VIDEO QUALITY EVALUATION METHOD AND DEVICE, AND STORAGE MEDIUM

[0001] This application is a SG national stage of international application NO. PCT/CN2019/123007, filed on December 4, 2019, which claims priority to Chinese Patent Application No. 201811627024.4 filed on December 28, 2018 and entitled "VIDEO QUALITY EVALUATION METHOD, APPARATUS AND DEVICE, AND STORAGE MEDIUM". Both applications are herein incorporated by reference in their entireties.

TECHNICAL FIELD

[0002] Embodiments of this application relate to the field of video detection technologies, in particular relate to a video quality evaluation method and device, and a storage medium.

BACKGROUND

[0003] With the development of Internet technologies, short video applications (APP) are more and more widely used. Short video apps acquire a large quantity of short videos produced and uploaded by users every day. After offending videos are filtered out of these short videos, high-quality videos need to be selected from the remaining videos and pushed to viewers.

SUMMARY

[0004] This application provides a video quality evaluation method and device, and a storage medium, to improve timeliness of video quality evaluation and evaluation accuracy of a high-quality video.

[0005] An embodiment of this application provides a video quality evaluation method. The method includes:

[0006] acquiring an image sequence and audio information by decoding a to-be-evaluated video, wherein the to-be-evaluated video is non-offending video;

[0007] extracting an action feature vector and a face feature vector from the image sequence, and extracting an audio feature vector from the audio information;

[0008] constructing a video feature vector according to at least one of the action feature vector, the face feature vector and the audio feature vector;

[0009] inputting the video feature vector into a first-layer evaluation network, and acquiring, by the first-layer evaluation network, a primary feature vector through recursive and non-linear operations;

[0010] inputting the primary feature vector into a second-layer evaluation network, and acquiring, by the second-layer evaluation network, an advanced feature vector through recursive and non-linear operations; and

[0011] acquiring a quality score of the to-be-evaluated video by performing, by an output layer, operations on the primary feature vector and the advanced feature vector;

[0012] wherein constructing the video feature vector according to at least one of the action feature vector, the face feature vector and the audio feature vector comprises:

[0013] setting the feature vector that fails to be extracted among the action feature vector, the face feature vector and the audio feature vector to zero; and

[0014] constructing the video feature vector by combining the action feature vector, the face feature vector and the audio feature vector.

[0015] An embodiment of this application further provides a video quality evaluation device, including:

[0016] one or more processors; and

[0017] a memory storing at least one program therein, wherein

[0018] the at least one program, when run by the at least one processor, causes the at least one processor to perform the video quality evaluation method according to any embodiment of this application.

[0019] An embodiment of this application further provides a non-transitory computer-readable storage medium storing a computer program, wherein the computer program, when run by a processing apparatus, causes the processing apparatus to perform the video quality evaluation method according to any embodiment of this application.

BRIEF DESCRIPTION OF THE DRAWINGS

[0020] FIG. 1 is a flowchart of a video quality evaluation method according to Embodiment 1 of this application;

[0021] FIG. 2 is a schematic diagram of an application scenario of the video quality evaluation method according to Embodiment 1 of this application;

[0022] FIG. 3 is a flowchart of a video quality evaluation method according to Embodiment 2 of this application;

[0023] FIG. 4 is a flowchart of a video quality evaluation method according to Embodiment 3 of this application;

[0024] FIG. 5 is a schematic structural diagram of a video quality evaluation apparatus according to Embodiment 4 of this application; and

[0025] FIG. 6 is a schematic structural diagram of a video quality evaluation device according to Embodiment 5 of this application.

DETAILED DESCRIPTION

[0026] High-quality videos are videos that are potentially popular with viewers. A capability to accurately and comprehensively select high-quality videos from remaining videos after filtering out the offending videos or non-compliance videos to show to viewers determines key indicators of a short video app, such as a viewing duration, volume saved, and volume downloaded by a viewer of the short video app.

[0027] Two main types of quality video evaluation methods are commonly used in related technologies. One is to evaluate whether a video is of high quality based on external information of the video, that is, to evaluate quality of the video based on feedback data from viewers after the video is released. This method has a disadvantage of lacking timeliness because the quality of the video is evaluated only after the video is released and feedbacks from the viewers are collected. The other is to evaluate whether content of the video is of high quality based on some simple image parameters in the video. For example, a cover or key frame of the video is used to classify or cluster images to form some simple description information about the content of the video, and determine whether the video content is of high quality based on the description information. The disadvantage of this method is that the image parameters used are too simple to accurately and comprehensively describe the content of the video, and therefore the quality of the video cannot be accurately evaluated.

[0028] The following describes this application with reference to the accompanying drawings and embodiments. The specific embodiments described herein are merely intended to explain this application, rather than to limit this application. For ease of description, only some structures rather than all structures related to this application are shown in the accompany drawings.

[0029] Embodiment 1

[0030] FIG. 1 is a flowchart of a video quality evaluation method according to Embodiment 1 of this application. This embodiment is applicable to a case of video quality evaluation. The method may be performed by a video quality evaluation apparatus, which may be implemented by hardware and/or software. The video quality evaluation apparatus may include a plurality of physical entities, or may include one physical entity. In an embodiment, the video quality evaluation apparatus is integrated in a computer device.

[0031] As shown in FIG. 1, the video quality evaluation method provided by this embodiment includes the following steps:

[0032] In S110, an image sequence and audio information of a to-be-evaluated video are acquired by decoding the to-be-evaluated video.

[0033] In this embodiment, a video server receives a large quantity of videos produced and uploaded by users every day. The video server needs to remove offending videos from all the videos uploaded by the users, and sifts out videos that viewers may like. Therefore, the videos uploaded by the users need to be detected and screened, and the videos are recommended to the viewers according to a screening result.

[0034] The video quality evaluation method according to this embodiment may be performed by the video quality evaluation apparatus. FIG. 2 is a schematic diagram of an application scenario of the video quality evaluation method according to Embodiment 1 of this application. As shown in FIG. 2, a user records a video through a user-side terminal 210 and uploads the video to a server 220. Then, a backend 221 or a recommendation system 222 in the server 220 invokes a video quality evaluation apparatus 223. The video quality evaluation apparatus 223 evaluates the video uploaded by the user, outputs an evaluation result, and returns the evaluation result to the backend 221 or the recommendation system 222. The backend 221 or the recommendation system 222 processes the video according to the evaluation result. In an embodiment, if the evaluation result is that the video is a high-quality video, the recommendation system 222 recommends the video to a viewer-side terminal 230. In an embodiment, the user-side terminal 210 and the viewer-sider terminal 230 may be a same terminal device, or may be two different terminal devices.

[0035] In this embodiment, the to-be-evaluated video is a video uploaded by the user to the video server and whose video quality is not known. To meet a data transmission requirement, the user needs to code the video before uploading the video. Coding refers to conversion of a video file in a format into a video file in another format through a specific compression technology. In an embodiment, the to-be-evaluated video at least includes video, audio and/or script. The to-be-

evaluated video may be considered as a set of a plurality of images. The coded video needs to be decoded before evaluation of the to-be-evaluated video. The video may be considered to include a plurality of fast-playing images. Due to visual persistence of naked eyes, the picture is perceived as continuous. Therefore, decoding the video can acquire a plurality of frames of image data, and each frame of image data includes one image. After the video is decoded, an image sequence including a plurality of images arranged in a particular order may be acquired.

[0036] In this embodiment, after the to-be-evaluated video is decoded, audio information of the to-be-evaluated video may be further acquired. The audio information includes background music and language information of a character in the to-be-evaluated video.

[0037] In S120, an action feature vector and a face feature vector are extracted from the image sequence, and an audio feature vector is extracted from the audio information.

[0038] In this embodiment, the action feature vector refers to an action feature represented in a form of a vector. The action feature includes at least a character action and a surrounding environment in the to-be-evaluated video. For example, the character action is a dance movement, a sports movement, or the like of a character. The surrounding environment is a building, a high mountain, a bush, or the like. The action feature largely determines popularity of a video.

[0039] In this embodiment, a pose vector of each frame of image in the image sequence is extracted. One frame of image corresponds to one piece of matrix information. The corresponding pose vector may be acquired by merging the matrix information. The pose vector reflects image information corresponding to each frame of image. One image sequence may include a plurality of frames of images, and each frame of image corresponds to one pose vector. A plurality of acquired pose vectors are clustered, to convert the plurality of pose vectors into action feature vectors. A method for extracting the action feature vector is only described but not limited in this embodiment. Any method for extracting the action feature vector may be used according to actual needs.

[0040] In an embodiment, an action feature extraction network may be established in advance, and the action feature vector is acquired from the image sequence by extracting features from on each frame of image in the image sequence by using the action feature extraction network. In an embodiment, an action feature extraction network is acquired by fine-tuning an action recognition network. The image sequence is input into the action feature extraction network, and an action feature vector is acquired from the image sequence by extracting features from each frame of image by using the action feature extraction network. The action recognition network may include at least one type of network that has achieved a favorable effect in an action

classification data set, including but not limited to C3D based on a three-dimensional (3D) convolutional network, an interactive three-dimensional (I3D) convolutional network, a temporal segment network (TSN) based on a two-stream (2-Stream) method of optical flow and space, and a plurality of action recognition networks based on a recurrent neural network (RNN).

[0041] The face feature vector is a vector used for describing a face feature. Extraction of a face feature vector refers to a process of acquiring a face feature vector in a face region based on face detection.

[0042] There are mainly two steps of extracting the face feature vector in this embodiment. First, face detection is performed on the image sequence, and then, features are extracted from the detected face information. In an embodiment, first, the face detection is performed on the acquired image sequence by a multi-task convolutional neural network (MTCNN) method, so as to acquire face detection data. Then, the face feature vector is acquired by extracting features from the face detection data by a FaceNet method. In this embodiment, the convolutional neural network is a major technology of deep learning, and the MTCNN method is a method for face detection. Face detection intends to determine whether a face appears in an image. FaceNet is a method, in which a face feature vector is acquired by mapping a face to a Euclidean feature space by using a convolutional neural network.

[0043] Background music or character voice in a video is also one of important contents in the video. Features of audio information of the background music or the character voice need to be extracted. Quality of the video is estimated according to the features of the audio information. After an audio signal is extracted from the video, some conventional audio signal processing methods need to be used to pre-process audio, to extract a spectrum or cepstrum feature of the audio, and then send the spectrum or cepstrum feature of the audio to the audio feature extraction network, to extract an audio feature vector that can be used for music recognition or character speech recognition.

[0044] The audio feature vector is a vector for describing sound information in a video. In an embodiment, extracting the audio feature vector from the audio information refers to extracting a component that reflect an essential feature of the audio information, such as a center frequency of the audio, or energy features of the audio in some frequency bands or energy distribution features of the audio over a plurality of periods of time, and combining such features to acquire the corresponding audio feature vector.

[0045] In S130, a video feature vector is constructed according to the action feature vector, the face feature vector and the audio feature vector.

[0046] In this embodiment, the video feature vector is a vector used for representing video quality. The video feature vector is a feature vector constructed according to the action feature vector, the face feature vector and the audio feature vector construct.

[0047] The video feature vector is acquired by combining the action feature vector, the face feature vector and the audio feature vector extracted in S120. In an embodiment, if only one type of feature vector is extracted from the to-be-evaluated video, other two feature vectors that fail to be extracted are set to 0, to form the video feature vector. For example, if only the action feature vector is extracted from the to-be-evaluated video, both the face feature vector and the audio feature vector are set to 0, to form the video feature vector. If only two types of feature vectors are extracted from the to-be-evaluated video, the other one feature vector that fails to be extracted is set to 0, to form the video feature vector. For example, if only the action feature vector and the face feature vector are extracted from the to-be-evaluated video, the audio feature vector is set to 0, to form the video feature vector.

[0048] In S140, a quality score of the to-be-evaluated video is determined according to the video feature vector.

[0049] In this embodiment, a video evaluation network is built and trained in advance, and the video feature vector is input into the video evaluation network. The video evaluation network outputs a score as the quality score of the to-be-evaluated video. In an embodiment, the video evaluation network may be a simple multilayer perceptron (MLP) or a relatively complex RNN network. In an embodiment, a high-quality video training data set is established, and corresponding features are extracted from the high-quality video training data set according to a manner in S120, and stored as training data of the video evaluation network.

[0050] In an embodiment, a case that there is no face in some videos needs to be considered during training of the video evaluation network. Therefore, during the training, face feature vectors need to be arbitrarily replaced with all zeros in the high-quality video set, as a data enhancement means, such that a trained network can be better adapted to actual data.

[0051] In the video quality evaluation method provided by this embodiment, an image sequence and audio information of a to-be-evaluated video is acquired by decoding the to-be-evaluated video; then an action feature vector and a face feature vector are extracted from the image sequence and an audio feature vector is extracted from the audio information; then a video feature vector is constructed according to the action feature vector, the face feature vector, and the audio feature vector; and finally, a quality score of the to-be-evaluated video is determined according to the video feature vector. According to this embodiment, the quality score of the to-be-evaluated video can be determined according to at least one extracted vector of the action

feature vector, the face feature vector and the audio feature vector in the to-be-evaluated video, thereby improving timeliness of video quality evaluation and evaluation accuracy of a high-quality video.

[0052] Embodiment 2

[0053] FIG. 3 is a flowchart of a video quality evaluation method according to Embodiment 2 of this application. This embodiment is applicable to a case of video quality evaluation. The video quality evaluation method in the foregoing embodiment is described in this embodiment. As shown in FIG. 3, the video quality evaluation method provided by this embodiment includes the following steps:

[0054] In S310, an image sequence and audio information of a to-be-evaluated video are acquired by decoding the to-be-evaluated video.

[0055] In S320, an action feature vector is acquired from the image sequence by extracting features from each frame of image in the image sequence by using an action feature extraction network.

[0056] In this embodiment, the action feature extraction network needs to be built and trained before extracting features from each frame of image by using the action feature extraction network. In an embodiment, the building and training the action feature extraction network mainly includes the following steps: acquiring a training video set, and extracting image sequences in each training video in the training video set; acquiring the action classification network by training in a video database; and acquiring the action feature extraction network by training the action classification network with the image sequences in each training video in the training video set and deleting a classification layer from the action classification network.

[0057] In an embodiment, the training video set is a high-quality video set acquired by a manual screening method. The more types and a larger quantity of high-quality videos are, the higher the accuracy of the action feature extraction network is and the more effective the video evaluation result is. First, the high-quality video set is acquired, and an image sequence corresponding to each high-quality video is acquired by decoding each high-quality video in the high-quality video set.

[0058] The action classification network is acquired by training in the video database based on a neural network. The neural network includes but is not limited to C3D based on a 3D convolutional network, an I3D convolutional network, a TSN based on a 2-stream method of an optical flow and space, and a plurality of action recognition networks based on an RNN. The video database includes but is not limited to Kinetics and YouTube-8M data sets. Kinetics is a large public data set that is a data set with precise labeled actions. YouTube-8M is also a large

public data set that is a large video data set. The action classification network may be acquired by training by videos in the Kinetics data set or the YouTube-8M data set. In this embodiment, a method for training the action classification network is not limited, and any model training method in the related technology can be used, so as to acquire the action classification network through training.

[0059] After the action classification network is acquired, the action classification network is fine-tuned with the image sequence extracted from each high-quality video in the high-quality video set, and the classification layer is deleted from the action classification network, so as to acquire the action feature extraction network.

[0060] In S330, a face feature vector is extracted from the image sequence by using a face feature extraction network.

[0061] In this embodiment, the face feature extraction network includes a face detection sub-network and a feature extraction sub-network. In an embodiment, extracting the face feature vector from the image sequence by using the face feature extraction network includes: detecting a face image in each frame of image in the image sequence by using the face detection sub-network; and acquiring the face feature vector from the image sequence by extracting features the face image in each frame of image in the image sequence by using the feature extraction sub-network.

[0062] In this embodiment, the face detection sub-network may be constructed by a common face detection network, for example, multi-task cascaded convolutional networks (MTCNN). In an embodiment, detecting a face image in each frame of image in the image sequence by using the face detection sub-network includes: first sifting out candidate face regions and corresponding candidate face region boarders in each frame of image by using a proposed network (P-Net) in the MTCNN network, and then merging highly overlapping candidate face region boarders by using a maximum suppression algorithm; then refining the candidate face regions by using a refining network (R-Net) in an MTCNN algorithm, so as to acquire a candidate face image, and deleting an incorrect candidate face region border by using a bounding box regression algorithm and a non-maximum suppression algorithm; and filtering the remaining candidate face region borders by using an output network (O-Net) in the MTCNN algorithm, and detecting a face from the candidate face image. In an embodiment, when there are a plurality of faces in the video, a face that is relatively large and close to the front of the image is selected.

[0063] In an embodiment, a public pre-trained face detection network may be directly used as the face detection sub-network, or a face detection network may be trained based on a high-

quality video set, such that the face detection sub-network is more accurate in character recognition in a high-quality video.

[0064] The feature extraction sub-network may include a public pre-trained face recognition network, including but not limited to, FaceNet, InsightFace, or the like, or a network may be autonomously designed with reference to the principle of the feature extraction sub-network, and then a dedicated face recognition model is acquired through training based on the high-quality video data set. A plurality of face images of a same person are needed in a training process of the feature extraction sub-network and may be acquired from a selfie video uploaded by the same user. The trained feature extraction sub-network outputs highly similar feature vectors for different face images of a same person. The feature vector includes face appearance information of the person, which may be used to identify different persons, or may be used as a basis for estimating popularity of looks.

[0065] This embodiment provides a feature extraction sub-network training method. The method includes the following steps: acquiring a training video set, and extracting face images in each training video in the training video set; acquiring a face recognition network by training in a face database; and acquiring the feature extraction sub-network by training the face recognition network using the face images in each training video in the training video set. The training video set is a high-quality video set acquired by a manual screening method.

[0066] In S340, a cepstral coefficient and/or a spectrum of the audio information is extracted, and an audio feature vector is acquired by extracting features from the cepstral coefficient and/or the spectrum by using an audio feature extraction network.

[0067] Music or character voice in a video is also one of important contents in the video. A feature vector of a sound of the music or the character voice needs to be extracted by using the audio feature extraction network, so as to estimate the quality of the video according to the feature vector of the sound.

[0068] In this embodiment, the cepstral coefficient is a mel frequency cepstral coefficient (MFCC). The mel frequency is proposed based on human hearing features, and the mel frequency and a hertz (Hz) frequency have a non-linear correspondence. The MFCC coefficient is a frequency feature acquired through calculation based on the relationship between the mel frequency and the Hz frequency. In this embodiment, the method for extracting the cepstral coefficient of the audio information is not limited, and any method can be used to extract the cepstral coefficient of the audio information.

[0069] In an embodiment, the spectrum refers to analysis of a frequency-related part rather than a time-related part in the audio information during analysis of the audio information. Extracting

the spectrum of the audio information is extracting the frequency-related part in the audio information. In this embodiment, the method for extracting the spectrum of the audio information is not limited, and any method can be used to extract the spectrum of the audio information. In this embodiment, any one of the cepstral coefficient or the spectrum of the audio information may be extracted, or both the cepstral coefficient and the spectrum of the audio information may be extracted.

[0070] In an embodiment, prior to extracting the cepstral coefficient and/or the spectrum of the audio information, the following is further included: after an audio signal is extracted from the video, audio needs to be pre-processed first by using some conventional audio signal processing methods. In an embodiment of Embodiment 1, pre-processing at least includes noise reduction, gain, data normalization, and other processing. Pre-processing the audio information can improve accuracy of audio recognition. An algorithm used for audio pre-processing needs to be determined according to a requirement of a pre-trained data set, and remain consistent during training and prediction.

[0071] In this embodiment, after the cepstral coefficient and/or the spectrum of the audio information is extracted, the audio feature vector is acquired by extracting features from the cepstral coefficient and/or the spectrum by using the audio feature extraction network.

[0072] In this embodiment, a method for training an audio feature extraction network is provided. The method for training an audio feature extraction network includes: acquiring a training video set, and extracting the audio information in each training video in the training video set; acquiring an audio recognition network by training in an audio database; and acquiring the audio feature extraction network by training the audio recognition network using the audio information in each training video in the training video set. In an embodiment, the audio feature extraction network may be any network based on a convolutional neural network (Convolutional Neural Networks, CNN) or an RNN. An audio recognition network is pre-trained based on a public data set used for music recognition or character speech recognition, and then the audio feature extraction network is acquired by fine-tuning the audio recognition network based on a high-quality video training set.

[0073] The three steps S320, S330, and S340 may be performed in any order, and the order in which the three steps are performed is not limited in this embodiment.

[0074] In S350, a video feature vector is constructed according to the action feature vector, the face feature vector, and the audio feature vector.

[0075] In S360, a quality score of the to-be-evaluated video is acquired by inputting the video feature vector into a video evaluation network.

[0076] In this embodiment, the video evaluation network may be a simple multilayer perceptron or a relatively complex RNN network. An input of the video evaluation network is a video feature vector, and an output is a score indicating quality of a video. In an embodiment, if the output score is greater than 0 points, the actual output score is used as the quality score of the to-be-evaluated video. If the output score is less than 0 points, 0 points are used as the quality score of the to-be-evaluated video.

[0077] In an embodiment, the video evaluation network includes a first-layer evaluation network, a second-layer evaluation network, and an output layer.

[0078] Acquiring the quality score of the to-be-evaluated video by inputting the video feature vector into the video evaluation network includes: inputting the video feature vector into the first-layer evaluation network, and acquiring, by the first-layer evaluation network, a primary feature vector through recursive and non-linear operations; inputting, by the first-layer evaluation network, the primary feature vector into the second-layer evaluation network, and acquiring, by the second-layer evaluation network, an advanced feature vector through recursive and non-linear operations and inputting the advanced feature vector into the output layer; and acquiring the quality score of the to-be-evaluated video by performing, by the output layer, operations on the primary feature vector and the advanced feature vector. In an embodiment, a case that there is no face in some videos needs to be considered during training of the video evaluation network. Therefore, during the training, face features need to be arbitrarily replaced with all zeros in the high-quality video set, as a data enhancement means, such that a trained model can be better adapted to actual data.

[0079] Based on the foregoing embodiment, S320 may be replaced by the following: extracting optical flow data of each frame of image in the image sequence; and acquiring the action feature vector from the image sequence by extracting features from the optical flow data of each frame of image in the image sequence by using an action feature extraction network.

[0080] In the video quality evaluation method provided by this embodiment, an image sequence and audio information of a to-be-evaluated video are first acquired by decoding the to-be-evaluated video; then an action feature vector is acquired from the image sequence by extracting features from each frame of image in the image sequence by using an action feature extraction network; a face feature vector is extracted from the image sequence by using a face feature extraction network; , an audio feature vector is acquired by extracting features from a cepstral coefficient and/or a spectrum of the audio information by using an audio feature extraction network; then a video feature vector is constructed according to the action feature vector, the face feature vector, and the audio feature vector; and finally, a quality score of the to-be-

evaluated video is determined according to the video feature vector. According to this embodiment, the quality score of the to-be-evaluated video can be determined according to at least one extracted vector of the action feature vector, the face feature vector, and the audio feature vector in the to-be-evaluated video, thereby improving timeliness of video quality evaluation and evaluation accuracy of a high-quality video.

[0081] Embodiment 3

[0082] Based on the foregoing embodiment, this embodiment provides an instance. FIG. 4 is a flowchart of a video quality evaluation method according to Embodiment 3 of this application. As shown in FIG. 4, the video quality evaluation method includes the following steps:

[0083] In S4010, a to-be-evaluated video is acquired. The to-be-evaluated video may be a video that a user has just recorded and uploaded to a server, or may be a video stored in a server and on which quality evaluation has not been performed. In an embodiment, the to-be-evaluated video in this embodiment is a short video recorded by a user.

[0084] In S4020, the to-be-evaluated video is decoded, such that an image sequence including a plurality of images arranged in a particular order may be acquired.

[0085] In S4030, after the to-be-evaluated video is decoded, audio information of the to-be-evaluated video may be acquired. The audio information includes background music and language information of a character in the to-be-evaluated video.

[0086] S4020 and S4030 may be performed in sequence or in a reverse order, that is, S4030 may be performed after S4020, or S4020 may be performed after S4030. In an embodiment, S4020 and S4030 may be performed at the same time. In this embodiment, an order in which S4020 and S4030 are performed is not limited.

[0087] In S4040, an optical flow sequence is acquired by extracting optical flow data of each frame of image in the image sequence.

[0088] In S4050, the image sequence acquired through decoding in S4020 or the optical flow sequence acquired S4040 is input into an action feature extraction network.

[0089] In S4060, the action feature extraction network outputs an action feature vector. The action feature vector is a vector of an action feature represented in a form of a vector. The action feature includes at least a character action and a surrounding environment in the to-be-evaluated video. For example, the character action is a dance movement, a sports movement, or the like of a character. The surrounding environment is a building, a high mountain, a bush, or the like. The action feature largely determines popularity of a video.

[0090] In S4070, a face image is acquired by performing, by using a face detection sub-network, face detection, on the image sequence acquired through extraction in S4020. A public pre-trained

face detection network may be directly used as the face detection sub-network, or a face detection network may be trained based on a high-quality video set, such that the face detection sub-network is more accurate in character recognition in a high-quality video.

[0091] In S4080, the face image acquired in S4070 is input into a feature extraction sub-network. The feature extraction sub-network may include a public pre-trained face recognition network, including but not limited to, FaceNet, InsightFace, or the like, or a network may be autonomously designed with reference to the principle of the feature extraction sub-network, and then a dedicated face recognition model is acquired through training based on the high-quality video data set.

[0092] In S4090, the feature extraction sub-network outputs a face feature vector. The trained feature extraction sub-network outputs highly similar feature vectors for different face images of a same person. The feature vector includes face appearance information of the person, which may be used to identify different persons, or may be used as a basis for estimating popularity of looks.

[0093] In S4100, after the audio information acquired by decoding in S4030 is pre-processed, spectrum in the audio information is extracted.

[0094] In S4110, after the audio information acquired by decoding in S4030 is pre-processed, a mel-frequency cepstral coefficient in the audio information is extracted.

[0095] Background music or character voice in a video is also one of important contents in the video. Features of audio information of the background music or the character voice need to be extracted. Quality of the video is estimated according to the features of the audio information.

[0096] In S4120, the spectrum acquired in S4100 and the mel-frequency cepstral coefficient acquired in S4110 are input into an audio feature extraction network. The audio feature extraction network may be any network based on a CNN or an RNN. An audio recognition network is pre-trained based on a public data set used for music recognition or character speech recognition, and then the audio recognition network is fine-tuned based on a high-quality video training set, to acquire the audio feature extraction network.

[0097] In S4130, the audio feature extraction network outputs an audio feature vector. The audio feature vector is a vector for describing sound information in a video.

[0098] In S4140, a video feature vector is constructed according to the action feature vector, the face feature vector, and the audio feature vector.

[0099] In S4150, the video feature vector is input into a video evaluation network. The video evaluation network may be a simple multilayer perceptron or a relatively complex RNN network.

An input of the video evaluation network is a video feature vector, and an output is a score indicating quality of a video.

[00100] In S4160, the video evaluation network outputs a video score. The video score is used for measuring quality of a video. A higher score indicates higher quality of the video.

[00101] Based on the foregoing embodiment, this embodiment provides an instance. A user uploads a one-minute short video through a short video application. In an embodiment, the short video is a video of a character A dancing, and background music is music B. It is assumed that the short video includes 24 frames per second. In this case, the one-minute short video is decoded, and 1440 frames of image data and one-minute audio information may be acquired. Each frame of image data includes one image, that is, an image sequence including 1440 continuous images may be acquired. The image sequence including the 1440 images is input into the action feature extraction network and the face feature extraction network. The action feature extraction network outputs an action feature vector for representing a dance movement of the character A in the video. The face feature extraction network outputs a face feature vector for representing face appearance information of the character A in the video. The audio information is input into the audio feature extraction network, and the audio feature extraction network outputs an audio feature vector. For example, the action feature vector is a vector of three rows and five columns, the face feature vector is a vector of five row and eight columns, and the audio feature vector is a vector of four rows and nine columns. A video feature vector is constructed according to at least one of the action feature vector, the face feature vector and the audio feature vector. A video feature vector of five rows and 22 columns may be acquired. The video feature vector is input into the video evaluation network. The video evaluation network outputs a score. For example, the score may be 95 points.

[00102] In the video quality evaluation method provided by this embodiment, an image sequence and audio information of a to-be-evaluated video is acquired by decoding the to-be-evaluated video; then an action feature vector and a face feature vector are extracted from the image sequence and an audio feature vector is extracted from the audio information; then a video feature vector is constructed according to the action feature vector, the face feature vector, and the audio feature vector; and finally, a quality score of the to-be-evaluated video is determined according to the video feature vector. According to this embodiment, the quality score of the to-be-evaluated video can be determined according to at least one extracted vector of the action feature vector, the face feature vector and the audio feature vector in the to-be-evaluated video, thereby improving timeliness of video quality evaluation and evaluation accuracy of a high-quality video.

[00103] Embodiment 4

[00104] FIG. 5 is a schematic structural diagram of a video quality evaluation apparatus according to Embodiment 4 of this application. This embodiment is applicable to a case of video quality evaluation. The video quality evaluation apparatus may be implemented by software and/or hardware. The video quality evaluation apparatus is set to be integrated in a terminal device. A structure of the video quality evaluation apparatus includes the following: a video decoding module 510, configured to acquire an image sequence and audio information of a to-be-evaluated video by decoding the to-be-evaluated video; a vector extracting module 520, configured to extract an action feature vector and a face feature vector from the image sequence, and extract an audio feature vector from the audio information; a vector constructing module 530, configured to construct a video feature vector according to the action feature vector, the face feature vector, and the audio feature vector; and a quality scoring module 540, configured to determine a quality score of the to-be-evaluated video according to the video feature vector.

[00105] In the video quality evaluation apparatus provided by this embodiment, an image sequence and audio information of a to-be-evaluated video is acquired by decoding the to-be-evaluated video; then an action feature vector and a face feature vector are extracted from the image sequence and an audio feature vector is extracted from the audio information; then a video feature vector is constructed according to the action feature vector, the face feature vector, and the audio feature vector; and finally, a quality score of the to-be-evaluated video is determined according to the video feature vector. According to this embodiment, the quality score of the to-be-evaluated video can be determined according to at least one extracted vector of the action feature vector, the face feature vector and the audio feature vector in the to-be-evaluated video, thereby improving timeliness of video quality evaluation and evaluation accuracy of a high-quality video.

[00106] Embodiment 5

[00107] FIG. 6 is a schematic structural diagram of a video quality evaluation device according to Embodiment 5 of this application. As shown in FIG. 6, the device includes a processor 610 and a memory 620. There may be one or more processors 610 in the device, and in FIG. 6, for example, there is one processor 610. The processor 610 and the memory 620 in the device may be connected through a bus or in another manner, and in FIG. 6, for example, the connection is implemented through a bus.

[00108] As a computer-readable storage medium, the memory 620 may be configured to store software programs, computer executable programs, and modules, such as program instructions/modules (for example, the video decoding module 510, the vector extracting module

520, the vector constructing module 530, and the quality scoring module 540 in the video quality evaluation apparatus) corresponding to the video quality evaluation method in the embodiments of this application. The processor 610 runs the software programs, instructions, and modules stored in the memory 620, to perform various functional applications of the device and data processing, that is, to implement the video quality evaluation method described above.

[00109] Embodiment 6

[00110] Embodiment 6 further provides a non-transitory computer-readable storage medium including computer executable instructions. The computer executable instructions, when run by a computer processor, causes the processing apparatus to perform a video quality evaluation method. The method includes: acquiring an image sequence and audio information of a to-be-evaluated video by decoding the to-be-evaluated video; extracting an action feature vector and a face feature vector from the image sequence, and extracting an audio feature vector from the audio information; constructing a video feature vector according to the action feature vector, the face feature vector and the audio feature vector; and determining a quality score of the to-be-evaluated video according to the video feature vector.

[00111] In the storage medium including computer executable instructions according to this embodiment of this application, in addition to performing the foregoing method operations, the computer executable instructions can also be used to perform related operations in the video quality evaluation method according to any embodiment of this application.

CLAIMS

What is claimed is:

1. A video quality evaluation method, comprising:

acquiring an image sequence and audio information by decoding a to-be-evaluated video, wherein the to-be-evaluated video is a non-offending video;

extracting an action feature vector and a face feature vector from the image sequence, and extracting an audio feature vector from the audio information;

constructing a video feature vector according to at least one of the action feature vector, the face feature vector and the audio feature vector;

inputting the video feature vector into a first-layer evaluation network, and acquiring, by the first-layer evaluation network, a primary feature vector through recursive and non-linear operations;

inputting the primary feature vector into a second-layer evaluation network, and acquiring, by the second-layer evaluation network, an advanced feature vector through recursive and non-linear operations; and

acquiring a quality score of the to-be-evaluated video by performing, by an output layer, operations on the primary feature vector and the advanced feature vector,

wherein constructing the video feature vector according to at least one of the action feature vector, the face feature vector and the audio feature vector comprises:

setting the feature vector that fails to be extracted among the action feature vector, the face feature vector and the audio feature vector to zero; and

constructing the video feature vector by combining the action feature vector, the face feature vector and the audio feature vector.

2. The method according to claim 1, wherein extracting the action feature vector from the image sequence comprises:

acquiring the action feature vector by extracting features from each frame of image by using an action feature extraction network;

and/or

extracting optical flow data of each frame of image; and

acquiring the action feature vector by extracting features from the optical flow data of each

frame of image by using an action feature extraction network.

3. The method according to claim 2, further comprising:

acquiring a training video set, and extracting image sequences in each training video;

acquiring an action classification network by training in a video database; and

acquiring the action feature extraction network by training the action classification network with the image sequences in each training video and deleting a classification layer from the action classification network.

4. The method according to claim 1, wherein extracting the face feature vector from the image sequence comprises:

extracting the face feature vector from the image sequence by using a face feature extraction network.

5. The method according to claim 4, wherein the face feature extraction network comprises a face detection sub-network and a feature extraction sub-network.

6. The method according to claim 5, wherein extracting the face feature vector from the image sequence by using the face feature extraction network comprises:

detecting a face image in each frame of image in the image sequence by using the face detection sub-network; and

acquiring the face feature vector by extracting features the face image in each frame of image by using the feature extraction sub-network.

7. The method according to claim 6, further comprising:

acquiring a training video set, and extracting face images in each training video;

acquiring a face recognition network by training in a face database; and

acquiring the feature extraction sub-network by training the face recognition network using the face images in each training video.

8. The method according to claim 1, wherein extracting the audio feature vector from the audio information comprises:

extracting at least one of a cepstral coefficient and a spectrum of the audio information; and

acquiring the audio feature vector by extracting features from at least one of the cepstral coefficient and the spectrum by using an audio feature extraction network.

9. The method according to claim 8, further comprising:

acquiring a training video set, and extracting the audio information in each training video;

acquiring an audio recognition network by training in an audio database; and

acquiring the audio feature extraction network by training the audio recognition network using the audio information in each training video.

10. A video quality evaluation device, comprising:

at least one processor; and

a memory storing at least one program therein, wherein

the at least one program, when run by the at least one processor, causes the at least one processor to perform a video quality evaluation method, comprising:

acquiring an image sequence and audio information by decoding a to-be-evaluated video, wherein the to-be-evaluated video is a short video recorded by a user of a short video application;

extracting an action feature vector and a face feature vector from the image sequence, and extracting an audio feature vector from the audio information;

constructing a video feature vector according to at least one of the action feature vector, the face feature vector and the audio feature vector;

inputting the video feature vector into a first-layer evaluation network, and acquiring, by the first-layer evaluation network, a primary feature vector through recursive and non-linear operations;

inputting the primary feature vector into a second-layer evaluation network, and acquiring, by the second-layer evaluation network, an advanced feature vector through recursive and non-linear operations; and

acquiring a quality score of the to-be-evaluated video by performing, by an output layer,

operations on the primary feature vector and the advanced feature vector,

wherein constructing the video feature vector according to at least one of the action feature vector, the face feature vector and the audio feature vector comprises:

setting the action feature vector, the face feature vector or the audio feature vector to zero in response its failed extraction; and

constructing the video feature vector by combining the action feature vector, the face feature vector and the audio feature vector.

11. The device according to claim 10, wherein extracting the action feature vector from the image sequence comprises:

acquiring the action feature vector by extracting features from each frame of image by using an action feature extraction network;

and/or

extracting optical flow data of each frame of image; and

acquiring the action feature vector by extracting features from the optical flow data of each frame of image by using an action feature extraction network.

12. The device according to claim 11, wherein the method further comprises:

acquiring a training video set, and extracting image sequences in each training video;

acquiring an action classification network by training in a video database; and

acquiring the action feature extraction network by training the action classification network with the image sequences in each training video and deleting a classification layer from the action classification network.

13. The device according to claim 10, wherein extracting the face feature vector from the image sequence comprises:

extracting the face feature vector from the image sequence by using a face feature extraction network.

14. The device according to claim 13, wherein the face feature extraction network

comprises a face detection sub-network and a feature extraction sub-network.

15. The device according to claim 14, wherein extracting the face feature vector from the image sequence by using the face feature extraction network comprises:

detecting a face image in each frame of image in the image sequence by using the face detection sub-network; and

acquiring the face feature vector by extracting features the face image in each frame of image by using the feature extraction sub-network.

16. The device according to claim 15, wherein the method further comprises:

acquiring a training video set, and extracting face images in each training video;

acquiring a face recognition network by training in a face database; and

acquiring the feature extraction sub-network by training the face recognition network using the face images in each training video.

17. The device according to claim 10, wherein extracting the audio feature vector from the audio information comprises:

extracting at least one of a cepstral coefficient and a spectrum of the audio information; and

acquiring the audio feature vector by extracting features from at least one of the cepstral coefficient and the spectrum by using an audio feature extraction network.

18. The device according to claim 17, wherein the method further comprises:

acquiring a training video set, and extracting the audio information in each training video;

acquiring an audio recognition network by training in an audio database; and

acquiring the audio feature extraction network by training the audio recognition network using the audio information in each training video.

19. A non-transitory computer-readable storage medium storing a computer program, wherein the computer program, when run by a processing apparatus, causes the processing apparatus to perform a video quality evaluation method, comprising:

acquiring an image sequence and audio information by decoding a to-be-evaluated video, wherein the to-be-evaluated video is non-offending video;

extracting an action feature vector and a face feature vector from the image sequence, and extracting an audio feature vector from the audio information;

constructing a video feature vector according to at least one of the action feature vector, the face feature vector and the audio feature vector;

inputting the video feature vector into a first-layer evaluation network, and acquiring, by the first-layer evaluation network, a primary feature vector through recursive and non-linear operations;

inputting the primary feature vector into a second-layer evaluation network, and acquiring, by the second-layer evaluation network, an advanced feature vector through recursive and non-linear operations; and

acquiring a quality score of the to-be-evaluated video by performing, by an output layer, operations on the primary feature vector and the advanced feature vector,

wherein constructing the video feature vector according to at least one of the action feature vector, the face feature vector and the audio feature vector comprises:

setting the feature vector that fails to be extracted among the action feature vector, the face feature vector and the audio feature vector to zero; and

constructing the video feature vector by combining the action feature vector, the face feature vector and the audio feature vector.