma correction coefficient. Here, $f(x)$ is the output pixel value. In this paper, the correction coefficient is 0.8, and the image after gamma correction is shown in Fig. 5



（a） (b) (c)

Fig. 5. The Image after gamma correction:(a) normal image; (b) tear image; (c) scratch image.

(2) Calculate image gradient

We use the statistical gradient method to obtain the longitudinal tearing profile information of the conveyor, the gradient in mathematics is actually the first derivative. The gradient of a continuous image at a certain pixel point can be calculated by the following formula.

$$G_x(x, y) = H(x+1, y) - H(x-1, y) \tag{3}$$

$$G_y(x, y) = H(x, y+1) - H(x, y-1) \tag{4}$$

Where $G_x(x, y)$is the vertical gradient at point(x, y), $G_y(x, y)$ is the horizontal gradient, $H(x, y)$ is the pixel value at point (x, y), the gradient amplitude $G(x, y)$and direction $a(x, y)$at point $(x, y)$are calculated as follows:

$$G(x, y) = \sqrt{G_x(x, y)^2 + G_y(x, y)^2} \tag{5}$$

$$a(x, y) = \tan^{-1}(\frac{G_y(x, y)}{G_x(x, y)}) \tag{6}$$

(3) Calculate HOG feature

First, the image was divided into 8 * 8 pixels of small cell, then the gradient direction of 360 degrees of cell can be divided into 9 bins, 9 bins of the histogram is used to statistics the 8 * 8 pixels gradient information, finally to calculate the pixel gradient direction projection to its corresponding histogram, In this way, the weighted projection of the gradient size and direction of each pixel in the cell on the histogram is the corresponding 9-dimensional feature vector of the cell. We use four cells to form a block of 16*16 pixels. It's going to join up to form a 36*1member vector. Then, the

block window moves with a fixed step size (8 pixels per step) to normalized the histogram, thus generating a standardized 36*1 vector for each move. The 36 features obtained after each movement are concatenated together as the final feature of our image. The HOG feature extraction flowchart is shown in the Fig. 6

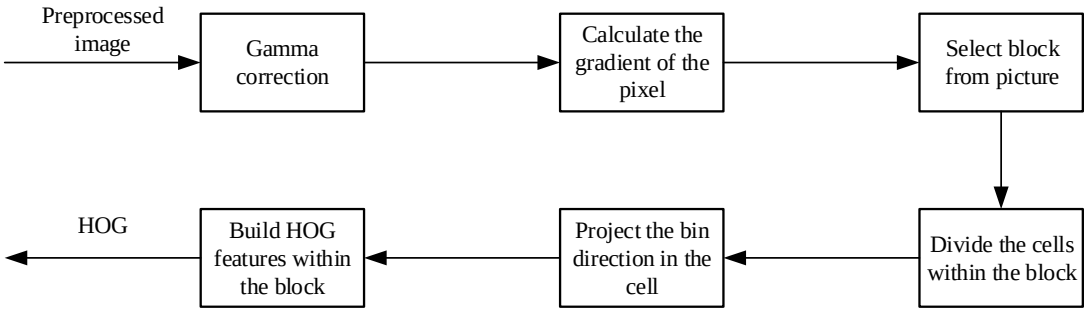


Fig. 6. HOG feature extraction flowchart

2.2.3. Sound preprocessing

(1) Pre-emphasis

Because of the conveyor belt longitudinal tear sound signal collection belt is often affected by various noises under the mine, the high-frequency acoustic signal will attenuate. Therefore, before the sound signal processing, We should enhance the high frequency part of the sound to effectively reduce the output noise, obtain more frequency-domain information, so as to facilitate sound feature extraction[28]. Pre-emphasis usually expression of the transfer function is:

$$H(Z) = 1 - aZ^{-1} \tag{7}$$

Where a is the pre-emphasis coefficient, which is 0.97 in this experiment (usually selected between 0.9 and 1)

(2) Framing and Windowing

Because of the sound signal remains unchanged and relatively stable in a short time range, it is possible to divide sound signal into some short segments for processing Each short segment is called a frame. We use overlapping segmentation method to divide frames. The overlapping segmentation method enhances the correlation between frames and facilitates the smooth transition between them, the overlapping part is called frame shift. In this paper, frame length N=1024 and frame shift T=256, the sampling frequency is set to 44.1KHZ, so each frame is 23.2ms. In order to obtain a smoother spectrum, we use hamming window for framing. The hamming window calculation formula is:

$$w(n) = \begin{cases} (1-a) - a\cos[2\pi n/(N-1)], & 0 \le n \le N \\ 0, & others \end{cases} \tag{8}$$

Where N is the frame length, and different a will generate different hamming windows and here we choose a= 0.46, after determining the window function, add a window to the pre-emphasis sound signal, the formulas for framing and windowing are as follows:

$$s'(n) = s(n) * w(n), \tag{9}$$

Where, $s(n)$ is the original sound of the n-the frame, $s'(n)$ is the sound signal of the n-the frame after windowing and framing processing.

2.2.4. Sound Feature extraction

(1) MFCC

MFCC has been widely used in audio signal analysis[29]. Although MFCC is mainly designed for voice processing, but we found that it can be used to analyze the damage of conveyor belt. The frequency perceived by the human auditory system of low-frequency sounds and the physical frequency of the sound are approximately linear; the frequency perceived by high-frequency sounds and the physical frequency of the sound are approximately logarithmic. The relationship between the Mel frequency and the physical frequency the relationship is shown in the formula.

$$f_{mel} = 2595 \log_{10}(1 + \frac{f_{hz}}{700}) \tag{10}$$

Where f is the frequency the specific steps to extract MFCC parameters from sound signals in different damage states is as follows:

- After the preprocessing of sound signal, we convert the audio signal in the time-frequency domain, it main implementation is Discrete Fourier Transform (DFT). The DFT input is a sequence of frames windowed to signal $s'(n)$, the output is the complex number $x_n(k)$ containing N frequency bands. The definition of DFT is as follows:

$$x_n(k) = \sum_{n=0}^{N-1} s'(n) e^{-j2\pi kn/N}, \qquad 0 \le k \le N \tag{11}$$

  Where k is the kth spectrum of the Fourier transform, and N is the number of points of the Fourier transform.
- Through the Mel filter. The Mel filter is composed of a triangular bandpass filter, which can convert the frequency spectrum into Smooth processing to remove the influence of harmonics and highlight the formants of the original sound, the frequency response of the triangle filter $H_m(K)$ is:

$$H_m(k)\begin{cases} 0, & k < f(m-1) \\ \dfrac{2(k-f(m-1))}{(f(m+1)-f(m-1))(f(m)-f(m-1))}, & f(m-1) \le k < f(m) \\ \dfrac{2(f(m+1)-k)}{(f(m+1)-f(m-1))(f(m+1)-f(m))}, & f(m) \le k \le f(m+1) \\ 0, & k > f(m+1) \end{cases} \tag{12}$$

Where $H_m(k)$ satisfies:

$$\sum_{m=0}^{M-1} H_m(k) = 1 \tag{13}$$

- Take the logarithmic energy, and perform logarithmic operation on the signal passing through the triangular filter bank. The logarithmic energy output by the m-th Mel filter is shown in the following formula.

$$S(m) = \ln\left(\sum_{k=0}^{N-1} \left|x_n(k)\right|^2 H_m(k)\right) \qquad 0 < m < M \tag{14}$$

Where M is the number of Mel scale triangle filters, we select 26 in this paper.

- Through the discrete cosine transform of the energy logarithm $S(m)$ to MFCC.

$$C(m) = \sum_{m=0}^{M-1} S(m)\cos\left(\frac{\pi l(i-0.5)}{M}\right), \qquad l = 0,1,2,\cdots,L \tag{15}$$

Perform discrete cosine transform on $S(m)$ to obtain the Mel-scale cepstral parameter$C(m)$, L is the order of the MFCC coefficient, usually set to 12-16. In this paper, L = 13.

（2） Short-Time Energy

As the energy of the sound signal changes with time, the sound energy of tear, scratches and normal operation of the conveyor belt are significantly different. Therefore, the analysis of short-time energy can describe the characteristic changes in different states of the conveyor belt. The short-time energy is calculated as follows:

$$E_n = \sum_{m=0}^{N-1} s_n^2(m) \tag{16}$$

Where $s_n(m)$is the nth frame sound signal, m is for window position.

（3）ZCR

The zero-crossing rate represents the rate at which the symbols of the sound signal change, the ZCR is mainly used for the recognition when the background noise is large. The calculation formula is as follows:

$$ZCR_n = \frac{1}{2}\sum_{m=0}^{N-1}[\mathrm{sgn}(s_n(m)) - \mathrm{sgn}(s_n(m+1))] \tag{17}$$

Where $sgn()$ is the sign function, the formula is as follows:

$$\operatorname{sgn}(x)=\begin{cases}1, & x\geq 0\\ -1, & x<0\end{cases} \tag{18}$$

（4）Spectral centroid

Spectral centroid is one of the important parameters to describe the timbre attributes. It is the frequency averaged by energy weighted within a certain frequency range, the spectral centroid describes the brightness of sound. It is the important information of the frequency distribution and energy distribution of the sound signal. The calculation formula is:

$$SC_n=\frac{\sum_{f=0}^{F_s/2} fS_n(f)}{\sum_{f=0}^{F_s/2} S_n(f)} \tag{19}$$

Where f is the sound signal frequency, $S_n(f)$ is the spectral energy formula of the corresponding frequency after the discrete Fourier transform of the continuous time domain signal $s_n(m)$ as follows:

$$S_n(f)=\sum_{m=0}^{N-1} s_n(m)e^{-j2\pi kn/N} \tag{20}$$

Where N is the frame length.

（5）Spectral roll-off

Spectrum roll-off is the change of the spectrum amplitude when the frequency is lower than a certain set value. According to the characteristics of the spectrum roll-off, the slope of the spectrum shape can be measured. The calculation formula of spectrum roll-off is as follows:

$$\sum_{k=0}^{m}\left|S_n(k)\right|=\theta\sum_{k=0}^{N-1}\left|S_n(k)\right| \tag{21}$$

Where θ is the threshold, and its value range is 0.85~0.99， In this paper it is 0.85

2.3. Feature fusion

In the feature layer, we fused the sound features and image features of the conveyor belt under different operating states, including data normalization processing and PCA dimensionality reduction.

2.3.1. Data normalization

We serialized the acquired image features and sound features to obtain the new audio-visual fusion eigen matrix X ,and then carried out data normalization processing on the acquired new audio-visual fusion eigen matrix. The calculation formula is

$$x_j^{(i)}=\frac{a_j^{(i)}-\mu_j}{s_j} \tag{22}$$

Where $a_j^{(i)}$ is the j eigenvalue of i samples, μ_j represents the mean value of the j feature, and s_j represents the range of the j feature, that is: $s_j = \max(a_j^{(i)}) - \min(a_j^{(i)})$.

2.3.2.PCA

PCA is a dimensionality reduction algorithm, which can convert high-dimensional data into low-dimensional data with minimal loss[30]. We use PCA algorithm to process audio-visual fusion features, extract useful information and remove redundant information. After the normalization of the data, the audio-visual fusion feature matrix X was obtained, and the covariance matrix was first calculated:

$$C = \frac{1}{m} X^T X \tag{23}$$

Where C is the covariance matrix, and then the singular value decomposition is used to calculate the eigenvectors of the covariance matrix.

$$[U, S, V] = svd(C) \tag{24}$$

After the characteristic matrix is obtained, dimensionality reduction can be carried out on the data. Assuming that the value before dimensionality reduction is $x^{(i)}$, the formula of $Z^{(i)}$ dimension reduction is as follows:

$$Z^{(i)} = U_{reduce}^T x^{(i)} \tag{25}$$

Where $U_{reduce} = [u^{(1)}, u^{(2)}, \cdots, u^{(k)}]$ is principal component characteristic matrix.

2.4. Machine learning model

We classify the collected audiovisual data of conveyor belt damage. In this paper, we have selected three machine learning algorithms, K-nearest neighbors(KNN) algorithm is simple to implement, suitable for multi-classification problems, and not sensitive to abnormal points[31], support vector machine (SVM) biggest characteristic is the maximum distance can be structured decision boundary, generalization error rate is low, It has high robustness, due to the small sample of audio-visual data we collect, SVM is better than other algorithm in the case of fewer data sets[32]. The Random Forest (RF) algorithm is a classification algorithm that combines multiple weak classifiers (decision trees) into a strong classifier. The fusion of the output of multiple classifiers not only helps to improve the accuracy of classification, but also factors such as outliers and noise in the data are well tolerated and are not prone to overfitting[33].

3.Experiment and analysis

In this section, we are simulating the actual environment of coal mines and show the experimental results of the AVF method on the audio-visual data set we collected.

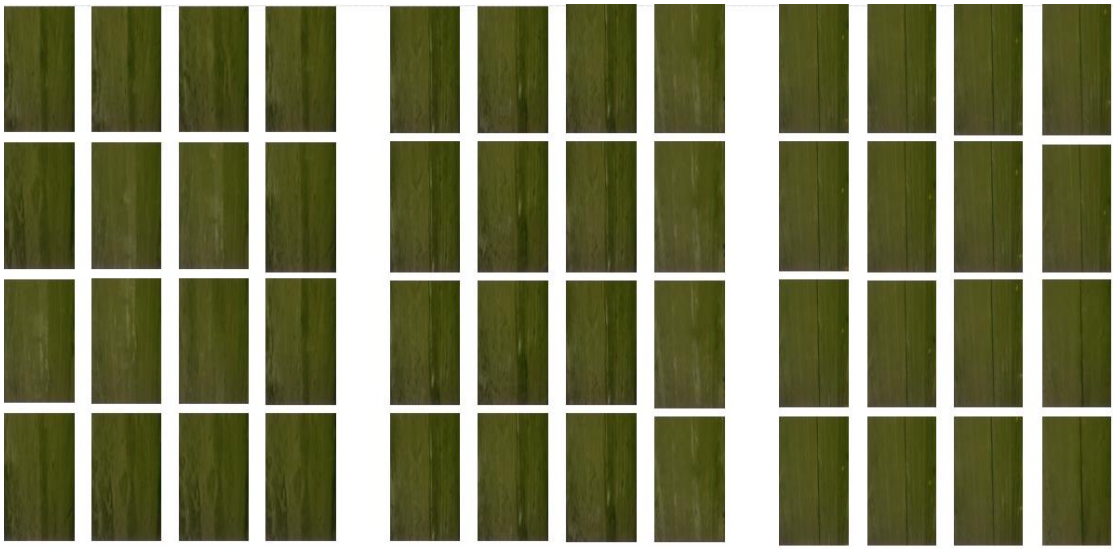
3.1. Experiment setup

In order to prove the effectiveness of the AVF method, we built an experimental platform and since the laboratory itself is located in the basement, we turn off the laboratory light, use only auxiliary light source, and blow dust around visible CCD and conveyor belt to simulate the underground environment of the mine to collect audio-visual data, the picture of the conveyor belt, visible light CCD and microphone array is shown in Fig.7. Using steel wire conveyor belt, the specific parameters of the conveyor belt are as follows: length 13 meters, width 1 meter, thickness 15 mm, longitudinal tensile strength 1250 N/mm, the speed is 4m/s. We installed anchor bolt between the upper and lower conveyor belts to simulate the tear of the upper belt. The visible light CCD is an area-array industrial camera with a resolution of 1280*1080 and a frame rate of 60 fps, the sound acquisition device is a four-array microphone, in the AVF method experiment, we simulate the tear and scratch of the conveyor belt to collect the images and sounds under different running states of the conveyor belt, we set the sampling frequency of the sound to 44.1KHz and the duration of each segment to 1.2s, all sounds are in wav format. A computer is used for sound and image processing and machine learning modeling. The specific parameters of the computer is: CPU is Inter Core i7-7700 3.6GHz, memory is 16GB. The experiment simulated the dark and dusty working environment of the coal mine.

We first place the metal anchor bolt on the surface of the conveyor belt, turn on the conveyor belt to allow it to run normally, then adjust the depth of the anchor bolt so that it slightly penetrates the conveyor belt to simulate the conveyor belt scratch, and then adjust the depth of the anchor bolt to penetrate the conveyor belt to make it through conveyor belt simulates the belt tear. According to the different depth of bolt insertion, the normal operation of the conveyor belt, the belt scratch and longitudinal tear of the conveyor belt are respectively presented, a total of 2,600 including 1000 pictures normal, 800 pictures of the tear and 800 pictures of the scratch were taken as the image data set, the image data set is shown in Fig.8. At the same time, added the collected noise of the conveyor belt running in the mine to the collected sound fragments to simulate the working environment of the coal mine. We collected a total of 2600 sound segments corresponding to the pictures of the conveyor belt in different states as our sound data set including 1000 sound segments during normal operation, 800 sound segments during longitudinal tearing and 800 sound segments during scratch. The schematic diagram of the AVF method experiment is shown in Fig.9.

（a） (b)

Fig. 7. The experiment platform:(a) Conveyor belt, (b) Visible light CCD and microphone array



（a） (b) (c)

Fig. 8. The experimental samples:(a) The normal sample, (b)The scratch sample, (c)The Tear sample

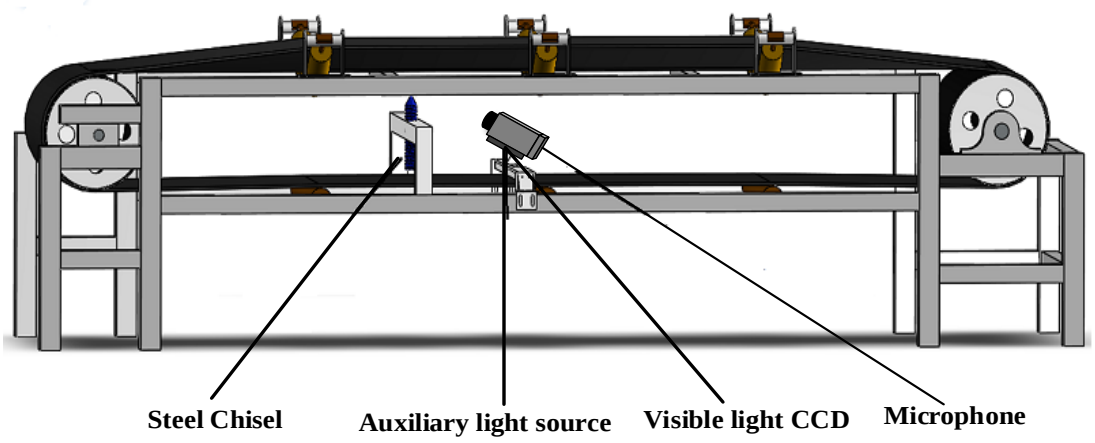


Fig. 9. Schematic diagram of AVF method experiment

All experiments are performing by python3.7, and development platform is Pycharm. The experimental data is divided into three parts, including conveyor belt normal operation of audio-visual data, conveyor belt longitudinal tear of audio-visual data and conveyor belt scratches of audio-visual data, then the three parts of audio-visual data respectively do image and sound of feature extraction and feature fusion to get the audio-visual fusion features of different running states of the conveyor belt. The audio-visual fusion features of different conveyor belt operation states were input into the classifier we selected. The main parameters of the three classifiers are shown in Table.1. We choose these parameters through experiments to optimize the classification performance of conveyor belt audio-visual fusion data as much as possible, and perform3-fold cross-validation to get the final audio-visual fusion detection model. The experimental results and analysis are as followed section:

Table 1

The main parameters of each machine learning model

Models	**Main parameters of the model**	
KNN	K: 1, 5, 10, 15, 20, 25	Distance: 'euclidean', 'cityblock', 'minkowski'
SVM	Kernel function: rbf	C:10^{-1} Gamma: 2^{-12}, 2^{-10}, 2^{-8}, 2^{-4}, 2^{-2}, 2^{0},2^{2}
RF	n_estimators : 2^{3}, 2^{4}, 2^{5}, 2^{6}	min_sample_leaf: 0, 5, 10, 20 max_features: 5

3.2. Experiment result and analysis

We performed PCA analysis on the audio-visual features and visual features, and the results are given in Fig.10, from Fig.10 We can see that the overall score of the audio-visual features is higher than that of the visual features alone, the audio-visual feature dimension has the highest score at 850. We compared three different machine learning classifiers. The results of the KNN-based AVF method are given in Table 2, from Table 2 we can see the average accuracy is 93.9%, the average detection time of 27.6ms. The results of the SVM-based AVF method are given in Table 3, from Table 3 we can see the average accuracy is 96.23%, the average detection time of 25.7ms. The results of the RF-based AVF method are given in Table 4, from Table 4 we can see the average accuracy is 95.63%, the average detection time of 29.99ms. The overall result is shown in Fig.11, from Fig.11, we can see the accuracy rate of the AVF method based on SVM is 2.23% higher than that based on KNN. The detection speed is 1.9ms faster, and the accuracy of detection is 0.6% higher than that based on RF. The detection speed is 4.92ms faster, SVM has advantages in detection accuracy and detection time compared with RF and KNN.

Comparing the image detection method with the AVF method are given in Table 5, from Table 5 we can see the overall detection accuracy of AVF method is 96.23%, while

that of image detection method is 92.25%. Therefore, the method of AVF is better than the method of image detection. As shown in the Fig.12, AVF method for all kinds of conveyor belt with damage detection accuracy higher than that of image detection especially for conveyor belt scratch detection, AVF detection method average accuracy is 93.66%, and the image detection method used alone is 87.16%. This shows that AVF detection method has higher accuracy in scratch detection, Due to the precursor of the longitudinal tear of the conveyor belt when the conveyor belt is scratched, it is very important for the safety production of the coal mine to detect the conveyor belt scratches, the timely detection of the conveyor belt scratches can reduce the occurrence of longitudinal tearing of the conveyor belt.

In order to better evaluate our method, and Compare with Qiao's IBVD method and Gong Xian Wang' method We calculated the accuracy, recall rate and FPR (false positive rate) of the tear sample and normal sample.

$$recall = \frac{TP}{TP + FN} \times 100\% \tag{26}$$

$$FPR = \frac{FP}{FP + TN} \times 100\% \tag{27}$$

$$acuracy = \frac{TP + TN}{TP + TN + FP + FN} \times 100\% \tag{28}$$

Where TP is the number of tear samples detected as tears, FN is the number of normal samples detected as tears, FP is the number of tear samples detected as normal, and TN is the number of normal samples detected as normal, the results are shown in Table 6, compared with the IBVD method, the AVF method has better accuracy, recall, and FPR. Compared with the Gong Xianwang'method, both have high detection accuracy, but the Gong Xianwang'method processing time is 96.5ms The AVF method processing time is 25.7ms, so the AVF method is more in line with the needs of real-time detection of coal mines.

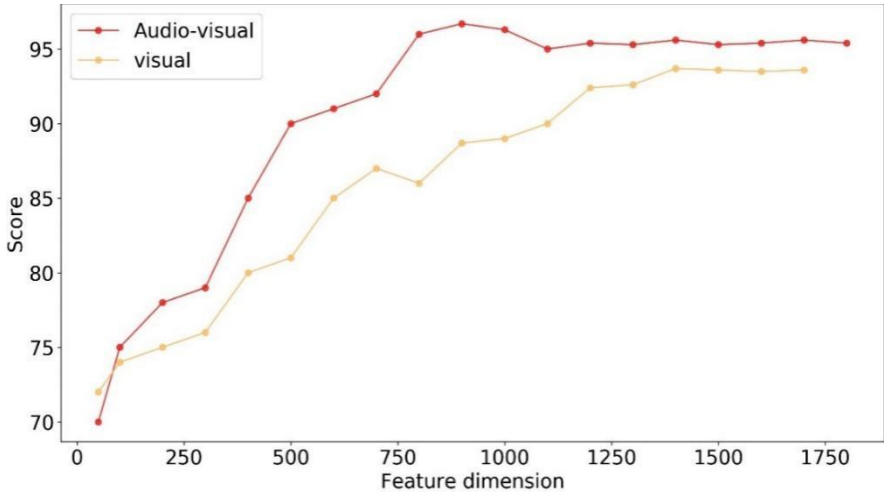


Fig.10. Score by applying PCA

Table 2

Detection results of AVF based on KNN

Experiment	Tear sample			Scratch sample			Normal sample			Accuray	Average detection time (ms)
	tear	scratch	normal	tear	scratch	normal	tear	scratch	normal		
1	190	8	2	17	182	1	1	9	190	93.6%	27.89
2	188	10	2	15	179	6	2	5	193	93.3%	26.90
3	193	5	2	10	184	6	1	7	192	94.8%	28 .24

Table 3

Detection results of AVF based on SVM

Experiment	Tear sample			Scratch sample			Normal sample			Accuray	Average detection time (ms)
	tear	scratch	normal	tear	scratch	normal	tear	scratch	normal		
1	191	8	1	13	185	2	1	2	197	95.5%	24.89
2	192	6	2	10	187	3	0	2	198	96.1%	25.79
3	193	6	1	9	190	1	0	0	200	97.1%	26.44

Table 4

Detection results of AVF based on RF

Experiment	Tear sample			Scratch sample			Normal sample			Accuray	Average detection time (ms)
	tear	scratch	normal	tear	scratch	normal	tear	scratch	normal		
1	192	7	1	13	184	3	2	2	196	95.3%	30.88
2	190	5	5	11	188	1	0	2	198	96.0%	29.47
3	192	6	2	10	187	3	0	5	195	95.6%	29.64

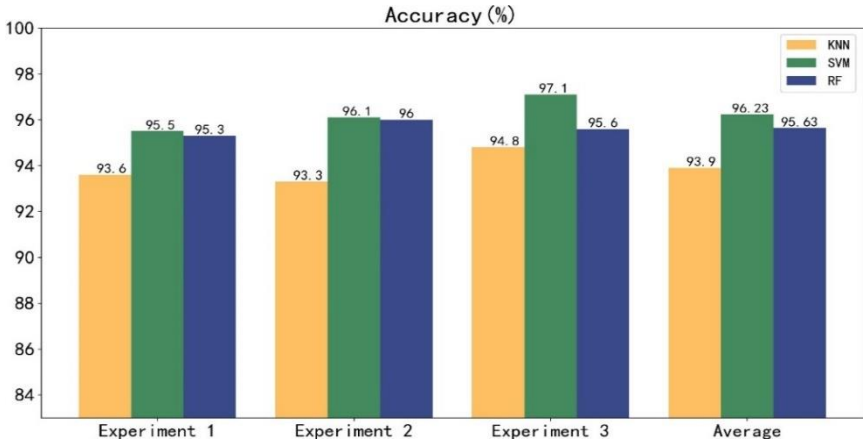


Fig.11. Accuracy comparison of three machine learning models.

Table 5

The image detection results based on SVM

Experiment	Tear sample			Scratch sample			Normal sample			Accuray	Average detection time (ms)
	tear	scratch	normal	tear	scratch	normal	tear	scratch	normal		
1	187	12	1	15	179	6	2	4	194	93.3%	22.88
2	188	10	2	20	174	6	0	5	195	92.8%	23.47
3	184	10	6	25	170	5	0	4	196	91.6%	24.64

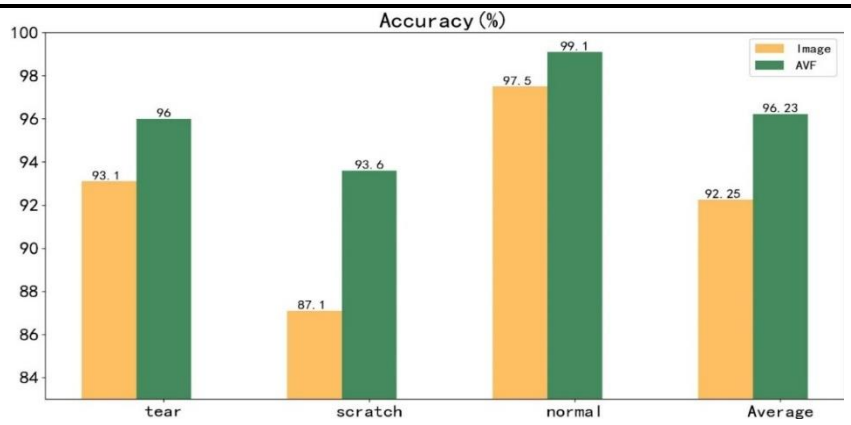


Fig. 12. Accuracy comparison between the AVF method and the image method

Table 6

Comparison of AVF method, IBVD method and Gong ian Wang method

Methods	recall	accuracy	FPR	Processing time (ms)
AVF	98.96%	96.23%	3.8%	25.7
Gong XianWang'method	93%	95.5%	5%	96.5
IBVD	90.16%	86.75%	12.07%	26.7

4. Conclusions

In this paper, we investigated a conveyor belt tear detection method based on sound and visual features, in order to adapt to the complex environment in coal mines and improve the detection accuracy. Visible light CCD is used to collect images of different running states of the conveyor belt, while the microphone array is used to collect sounds corresponding to different images. By processing and analyzing the collected images and sounds, we extracted normal, tear and scratch image features and sound features respectively. Then the extracted image and sound features fused and the machine learning algorithm is used for training and classification. The experimental findings the overall accuracy of AVF method is above 96.23%, and the recall rate is 98.96%, and FPR is 3.8% and the detection accuracy of scratches on conveyor belt is 93.6%. In terms of longitudinal tear detection Compared with the IBVD method and Gong Xian wang'method, the AVF method has better accuracy, recall and FPR. The average detection time of the AVF method is 25.7ms. In the next step of research, we can use deep learning to automatically extract the features of sound and images, and further improve the detection accuracy and generalization of the AVF method and we can also use better experimental platforms and equipment to improve the real-time performance of our algorithms. To sum up, AVF method can not only adapt to the complex environment under coal mine, but also to better achieve early warning of longitudinal tear.

Acknowledgements

This work is supported by the National Natural Science Foundation of China-Shanxi coal-based low-carbon joint fund (Grant No. U1810121) ； Funds for Local Scientific and Technological Development under the Guidance of the Central Government(Grant No.YDZX20201400001796)

References

[1] D. He, Y. Pang, G. Lodewijks, Green operations of belt conveyors by means of speed control, Applied Energy, 188 (2017) 330-341.

[2] M. Andrejiova, A. Grincova, D. Marasova, ANALYSIS OF TENSILE PROPERTIES OF WORN FABRIC CONVEYOR BELTS WITH RENOVATED COVER AND WITH THE DIFFERENT CARCASS TYPE, Eksploatacja I Niezawodnosc-Maintenance and Reliability, 22 (2020) 472-481.

[3] G. Fedorko, V. Molnar, D. Marasova, A. Grincova, M. Dovica, J. Zivcak, T. Toth, N. Husakova, Failure analysis of belt conveyor damage caused by the falling material. Part I: Experimental measurements and regression models, Engineering Failure Analysis, 36 (2014) 30-38.

[4] T. Qiao, Y. Duan, B. Jin, Infrared spectra imaging mechanism and modelling of the transport of hazard belt, Materials Research Innovations, 19 (2015) 92-97.

[5] M. Kuntoglu, H. Saglam, Investigation of progressive tool wear for determining of optimized machining parameters in turning, Measurement, 140 (2019) 427-436.

[6] A. Aslan, Optimization and analysis of process parameters for flank wear, cutting forces and vibration in turning of AISI 5140: A comprehensive study, Measurement, 163 (2020).
[7] M. Barburski, Analysis of the mechanical properties of conveyor belts on the three main stages of production, Journal of Industrial Textiles, 45 (2016) 1322-1334.
[8] T. Qiao, X. Li, Y. Pang, Y. Lu, F. Wang, B. Jin, Research on conditional characteristics vision real-time detection system for conveyor belt longitudinal tear, Iet Science Measurement & Technology, 11 (2017) 955-960.
[9] Y. Yang, C. Miao, X. Li, X. Mei, On-line conveyor belts inspection based on machine vision, Optik, 125 (2014) 5803-5807.
[10] C. Wang, J. Zhang, The research on the monitoring system for conveyor belt based on pattern recognition, in: J.H. Wu, M. Zhao, B. Wu (Eds.) Intelligent System and Applied Material, Pts 1 and 22012, pp. 622-625.
[11] G. Wang, L. Zhang, H. Sun, C. Zhu, Longitudinal tear detection of conveyor belt under uneven light based on Haar-AdaBoost and Cascade algorithm, Measurement, 168 (2021).
[12] T. Qiao, L. Chen, Y. Pang, G. Yan, C. Miao, Integrative binocular vision detection method based on infrared and visible light fusion for conveyor belts longitudinal tear, Measurement, 110 (2017) 192-201.
[13] Y. Yang, C. Hou, T. Qiao, H. Zhang, L. Ma, Longitudinal tear early-warning method for conveyor belt based on infrared vision, Measurement, 147 (2019).
[14] R. Yang, T. Qiao, Y. Pang, Y. Yang, H. Zhang, G. Yan, Infrared spectrum analysis method for detection and early warning of longitudinal tear of mine conveyor belt, Measurement, 165 (2020).
[15] A. Qayyum, S.M. Anwar, M. Awais, M. Majid, Medical image retrieval using deep convolutional neural network, Neurocomputing, 266 (2017) 8-20.
[16] D. Kumar, P. Carvalho, M. Antunes, R.P. Paiva, J. Henriques, Noise detection during heart sound recording using periodicity signatures, Physiological Measurement, 32 (2011) 599-618.
[17] Y. Kato, T. Hayashi, T. Kitagawa, Detection of extraneous abnormal sounds affecting road traffic noise by use of a necessary condition method, Applied Acoustics, 67 (2006) 1009-1021.
[18] X. Hu, X. Ye, D. Zhang, L. Wu, Vehicle Detection Technology Based on Cascading Classifiers of Multi-Feature Integration, International Journal of Pattern Recognition and Artificial Intelligence, 31 (2017).
[19] M. Guerrieri, G. Parla, C. Celauro, Digital image analysis technique for measuring railway track defects and ballast gradation, Measurement, 113 (2018) 137-147.
[20] H. Kim, E. Ahn, S. Cho, M. Shin, S.-H. Sim, Comparative analysis of image binarization methods for crack identification in concrete structures, Cement and Concrete Research, 99 (2017) 53-61.
[21] S. Zhang, S. Zhang, T. Huang, W. Gao, Q. Tian, Learning Affective Features With a Hybrid Deep Model for Audio–Visual Emotion Recognition, IEEE Transactions on Circuits and Systems for Video Technology, 28 (2018) 3030-3043.
[22] L. Yang, L. Tao, X. Chen, X. Gu, Multi-scale semantic feature fusion and data augmentation for acoustic scene classification, Applied Acoustics, 163 (2020).
[23] Y. Li, D.K. Jha, A. Ray, T.A. Wettergren, Information Fusion of Passive Sensors for Detection of Moving Targets in Dynamic Environments, Ieee Transactions on Cybernetics, 47 (2017) 93-104.
[24] A.J. Eronen, V.T. Peltonen, J.T. Tuomi, A.P. Klapuri, S. Fagerlund, T. Sorsa, G. Lorho, J. Huopaniemi, Audio-based context recognition, Ieee Transactions on Audio Speech and Language Processing, 14 (2006) 321-329.
[25] A. Grincova, M. Andrejiova, D. Marasova, S. Khouri, Measurement and determination of the absorbed impact energy for conveyor belts of various structures under impact loading, Measurement,

131 (2019) 362-371.
[26] M. Storath, A. Weinmann, Fast Median Filtering for Phase or Orientation Data, Ieee Transactions on Pattern Analysis and Machine Intelligence, 40 (2018) 639-652.
[27] N. Dalal, B. Triggs, Histograms of oriented gradients for human detection, in: C. Schmid, S. Soatto, C. Tomasi (Eds.) 2005 Ieee Computer Society Conference on Computer Vision and Pattern Recognition, Vol 1, Proceedings2005, pp. 886-893.
[28] X. Zhou, J. Wang, H. Hu, W. Dai, L. Wei, H. Mao, Ieee, Recognition of Infant's Emotions and Needs from Speech Signals, 2016 Ieee International Conference on Systems, Man, and Cybernetics2016, pp. 4620-4625.
[29] A. Maurya, D. Kumar, R.K. Agarwal, Speaker Recognition for Hindi Speech Signal using MFCC-GMM Approach, in: J. Mathew, A.K. Singh (Eds.) 6th International Conference on Smart Computing and Communications2018, pp. 880-887.
[30] X. Zeng, Q. Wang, C. Zhang, H. Cai, Ieee, Feature Selection Based on ReliefF and PCA for Underwater Sound Classification, 2013.
[31] K. Hwang, S.-Y. Lee, Environmental Audio Scene and Activity Recognition through Mobile-based Crowdsourcing, Ieee Transactions on Consumer Electronics, 58 (2012) 700-705.
[32] C.-W. Hsu, C.-J. Lin, A comparison of methods for multiclass support vector machines, IEEE transactions on neural networks, 13 (2002) 415-425.
[33] V. Svetnik, A. Liaw, C. Tong, J.C. Culberson, R.P. Sheridan, B.P. Feuston, Random forest: a classification and regression tool for compound classification and QSAR modeling, Journal of chemical information and computer sciences, 43 (2003) 1947-1958.