

FRAUD IDENTIFICATION METHOD AND DEVICE COMBINING AUDIO ANALYSIS AND VIDEO ANALYSIS AND STORAGE MEDIUM

This application claims priority to the Chinese Patent Application No. 201711252009.1, filed with the National Intellectual Property Administration, PRC on December 1, 2017, and entitled
5 "FRAUD IDENTIFICATION METHOD AND DEVICE COMBINING AUDIO ANALYSIS AND VIDEO ANALYSIS AND STORAGE MEDIUM", which is incorporated herein by reference in its entirety.

TECHNICAL FIELD

This application relates to the field of computer information processing technologies, and in
10 particular, to a fraud identification method and device which combine audio analysis and video analysis and a storage medium.

BACKGROUND

At present, fraud identification is generally achieved through face-to-face review, which relies heavily on the experience and judgment of analysts. It takes a lot of time and manpower, and analysis
15 results are often inaccurate and not objective. A professional instrument may also be utilized to determine whether a tested person is suspected of fraud by detecting a series of indicators such as breath, pulse, blood pressure, and skin resistance. However, such instruments are usually expensive and easily infringe on the human rights of the tested person.

SUMMARY

20 In order to solve the deficiencies of the prior art, this application provides a fraud identification method and device which combine audio analysis and video analysis and a computer readable storage medium. By analyzing audio/video data of an object to be identified, whether the object to be identified is suspected of fraud can be objectively and accurately determined.

To achieve the foregoing objective, this application provides a fraud identification method
25 combining audio analysis and video analysis, which is applied to an electronic device, and the method includes:

a sample preparation step: acquiring an audio/video sample of a person, cutting the audio/video

sample to obtain audio/video clips, assigning a fraud label for each audio/video clip, and decoding and preprocessing each audio/video clip to obtain an audio clip and a video clip for each audio/video clip;

5 a feature extraction step: extracting a voice feature from each audio clip, and extracting an emotional expression feature from each video clip;

a model training step: training a first support vector machine by using the voice feature of each audio clip and the fraud label as sample data to obtain a voice analysis model, and training a second support vector machine by using the emotional expression feature of each video clip and the fraud label as sample data to obtain an emotional expression analysis model;

10 a model application step: acquiring audio/video data of an object to be identified; utilizing the voice analysis model and the emotional expression analysis model to analyze the audio/video data, and outputting an audio fraud probability P1 and a video fraud probability P2 of the object to be identified; and

15 a weighted calculation step: calculating the weight of P1 and P2 according to the weight of the voice analysis model and the emotional expression analysis model to obtain a fraud identification result of the object to be identified.

This application also provides an electronic device including a memory and a processor, and the memory includes a fraud identification program. The electronic device is directly or indirectly connected to a photographing device, and the photographing device transmits acquired audio/video data to the electronic device. When the processor of the electronic device executes the fraud identification program in the memory, the following steps are implemented:

20 a sample preparation step: acquiring an audio/video sample of a person, cutting the audio/video sample to obtain audio/video clips, assigning a fraud label for each audio/video clip, and decoding and preprocessing each audio/video clip to obtain an audio clip and a video clip for each audio/video clip;

a feature extraction step: extracting a voice feature from each audio clip, and extracting an emotional expression feature from each video clip;

a model training step: training a first support vector machine by using the voice feature of each

audio clip and the fraud label as sample data to obtain a voice analysis model, and training a second support vector machine by using the emotional expression feature of each video clip and the fraud label as sample data to obtain an emotional expression analysis model;

a model application step: acquiring audio/video data of an object to be identified, utilizing the voice analysis model and the emotional expression analysis model to analyze the audio/video data, and outputting an audio fraud probability P1 and a video fraud probability P2 of the object to be identified; and

a weighted calculation step: calculating the weight of P1 and P2 according to the weight of the voice analysis model and the emotional expression analysis model to obtain a fraud identification result of the object to be identified.

In addition, in order to achieve the foregoing objective, this application also provides a computer readable storage medium including a fraud identification program, and when the fraud identification program is executed by a processor, any step in the fraud identification method combining audio analysis and video analysis described above is implemented.

According to the fraud identification method and device combining audio analysis and video analysis and the computer readable storage medium provided by this application, a support vector machine is trained by using a voice feature of an audio clip and an emotional expression feature of a video clip of an audio/video sample in combination with a fraud label so as to obtain a voice analysis model and an emotional expression analysis model. Then, the trained models are applied to real-time fraud identification links: acquiring audio/video data of an object to be identified, extracting a voice feature and an emotional expression feature of the audio/video data, respectively inputting the voice features and the emotional expression features into the voice analysis model and the emotional expression analysis model which are obtained by training, outputting an audio fraud probability P1 and a video fraud probability P2 of the object to be identified, and performing weighted fusion on P1 and P2 to obtain a fraud identification result of the object to be identified. With this application, whether a person is suspected of fraud can be objectively and accurately identified.

BRIEF DESCRIPTION OF DRAWINGS

FIG. 1 is an application environment diagram of an electronic device according to a first

preferred embodiment of this application;

FIG. 2 is an application environment diagram of an electronic device according to a second preferred embodiment of this application;

FIG. 3 is a program module diagram of a fraud identification program described in FIG. 1 and
5 FIG. 2; and

FIG. 4 is a flowchart of a fraud identification method which combines audio analysis and video analysis according to a preferred embodiment of this application.

The implementation, functional features and advantages of this application will be further described by combining with embodiments with reference to the accompanying drawings.

10 DESCRIPTION OF EMBODIMENTS

The principles and spirit of this application will be described below with reference to several specific embodiments. It should be understood that the specific embodiments described herein are merely illustrative of this application and are not intended to limit this application.

Referring to FIG. 1, an application environment diagram of an electronic device according to a
15 first preferred embodiment of this application is illustrated. In this embodiment, a photographing device 3 is connected to an electronic device 1 via a network 2, the photographing device 3 acquires audio/video data of a person and transmits the audio/video data to the electronic device 1 via the network 2, and the electronic device 1 analyzes the audio/video data using a fraud identification program 10 provided by this application to obtain a fraud identification result of the person.

20 The electronic device 1 may be a terminal device having storage and computing functions, such as a server, a smart phone, a tablet computer, a portable computer, and a desktop computer.

The electronic device 1 includes a memory 11, a processor 12, a network interface 13, and a communication bus 14. The photographing device 3 is installed in a specific place, such as an office and a monitoring area, and configured to acquire audio/audit data of a person and then transmit the
25 audio/audit data to the memory 11 via the network 2. The network interface 13 may include a standard wired interface and a wireless interface (such as a Wi-Fi interface). The communication bus 14 is configured to implement connection communication between these components.

The memory 11 includes at least one type of readable storage medium. The at least one type of

readable storage medium may be a non-volatile storage medium such as a flash memory, a hard disk, a multimedia card, or a card type memory. In some embodiments, the readable storage medium may be an internal storage unit of the electronic device 1, such as a hard disk of the electronic device 1. In some other embodiments, the readable storage medium may also be an external memory 11 of the electronic device 1, such as a plug-in hard disk disposed on the electronic device 1, a Smart Media Card (SMC), a Secure Digital (SD) card, or a flash card.

In this embodiment, the memory 11 stores a program code of the fraud identification program 10, audio/video data acquired by the photographing device 3, other data applied when the processor 12 executes the program code of the fraud identification program 10, data output finally, and the like.

10 In some embodiments, the processor 12 may be a Central Processing Unit (CPU), a microprocessor or other data processing chips.

FIG. 1 shows only the electronic device 1 with components 11-14, but it should be understood that it is not required to implement all illustrated components, and more or fewer components may be implemented instead.

15 Optionally, the electronic device 1 may also include a user interface, and the user interface may include an input unit such as a keyboard, a voice input device such as a microphone that has a voice identification function, and a voice output device such as a loudspeaker box or a headset. Optionally, the user interface may also include a standard wired interface and a wireless interface.

20 Optionally, the electronic device 1 may further include a display. In some embodiments, the display may be an LED display, a liquid crystal display, a touch liquid crystal display, an Organic Light-Emitting Diode (OLED) touch sensor or the like. The display is configured to display information processed by the electronic device 1 and a visualized user interface.

25 Optionally, the electronic device 1 may further include a touch sensor. An area provided by the touch sensor for a user to perform a touch operation is referred to as a touch area. In addition, the touch sensor described herein may be a resistive touch sensor, a capacitive touch sensor, or the like. Moreover, the touch sensor includes not only a contact type touch sensor but also a proximity type touch sensor or the like. In addition, the touch sensor may be a single sensor or a plurality of sensors arranged in an array, for example. A user, such as a psychological consultant or a credit interviewing

person, can start the fraud identification program 10 by touch.

The electronic device 1 may further include a Radio Frequency (RF) circuit, a sensor, an audio circuit, and the like, and details are not described herein.

Referring to FIG. 2, an application environment diagram of an electronic device according to a second preferred embodiment of this application is illustrated. The user implements the fraud identification process through the terminal 3, and the photographing device 30 of the terminal 3 acquires audio/video data of the object to be identified, and transmits the audio/video data to the electronic device 1 via the network 2, and the processor 12 of the electronic device 1 implements the program code of the fraud identification program 10 stored in the memory 11. Audio data and video data of the audio/video data are analyzed, an audio fraud probability P1 and a video fraud probability P2 of the object to be identified are output, the weight of P1 and P2 is calculated to obtain a fraud identification result of the object to be identified, for reference by the object to be identified, an examiner, or the like.

A reference is made to the introduction of FIG. 1 for the components of the electronic device 1 in FIG. 2, such as the memory 11, the processor 12, the network interface 13 and the communication bus 14 shown in the drawing, as well as components not shown.

The terminal 3 may be a terminal device having storage and computing functions, such as a smart phone, a tablet computer, a portable computer, and a desktop computer.

When the fraud identification program 10 in FIG. 1 and FIG. 2 is executed by the processor 12, the following steps are implemented:

a sample preparation step: acquire an audio/video sample of a person, cut the audio/video sample to obtain audio/video clips, assign a fraud label for each audio/video clip, and decode and preprocess each audio/video clip to obtain an audio clip and a video clip for each audio/video clip;

a feature extracting step: extract a voice feature from each audio clip, and extract an emotional expression feature from each video clip;

a model training step: train a first support vector machine by using the voice feature of each audio clip and the fraud label as sample data to obtain a voice analysis model; and train a second support vector machine by using the emotional expression feature of each video clip and the fraud

label as sample data to obtain an emotional expression analysis model;

a model application step: acquire audio/video data of an object to be identified; utilize the voice analysis model and the emotional expression analysis model to analyze the audio/video data, and output an audio fraud probability P1 and a video fraud probability P2 of the object to be identified;

5 and

a weighted calculation step: calculating the weight of P1 and P2 according to the weight of the voice analysis model and the emotional expression analysis model to obtain a fraud identification result of the object to be identified.

For a detailed description of the foregoing steps, please refer to the following description of a program module diagram of a fraud identification program 10 in FIG. 3 and a flowchart of a fraud identification method which combines audio analysis and video analysis according to a preferred embodiment in FIG. 4.

Referring to FIG. 3, a program module diagram of a fraud identification program 10 in FIG. 1 and FIG. 2 is illustrated. In this embodiment, the fraud identification program 10 is cut into a plurality of modules, and the plurality of modules are stored in the memory 11 and executed by the processor 12 to complete this application. A module as referred to in this application refers to a series of computer program instruction segments that can perform particular functions.

The fraud identification program 10 may be divided into an acquisition module 110, an extraction module 120, a training module 130, a model application module 140, and a weighted calculation module 150.

The acquisition module 110 is configured to acquire an audio/video of a person and decode and preprocess the same to obtain a corresponding audio part and a corresponding video part. The audio/video may be acquired by a photographing device 3 in FIG. 1 or a photographing device 30 in FIG. 2, or may be an audio/video that has an obvious fraudulent conduct selected from network information or an audio/video database, and an audio/video having no fraudulent conduct. An audio/video sample for training a support vector machine is cut in units of emotions to obtain audio/video clips, and a fraud label is assigned for each audio/video clip. The fraud label indicates that a person in the audio and video clip has a fraud suspect. For example, 1 indicates fraud suspect,

and 0 indicates no fraud suspect. The audio/video is decoded and preprocessed to obtain a corresponding audio part and a corresponding video part.

The extraction module 120 is configured to extract a voice feature from the audio part and an emotional expression feature from the video part. The extraction module 120 extracts the voice
5 feature from each audio part obtained by the acquisition module 110, and extracts the emotional expression feature from the video part obtained by the acquisition module 110.

When the extraction module 120 extracts voice features from an audio part, low-order audio features such as a Mel frequency cepstrum coefficient, a pitch, and the zero-crossing rate are first extracted from the audio part, then a dynamic regression coefficient is extracted from these low-order
10 audio features to obtain dynamic audio features of the audio part, and then a statistical function is utilized to extract high-order audio features from the low-order audio features and the dynamic audio features, and finally a feature screening algorithm is used to screen out a high-order audio feature subset from the higher-order audio features, and the high-order audio feature subset is used as voice features of the audio part.

In this embodiment, the low-order audio features such as the Mel frequency cepstrum coefficient,
15 the pitch, and the zero-crossing rate can be extracted from the audio part by using OpenSMILE software. The dynamic regression coefficient is used to indicate the importance of low-order audio features. For example, if a low-order audio feature (such as a pitch feature) of an audio part is represented as a waveform file, the waveform file can be expressed in a multiple linear regression
20 mode as:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_K X_K$$

where k is the number of the lower-order audio features in the audio part, and β_j ($j=1,2,\cdots,k$) is a dynamic regression coefficient of the low-order audio features.

The statistical function includes functions for extracting a maximum value, a minimum value, a
25 kurtosis, the skewness and the like of low-order audio features and dynamic audio features, and the extraction module 120 combines and transforms data extracted by using the statistical function to obtain high-order audio features. The number of high-order audio features extracted from each audio part is often very large, but usually only a small number of high-order audio features have a

significant impact on the results of fraud identification, so a feature screening algorithm is used to reduce the number of high-order audio features, and the fraud identification speed is improved. In this embodiment, the feature screening algorithm may be a Sequential Forward Selection (SFS) algorithm, a Sequential Backward Selection (SBS) algorithm, a Bidirectional Search (BDS) algorithm, or a filter
5 feature selection algorithm, and may also be other feature screening algorithms.

Similarly, when the extraction module 120 extracts an emotional expression feature from a video part, low-order action features such as a head orientation, an eye orientation and a facial Action Unit (AU) are first extracted from the video part, and then the number of occurrences of each low-order action feature in the video part and the duration are counted, high-order action features of
10 the video part are constructed according to statistical results, and then a feature screening algorithm is used to screen out a high-order action feature subset from the high-order action features, and the high-order action feature subset is used as an emotional expression feature of the video part.

The training module 130 is configured to train a support vector machine to obtain a voice analysis model and an emotional expression analysis model. A first support vector machine is trained
15 by using the voice feature of each audio part of the audio/video sample extracted by the extraction module 120 and a fraud label assigned by the acquisition module 110 as sample data, to obtain a voice analysis model. A second support vector machine is trained by using the voice feature of each audio part of the audio/video sample extracted by the extraction module 120 and a fraud label assigned by the acquisition module 110 as sample data, to obtain the emotional expression analysis model.

20 The model application module 140 is configured to analyze audio/video data of an object to be identified, and obtain an audio fraud probability and a video fraud probability of the object to be identified. The voice feature of the audio part of the audio and video of the object to be identified, which is extracted by the extraction module 120, is input into the voice analysis model trained by the training module 130, and the audio fraud probability P1 of the object to be identified is output. The
25 emotional expression feature of the video part of the object to be identified is input into the trained emotional expression analysis model, and the video fraud probability P2 of the object to be identified is output.

The weighted calculation module 150 is configured to calculate the weight of the audio fraud

probability P1 and the video fraud probability P2 of the object to be identified to obtain a fraud identification result of the object to be identified. When the training module 130 uses sample data to train a support vector machine to obtain the voice analysis model and the emotional expression analysis model, the precision of the two models can be calculated, thereby calculating the weight of the voice analysis model and the emotional expression analysis model, and calculating a final fraud probability of the object to be identified.

For example, assuming that the precision of the voice analysis model is 85% and the precision of the emotional expression analysis model is 95%, the calculation of the weight of the voice analysis model and the emotional expression analysis model can be expressed as follows:

$$P(\text{Audio}) = a = 0.85$$

$$P(\text{Video}) = b = 0.95$$

$$W(\text{Audio}) = \frac{a}{a+b} = (0.85/1.8)$$

$$W(\text{Video}) = \frac{b}{a+b} = (0.95/1.8)$$

where P(Audio) represents the precision of the voice analysis model, P(Video) represents the precision of the emotional expression analysis model, W(Audio) represents the weight of the voice analysis model, and W(Video) represents the weight of the emotional expression analysis model.

Assuming that the audio/video data of the object to be identified is analyzed by the voice analysis model and the emotional expression analysis model to obtain that the audio fraud probability of the object to be identified is 0.8 and the video fraud probability is 0.7, then the weighted fusion calculation is performed according to the W(Audio) and the W(Video), and it is finally obtained that the fraud probability of the object to be identified is:

$$P = (0.85/1.8) * 0.8 + (0.95/1.8) * 0.7.$$

Referring to FIG. 4, a flowchart of a fraud identification method which combines audio analysis and video analysis according to a preferred embodiment of this application is illustrated. The electronic device 1 is started by using a framework shown in FIG. 1 or 2, and the processor 12 executes the fraud identification program 10 stored in the memory 11 to implement the following steps.

Step S10: Use an acquisition module 110 to acquire an audio/video sample of a person, cut the audio/video sample in units of emotions to obtain audio/video clips, and assign a fraud label for each

audio/video clip. The audio/video sample may be acquired by a photographing device 3 in FIG. 1 or a photographing device 30 in FIG. 2, or may be an audio/video that has an obvious fraudulent conduct selected from network information or an audio/video database, and an audio/video having no fraudulent conduct.

5 Step S20: Use the acquisition module 110 to decode and preprocess each audio/video clip to obtain an audio clip and a video clip for each audio/video clip. A fraud label of each audio/video clip is still used as a fraud label for the corresponding audio clip and video clip.

10 Step S30: Use an extraction module 120 to extract a voice feature and an emotional expression feature from each audio clip and video clip respectively. For a specific method for extracting a voice feature and an emotional expression feature, refer to the detailed description of the extraction module 120 described above.

15 Step S40: Train a first support vector machine according to the voice feature of each audio clip and the fraud label to obtain a voice analysis model, and train a second support vector machine according to the emotional expression feature of each video clip and the fraud label to obtain an emotional expression analysis model. By using a training module 130, a first support vector machine is trained by using the voice feature of each audio clip and the fraud label as sample data to obtain a voice analysis model, and a second support vector machine is trained by using the emotional expression feature of each video clip and the fraud label as sample data to obtain an emotional expression analysis model.

20 Step S50: the acquisition module 110 is utilized to acquire audio/video data of an object to be identified, and decode and preprocess the audio/video data to obtain audio data and video data of the object to be identified. The audio/video data is acquired by the photographing device 3 in FIG. 1 or the photographing device 30 in FIG. 2 in real time.

25 Step S60: Utilize the extraction module 120 to extract a voice feature from the audio data and an emotional expression feature from the video data of the object to be identified. For a specific method for extracting a voice feature and an emotional expression feature, refer to the detailed description of the extraction module 120 described above.

Step S70: Input the voice features from the audio data and the emotional expression features

from the video data of the object to be identified into the voice analysis model and the emotional expression analysis model respectively, to obtain an audio fraud probability and a video fraud probability of the object to be identified. By using a model application module 140, the voice features of the audio data of the object to be identified, which are extracted by the extraction module 120, are input into the voice analysis model, and the audio fraud probability P1 of the object to be analyzed is output; the emotional expression features of the video data of the object to be identified, which are extracted by the extraction module 120, are input into the emotional expression analysis model, and the video fraud probability P2 of the object to be analyzed is output.

Step S80: Calculate the weight of P1 and P2 according to the weight of the voice analysis model and the emotional expression analysis model to obtain a fraud identification result of the object to be identified. For a method for determining the weight of the voice analysis model and the emotional expression analysis model and the specific process for calculating the weight of P1 and P2, refer to the detailed description of the weighted calculation module 150 described above.

Moreover, an embodiment of this application also provides a computer readable storage medium, which may be any one or any combination of a hard disk, a multimedia card, an SD card, a flash card, an SMC, a Read-Only Memory (ROM), an Erasable Programmable Read-Only Memory (EPROM), a portable Compact Disc Read-Only Memory (CD-ROM), a USB memory, and the like. The computer readable storage medium includes an audio/video sample and a fraud identification program 10, and when the fraud identification program 10 is executed by a processor, the following operations are implemented:

a sample preparation step: acquire an audio/video sample of a person, cut the audio/video sample to obtain audio/video clips, assign a fraud label for each audio/video clip, and decode and preprocess each audio/video clip to obtain an audio clip and a video clip for each audio/video clip;

a feature extraction step: extract a voice feature from each audio clip, and extract an emotional expression feature from each video clip;

a model training step: train a first support vector machine by using the voice feature of each audio clip and the fraud label as sample data to obtain a voice analysis model; train a second support vector machine by using the emotional expression feature of each video clip and the fraud label as

sample data to obtain an emotional expression analysis model;

a model application step: acquire audio/video data of an object to be identified; utilize the voice analysis model and the emotional expression analysis model to analyze the audio/video data, and output an audio fraud probability P1 and a video fraud probability P2 of the object to be identified;
5 and

a weighted calculation step: calculate the weight of P1 and P2 according to the weight of the voice analysis model and the emotional expression analysis model to obtain a fraud identification result of the object to be identified.

The specific implementation of the computer readable storage medium of this application is
10 substantially the same as the specific implementation of the foregoing fraud identification method which combines audio analysis and video analysis and the electronic device 1, and details are not described herein again.

It should be noted that the term "comprise", "include", or any other variant thereof is intended to encompass a non-exclusive inclusion, such that a process, device, article, or method that includes a
15 series of elements including not only those elements, but also other elements not explicitly listed, or elements that are inherent to such a process, device, article, or method. Without more restrictions, an element defined by the phrase "including a ..." does not exclude the presence of another same element in a process, device, article, or method that includes the element.

The serial numbers of the embodiments of this application described above are merely for
20 description, and do not represent the advantages and disadvantages of the embodiments. Through the description of the foregoing embodiments, those skilled in the art can clearly understand that the method of the foregoing embodiments can be implemented by means of software plus a necessary universal hardware platform, and certainly, can also be implemented through hardware, but in many cases, the former is a better implementation. Based on such understanding, the technical solution of
25 this application which is essential or a part contributing to the prior art may be embodied in the form of a software product, the computer software product is stored in a storage medium described above (such as an ROM/RAM, a magnetic disk or an optical disc), including a plurality of instructions for causing a terminal device (which may be a mobile phone, a computer, a server, or a network device)

to perform the methods described in the embodiments of this application.

The above are only preferred embodiments of this application, and are not intended to limit the patent scope of this application. Any equivalent structure or equivalent process transformation performed using the specification and the accompanying drawings of this application may be directly or indirectly applied to other related technical fields and similarly falls within the patent protection scope of this application.

5

CLAIMS

1. A fraud identification method combining audio analysis and video analysis applied to an electronic device, comprising:

a sample preparation step: acquiring an audio/video sample of a person, cutting the audio/video sample to obtain audio/video clips, assigning a fraud label for each audio/video clip, and decoding and preprocessing each audio/video clip to obtain an audio clip and a video clip for each audio/video clip;

a feature extraction step: extracting a voice feature from each audio clip, and extracting an emotional expression feature from each video clip;

a model training step: training a first support vector machine by using the voice feature of each audio clip and the fraud label as sample data to obtain a voice analysis model, and training a second support vector machine by using the emotional expression feature of each video clip and the fraud label as sample data to obtain an emotional expression analysis model;

a model application step: acquiring audio/video data of an object to be identified, utilizing the voice analysis model and the emotional expression analysis model to analyze the audio/video data, and outputting an audio fraud probability P1 and a video fraud probability P2 of the object to be identified; and

a weighted calculation step: calculating the weight of P1 and P2 according to the weight of the voice analysis model and the emotional expression analysis model to obtain a fraud identification result of the object to be identified.

2. The fraud identification method according to claim 1, wherein in the feature extraction step, the extracting a voice feature comprises:

a first feature extraction step: extracting a low-order audio feature from each audio clip;

a second feature extraction step: extracting a dynamic regression coefficient from each low-order voice feature to obtain a dynamic audio feature of each audio clip;

a third feature extraction step: extracting high-order audio features of each audio clip from the low-order audio features and the dynamic audio features by using a statistical function; and

a screening step: using a feature screening algorithm to screen out a high-order audio feature subset from the high-order audio features of each audio clip, and using the high-order audio feature subset as voice features of each audio clip.

3. The fraud identification method according to claim 2, wherein the low-order audio features
5 comprise a Mel frequency cepstrum coefficient, a pitch, and the zero-crossing rate.

4. The fraud identification method according to claim 1, wherein in the feature extraction step, the extracting an emotional expression feature comprises:

a low-order feature extraction step: extracting a low-order action feature from each video clip;

a high-order feature construction step: collecting statistics about the number of occurrences of
10 each low-order action feature in each video clip and the duration, and constructing a high-order action feature of each video clip according to a statistical result; and

a screening step: using a feature screening algorithm to screen out a high-order action feature subset from the high-order action features of each video clip, and using the high-order action feature subset as emotional expression features of each video clip.

15 5. The fraud identification method according to claim 2, wherein in the feature extraction step, the extracting an emotional expression feature comprises:

a low-order feature extraction step: extracting a low-order action feature from each video clip;

a high-order feature construction step: collecting statistics about the number of occurrences of
each low-order action feature in each video clip and the duration, and constructing a high-order action
20 feature of each video clip according to a statistical result; and

a screening step: using a feature screening algorithm to screen out a high-order action feature subset from the high-order action features of each video clip, and using the high-order action feature subset as emotional expression features of each video clip.

6. The fraud identification method according to claim 4 or 5, wherein the low-order action
25 features comprise a head orientation, an eye orientation, and a facial Action Unit (AU).

7. The fraud identification method according to any one of claims 1 to 5, wherein the model application step further comprises the following steps:

decoding and preprocessing audio/video data of an object to be identified to obtain audio data and video data of the object to be identified; and

extracting voice features from the audio data of the object to be identified, and extracting emotional expression features from the video data of the object to be identified.

5 8. An electronic device, comprising a memory and a processor, wherein the memory comprises a fraud identification program, and when the fraud identification program is executed by the processor, the following steps are implemented:

10 a sample preparation step: acquiring an audio/video sample of a person, cutting the audio/video sample to obtain audio/video clips, assigning a fraud label for each audio/video clip, and decoding and preprocessing each audio/video clip to obtain an audio clip and a video clip for each audio/video clip;

 a feature extraction step: extracting a voice feature from each audio clip, and extracting an emotional expression feature from each video clip;

15 a model training step: training a first support vector machine by using the voice feature of each audio clip and the fraud label as sample data to obtain a voice analysis model, and training a second support vector machine by using the emotional expression feature of each video clip and the fraud label as sample data to obtain an emotional expression analysis model;

20 a model application step: acquiring audio/video data of an object to be identified, utilizing the voice analysis model and the emotional expression analysis model to analyze the audio/video data, and outputting an audio fraud probability P1 and a video fraud probability P2 of the object to be identified; and

 a weighted calculation step: calculating the weight of P1 and P2 according to the weight of the voice analysis model and the emotional expression analysis model to obtain a fraud identification result of the object to be identified.

25 9. The electronic device according to claim 8, wherein in the feature extraction step, the extracting a voice feature comprises:

 a first feature extraction step: extracting a low-order audio feature from each audio clip;

a second feature extraction step: extracting a dynamic regression coefficient from each low-order voice feature to obtain a dynamic audio feature of each audio clip;

a third feature extraction step: extracting high-order audio features of each audio clip from the low-order audio features and the dynamic audio features by using a statistical function; and

5 a screening step: using a feature screening algorithm to screen out a high-order audio feature subset from the high-order audio features of each audio clip, and using the high-order audio feature subset as voice features of each audio clip.

10 10. The electronic device according to claim 9, wherein the low-order audio features comprise a Mel frequency cepstrum coefficient, a pitch, and the zero-crossing rate.

11. The electronic device according to claim 8, wherein in the feature extraction step, the extracting an emotional expression feature comprises:

a low-order feature extraction step: extracting a low-order action feature from each video clip;

15 a high-order feature construction step: collecting statistics about the number of occurrences of each low-order action feature in each video clip and the duration, and constructing a high-order action feature of each video clip according to a statistical result; and

a screening step: using a feature screening algorithm to screen out a high-order action feature subset from the high-order action features of each video clip, and using the high-order action feature subset as emotional expression features of each video clip.

20 12. The electronic device according to claim 9, wherein in the feature extraction step, the extracting an emotional expression feature comprises:

a low-order feature extraction step: extracting a low-order action feature from each video clip;

a high-order feature construction step: collecting statistics about the number of occurrences of each low-order action feature in each video clip and the duration, and constructing a high-order action feature of each video clip according to a statistical result; and

25 a screening step: using a feature screening algorithm to screen out a high-order action feature subset from the high-order action features of each video clip, and using the high-order action feature subset as emotional expression features of each video clip.

13. The electronic device according to claim 11 or 12, wherein the low-order action features

comprise a head orientation, an eye orientation, and a facial Action Unit (AU).

14. The electronic device according to any one of claims 8 to 12, wherein the model application step further comprises the following steps:

decoding and preprocessing audio/video data of an object to be identified to obtain audio data
5 and video data of the object to be identified;

extracting voice features from the audio data of the object to be identified, and extracting emotional expression features from the video data of the object to be identified.

15. A computer readable storage medium, comprising a fraud identification program, wherein when the fraud identification program is executed by a processor, the following steps are
10 implemented:

a sample preparation step: acquiring an audio/video sample of a person, cutting the audio/video sample to obtain audio/video clips, assigning a fraud label for each audio/video clip, and decoding and preprocessing each audio/video clip to obtain an audio clip and a video clip for each audio/video clip;

15 a feature extraction step: extracting a voice feature from each audio clip, and extracting an emotional expression feature from each video clip;

a model training step: training a first support vector machine by using the voice feature of each audio clip and the fraud label as sample data to obtain a voice analysis model, and training a second support vector machine by using the emotional expression feature of each video clip and the fraud
20 label as sample data to obtain an emotional expression analysis model;

a model application step: acquiring audio/video data of an object to be identified, utilizing the voice analysis model and the emotional expression analysis model to analyze the audio/video data, and outputting an audio fraud probability P1 and a video fraud probability P2 of the object to be identified; and

25 a weighted calculation step: calculating the weight of P1 and P2 according to the weight of the voice analysis model and the emotional expression analysis model to obtain a fraud identification result of the object to be identified.

16. The computer readable storage medium according to claim 15, wherein in the feature

extraction step, the extracting a voice feature comprises the following steps:

a first feature extraction step: extracting a low-order audio feature from each audio clip;

a second feature extraction step: extracting a dynamic regression coefficient from each low-order voice feature to obtain a dynamic audio feature of each audio clip;

5 a third feature extraction step: extracting high-order audio features of each audio clip from the low-order audio features and the dynamic audio features by using a statistical function; and

a screening step: using a feature screening algorithm to screen out a high-order audio feature subset from the high-order audio features of each audio clip, and using the high-order audio feature subset as voice features of each audio clip.

10 17. The computer readable storage medium according to claim 16, wherein the low-order audio features comprise a Mel frequency cepstrum coefficient, a pitch, and the zero-crossing rate.

18. The computer readable storage medium according to claim 15 or 16, wherein in the feature extraction step, the extracting an emotional expression feature comprises the following steps:

a low-order feature extraction step: extracting a low-order action feature from each video clip;

15 a high-order feature construction step: collecting statistics about the number of occurrences of each low-order action feature in each video clip and the duration, and constructing a high-order action feature of each video clip according to a statistical result; and

20 a screening step: using a feature screening algorithm to screen out a high-order action feature subset from the high-order action features of each video clip, and using the high-order action feature subset as emotional expression features of each video clip.

19. The computer readable storage medium according to claim 18, wherein the low-order action features comprise a head orientation, an eye orientation, and a facial Action Unit (AU).

20. The computer readable storage medium according to claim 19, wherein the model application step further comprises the following steps:

25 decoding and preprocessing audio/video data of an object to be identified to obtain audio data and video data of the object to be identified; and

extracting voice features from the audio data of the object to be identified, and extracting emotional expression features from the video data of the object to be identified.