

VIDEO QUALITY DETECTION METHOD AND APPARATUS, ELECTRONIC DEVICE AND STORAGE MEDIUM

TECHNICAL FIELD

[0001] The present disclosure relates to intelligent terminal technology, and in particular to a video quality detection method and apparatus, an electronic device and a computer-readable storage medium.

BACKGROUND

[0002] With the rapid development of smartphones and portable cameras, a large number of videos are created by a variety of consumers every day. The videos, which contain contents generated by users, are uploaded to social media sites or applications for the public to access via the Internet. Therefore, it is increasingly important to evaluate the quality of these videos in an objective way.

[0003] There are some methods for quality evaluation of videos in the related art, but the quality evaluation methods in the related art are to perform quality evaluation mainly based on individual video frames, which results in inaccurate quality evaluation of videos.

SUMMARY

[0004] Embodiments of the present disclosure provide a video quality detection method and apparatus, an electronic device and a computer-readable storage medium, which can improve the accuracy of video quality detection.

[0005] The technical solution of the embodiments of the present disclosure is implemented as follows.

[0006] The embodiments of the present disclosure provide a video quality detection method, which may include the following operations.

[0007] Feature extraction of different levels is performed on a video to be detected including multiple video frames to obtain multiple multilevel features corresponding to the multiple video frames.

[0008] A temporal difference feature between adjacent multilevel features of the multiple multilevel features is determined, and multiple first fusion features are obtained based on the temporal difference feature and the multiple multilevel features, the temporal difference feature being used for representing temporal distortion content in the video to be detected.

[0009] A quality parameter for reflecting the video quality of the video to be detected is obtained based on the multiple first fusion features.

[0010] The embodiments of the present disclosure provide a video quality detection apparatus, which may include an extraction unit and a determination unit.

[0011] The extraction unit is configured to perform feature extraction of different levels on the video to be detected including multiple video frames to obtain multiple multilevel features corresponding to the multiple video frames.

[0012] The determination unit is configured to determine a temporal difference feature between the adjacent multilevel features of the multiple multilevel features; obtain

multiple first fusion features based on the temporal difference feature and the multiple multilevel features, the temporal difference feature being used for representing temporal distortion content in the video to be detected; and obtain a quality parameter for reflecting the video quality of the video to be detected based on the multiple first fusion features.

- [0013] The embodiments of the present disclosure provide an electronic device, which may include: a memory configured to store an executable computer program; and a processor configured to implement, when executing the executable computer program stored in the memory, the above video quality detection method.
- [0014] The embodiments of the present disclosure provide a computer-readable storage medium, in which a computer program is stored. The computer program is configured to implement, when executed by a processor, the above video quality detection method.
- [0015] Through the video quality detection method and apparatus, the electronic device and the computer-readable storage medium provided by the embodiments of the present disclosure, because the feature which is used for representing the temporal distortion content in the video to be detected can be obtained, the video quality of a video to be detected in terms of temporal distortion between video frames can be determined based on this. Compared with video quality detection based on a single video frame, the detection dimensions of video quality detection are increased, which makes the detected video quality more accurate.
- [0016] It is to be understood that the above general descriptions and detailed descriptions below are only exemplary and explanatory and not intended to limit the present disclosure.

BRIEF DESCRIPTION OF THE DRAWINGS

- [0017] The accompanying drawings, which are incorporated in and constitute a part of this specification, illustrate embodiments consistent with the present disclosure and, together with the specification, serve to illustrate the technical solutions of the present disclosure.
- [0018] FIG. 1A illustrates all video frames in an exemplary video without an image distortion problem provided by an embodiment of the present disclosure.
- [0019] FIG. 1B illustrates all video frames in an exemplary video with an image distortion problem provided by an embodiment of the present disclosure.
- [0020] FIG. 2A illustrates all video frames in an exemplary video without a content inconsistency problem provided by an embodiment of the present disclosure.
- [0021] FIG. 2B illustrates all video frames in an exemplary video with a content inconsistency problem provided by an embodiment of the present disclosure.
- [0022] FIG. 3 illustrates an optional flowchart of a video quality detection method provided by an embodiment of the present disclosure.
- [0023] FIG. 4 illustrates an optional flowchart of a video quality detection method provided by an embodiment of the present disclosure.
- [0024] FIG. 5 illustrates a flowchart of an exemplary video quality detection method provided by an embodiment of the present disclosure.
- [0025] FIG. 6 illustrates an optional flowchart of a video quality detection method provided by an embodiment of the present disclosure.
- [0026] FIG. 7 illustrates an optional flowchart of a video quality detection method provided by an embodiment of the present disclosure.

- [0027] FIG. 8 illustrates an optional flowchart of a video quality detection method provided by an embodiment of the present disclosure.
- [0028] FIG. 9 illustrates an optional flowchart of a video quality detection method provided by an embodiment of the present disclosure.
- [0029] FIG. 10 illustrates an optional flowchart of a video quality detection method provided by an embodiment of the present disclosure.
- [0030] FIG. 11 illustrates another flowchart of an exemplary video quality detection method provided by an embodiment of the present disclosure.
- [0031] FIG. 12 illustrates an optional flowchart of a video quality detection method provided by an embodiment of the present disclosure.
- [0032] FIG. 13 illustrates an optional flowchart of a video quality detection method provided by an embodiment of the present disclosure.
- [0033] FIG. 14A illustrates a principle diagram of an exemplary first sampling mode provided by an embodiment of the present disclosure.
- [0034] FIG. 14B illustrates a principle diagram of an exemplary second sampling mode provided by an embodiment of the present disclosure.
- [0035] FIG. 15 illustrates a structural diagram of a video quality detection apparatus provided by an embodiment of the present disclosure.
- [0036] FIG. 16 illustrates a structural diagram of an electronic device provided by an embodiment of the present disclosure.

DETAILED DESCRIPTION

- [0037] In order to make the objectives, technical solutions and advantages of the disclosure clearer, the disclosure will further be described in detail below in combination with the drawings. The described embodiments should not be considered as limits to the disclosure. All other embodiments obtained by those of ordinary skill in the art without creative work shall fall within the scope of protection of the disclosure.
- [0038] There are some quality evaluation methods for videos in the related art, but the quality evaluation methods in the related art evaluate the quality of a whole video mainly based on individual video frames in the video, without considering the problems of image distortion (for example, wobbly shooting angle and unstable exposure of the video) and content inconsistency (for example, content uncorrelation between adjacent video frames) among multiple video frames of the video. Research indicates that both image distortion and content inconsistency will lead to the degradation of video quality. Exemplarily, FIG. 1A and FIG. 1B show a video without an image distortion problem and a video with an image distortion problem, respectively; FIG. 2A and FIG. 2B show a video without a content inconsistency problem and a video with a content inconsistency problem, respectively. As shown in FIG. 1A, the shooting angle continuously moves to the bottom right among three video frames from top to bottom in the a1-th group of video frames, making the shooting angle change smoothly in the a1-th group of video frames, that is, the shooting angle of the video does not wobble; the shooting angle continuously moves to the left among three video frames from top to bottom in the b1-th group of video frames, making the shooting angle also change smoothly in the b1-th group of video frames. As shown in FIG. 1B, the shooting angle moves first to the left and then to the right among three video frames from top to bottom in the c1-th group of video frames, making the shooting angle change unsmoothly in the c1-th group of video frames, that is, the shooting angle of the video wobbles; the shooting angle moves continuously to the bottom among three video frames from top to bottom in the d1-th

group of video frames, making the shooting angle also change smoothly in the d1-th group of video frames, that is, the shooting angle of the video does not wobble; however, the exposure is significantly different between the second video frame and the third video frame in the d1-th group of video frames, that is, the exposure is not stable. It can be known from comparison between FIG. 1A and FIG. 1B that when the video has an image distortion problem, a visual gap is brought to a viewer, which affects the viewer's viewing experience, and thus affects the viewer's evaluation on the video quality. As shown in FIG. 2A, the contents of adjacent video frames of three video frames in the a2-th group of video frames are correlated, and the contents of adjacent video frames of three video frames in the b2-th group of video frames are also correlated. As shown in FIG. 2B, the contents of the first video frame and the second video frame in the c2-th group of video frames are uncorrelated, and the contents of the first video frame and the second video frame, as well as the contents of the second video frame and the third video frame in the d2-th group of video frames are also uncorrelated. It can be known from comparison between FIG. 2A and FIG. 2B that when the video has a content inconsistency problem, a visual gap is brought to the viewer, which affects the viewer's viewing experience, and thus affects the viewer's evaluation on the video quality.

[0039] It can be known from the above that the method for evaluating the quality of a whole video mainly based on the individual video frames in the video in the related art makes the quality evaluation of videos inaccurate.

[0040] Based on this, the embodiments of the present disclosure provide a video quality detection method, which can improve the accuracy of video quality detection. The video quality detection method provided by the embodiments of the present disclosure is applied to an electronic device. The electronic device provided by the embodiments of the present disclosure may be implemented as augmented reality (AR) glasses, a laptop computer, a tablet computer, a desktop computer, a set-top box, a mobile device (for example, a mobile phone, a portable music player, a personal digital assistant, a dedicated message device, and a portable game device) and other types of user terminals (terminals for short), and may also be implemented as a server.

[0041] The video quality detection method provided by the embodiments of the present disclosure may be applied to video screening, for example, to social media sites or applications (e.g., a video viewing application) related to video browsing, thus, the social media sites or the applications related to video browsing may use the video quality detection method provided by the embodiments of the present disclosure to screen for videos that meet quality requirements from multiple videos uploaded by users and recommend them to users for browsing, or store them to related repositories of the social media sites or the applications related to video browsing for the users to download. Moreover, the video quality detection method provided by the embodiments of the present disclosure may also be applied to the evaluation of video recovery, enhancement real video generation, etc.; for example, after a certain video is recovered by using a certain algorithm, the video quality detection method provided by the embodiments of the present disclosure may be used to detect both the video quality of the recovered video and the video quality of the original video of the recovered video, and whether the algorithm is effective will be judged by comparing two detection results. The above two aspects are only illustrative description, and do not represent limitations on the application scenarios of the video quality detection method provided by the embodiments of the present disclosure. The video quality detection method provided by the embodiments of the present disclosure may also be applied to other aspects, which are

not listed here.

[0042] FIG. 3 illustrates an optional flowchart of a video quality detection method provided by an embodiment of the present disclosure, which will be illustrated in combination with the steps shown in FIG. 3.

[0043] At S101, feature extraction of different levels is performed on a video to be detected including multiple video frames to obtain multiple multilevel features corresponding to the multiple video frames.

[0044] In the embodiments of the present disclosure, the video to be detected may be any type of video, for example, a landscape video or a story video, which is not limited by the embodiments of the present disclosure.

[0045] In the embodiments of the present disclosure, for the video to be detected with N video frames $\{r_0, r_1, \dots, r_{N-1}\}$, the electronic device may use a backbone network to perform feature extraction of different levels (dimensions) on each video frame $r_i \Big|_{i=0}^{N-1}$, so as to obtain multiple features of different levels of each video frame $r_i \Big|_{i=0}^{N-1}$ at multiple levels, and fuse the multiple features of different levels to obtain a multilevel feature corresponding to each video frame $r_i \Big|_{i=0}^{N-1}$. Therefore, for N video frames, N multilevel features are obtained correspondingly, where N is an integer greater than or equal to 2.

[0046] In some embodiments, the multilevel features of each video frame may include low-level features, such as color, boundary division, contrast and brightness, and include high-level semantic information.

[0047] In some embodiments, as shown in FIG. 4, S101 may be implemented by S1011 to S1012, which will be illustrated in combination with the steps shown in FIG. 4.

[0048] At S1011, feature extraction of different levels is performed on the video to be detected including multiple video frames by using a pre-trained backbone network, so as to obtain multiple features of different levels of each video frame at multiple levels.

[0049] At S1012, the multiple features of different levels of each video frame are fused, so as to obtain multiple multilevel features corresponding to the multiple video frames.

[0050] In some embodiments, the electronic device may use the pre-trained backbone network to perform feature extraction of multiple levels on each video frame, so as to obtain multiple features of different levels of each video frame, and fuse the multiple features of different levels to obtain the multilevel feature of each video frame.

[0051] Exemplarily, the electronic device may use a pre-trained Video Swin Transformer (VSwin-T) as the backbone network to perform feature extraction of five levels to each video frame $r_i \Big|_{i=0}^{N-1}$, so as to obtain five features of different levels, and fuse the five features of different levels to obtain the multilevel feature of each video frame $r_i \Big|_{i=0}^{N-1}$, as shown in the following formula (1):

$$M_i^l = \psi_l \left(\dots \left(\psi_0 \left(r_j \Big|_{j=i-\frac{w_l}{2}}^{i+\frac{w_l}{2}-1} \right) \right) \right) \quad (1),$$

where ψ_l represents each layer of the backbone network, $l \in [0, 4]$, and w_l is a

cumulative time window size from ψ_0 to ψ_1 , which increases as l increases. Here, the electronic device may adjust the features of different sizes to the same size first, and then fuse the five features of adjusted sizes to obtain the multilevel feature M_i^l .

[0052] At S102, a temporal difference feature between adjacent multilevel features of the multiple multilevel features is determined, and multiple first fusion features are obtained based on the temporal difference feature and the multiple multilevel features, the temporal difference feature being used for representing temporal distortion content in the video to be detected.

[0053] In the embodiments of the present disclosure, for the multiple multilevel features corresponding to each video to be detected, the electronic device may determine the temporal difference feature between the adjacent multilevel features, and when the temporal difference feature between each group of adjacent multilevel features is obtained, fuse the temporal difference feature with the first multilevel feature of the group of adjacent multilevel features to obtain a first fusion feature corresponding to the group of adjacent multilevel features.

[0054] In some embodiments, S102 may be implemented in the following mode: Global Average Pooling (GAP) processing is performed on each multilevel feature of the multiple multilevel features to obtain multiple pooled features; and a difference feature between adjacent pooled features of the multiple pooled features is determined to obtain the temporal difference feature. Here, the electronic device may first perform GAP processing on each multilevel feature M_i^l , so as to pool M_i^l into one-dimensional features, combine the one-dimensional features into a multilevel feature vector to obtain the corresponding pooled feature $T_i = \oplus_{l=0}^4 \text{GAP}(M_i^l)$, and take the difference feature $T_i - T_{i+1}$ between each group of adjacent pooled features T_i and T_{i+1} as an obtained temporal difference feature. Moreover, when the temporal difference feature $T_i - T_{i+1}$ between the group of adjacent pooled features T_i and T_{i+1} is obtained, the electronic device fuses the temporal difference feature $T_i - T_{i+1}$ and the pooled feature T_i to obtain the corresponding first fusion feature $\hat{T}_i$. $\hat{T}_i$ may be expressed by the following formula (2):

$$\hat{T}_i = T_i \oplus (T_i - T_{i+1}) \quad (2),$$

where $\oplus$ represents feature fusion, and $T_i - T_{i+1}$ represents the temporal difference feature.

[0055] At S103, a quality parameter for reflecting the video quality of the video to be detected is obtained based on the multiple first fusion features.

[0056] Here, the quality parameter may be expressed by a video quality score. In some embodiments, when the multiple first fusion features corresponding to each video to be detected are obtained, the electronic device may determine a temporal distortion degree among multiple video frames of the video to be detected based on the multiple first fusion features obtained, and take the temporal distortion degree as the quality parameter for reflecting the video quality of the video to be detected, that is, the temporal distortion degree is the video quality score for reflecting the video quality of the video to be

detected.

[0057] Exemplarily, the electronic device uses a first Multilayer Perceptron (MLP) network to perform regression processing on multiple first fusion features to obtain the temporal distortion degree among multiple video frames. The first MLP network may be a two-layer MLP network.

[0058] FIG. 5 illustrates a flowchart of an exemplary video quality detection method provided by an embodiment of the present disclosure. As shown in FIG. 5, for a video I to be detected, multilevel features of the video I to be detected are extracted through a backbone network including five network layers; GAP processing is performed on each extracted multilevel feature to obtain a pooled feature; then, a difference feature between each two adjacent pooled features is calculated to obtain a temporal difference feature; the obtained temporal difference feature is fused with the pooled feature to obtain the first fusion feature; finally, regression processing is performed on first fusion features through the first MLP network to obtain the quality parameter for reflecting the video quality of the video I to be detected.

[0059] In the embodiments of the present disclosure, because the feature which is used for representing the temporal distortion content in the video to be detected can be obtained, the video quality of a video to be detected in terms of temporal distortion between video frames can be determined based on this. Compared with video quality detection based on a single video frame, the detection dimensions of video quality detection are increased, which makes the detected video quality more accurate.

[0060] In some embodiments, before S101, the electronic device may use a video in a training video set to train an initial backbone network and an initial first MLP network. The training video set is a training set composed of multiple videos, and the multiple videos may be of different types or the same type, which is not limited by the embodiments of the present disclosure. Here, for one video in the training video set, the electronic device may use an initial backbone network to perform initial multilevel feature extraction and GAP pooling, determine an initial first fusion feature corresponding to the initial multilevel feature, use the initial first MLP network to perform regression processing on the initial first fusion feature, so as to obtain an initial temporal distortion degree, and then use a loss function to calculate the current training loss according to the initial temporal distortion degree and a preset temporal distortion degree, and adjust network parameters of the initial backbone network and the initial first MLP network based on the current training loss to complete the current training process; thereafter, the electronic device continues to use the video and the adjusted network parameters to perform the next iterative training to the initial backbone network and the initial first MLP network until a preset training condition is met, for example, when the final loss is less than a preset loss threshold, the backbone network and the first MLP network are obtained.

[0061] In some other embodiments, the initial backbone network and the initial first MLP network may also be separately trained to obtain the backbone network and the first MLP network respectively.

[0062] In some embodiments, when multiple first fusion features corresponding to each video to be detected are obtained, the electronic device may continue to perform content correlation processing on the multiple first fusion features, obtain some features for representing the content with low correlation with the whole content of the video to be detected, and determine the quality parameter for reflecting the video quality of the video to be detected according to these features and the first fusion features. FIG. 6

illustrates an optional flowchart of a video quality detection method provided by an embodiment of the present disclosure. As shown in FIG. 6, S103 may be implemented by S1031 to S1032, which will be illustrated in combination with the steps shown in FIG. 6.

[0063] At S1031, content correlation processing is performed on multiple first fusion features based on a transformation network to obtain an output feature, the output feature being used for representing screened content, whose correlation with the whole content is lower than the remaining content in the whole content, in the whole content of the video to be detected, and the remaining content being the content in the whole content except the screened content.

[0064] At S1032, the quality parameter for reflecting the video quality of the video to be detected is obtained based on multiple first fusion features and the output feature.

[0065] In the embodiments of the present disclosure, the electronic device may use a pre-trained transformation network to perform content correlation feature processing on the obtained multiple first fusion features to obtain the output feature which can represent the screened content, whose correlation with the whole content is lower than other content in the whole content (that is, the content whose correlation with the whole content of the video to be detected is low), in the whole content of the video to be detected, and obtain the quality parameter for reflecting the video quality of the video to be detected according to the output feature and the multiple first fusion features.

[0066] In some embodiments, the transformation network may be a Transformer network. The electronic device may use a pre-trained Transformer network to perform content correlation encoding processing and decoding processing on the obtained multiple first fusion features, so as to obtain the output feature.

[0067] In some embodiments, S1032 may be implemented by S201 to S202.

[0068] At S201, a temporal distortion degree among multiple video frames is determined based on multiple first fusion features, and a content correlation degree among multiple video frames is determined based on the output feature.

[0069] At S202, the quality parameter for reflecting the video quality of the video to be detected is obtained based on the temporal distortion degree and the content correlation degree.

[0070] The electronic device may determine the temporal distortion degree between the video frames of the video to be detected according to the obtained multiple first fusion features, and determine the content correlation degree between the video frames of the video to be detected according to the obtained output feature; thereafter, the electronic device determines the quality parameter of the video quality of the video to be detected according to the obtained temporal distortion degree and content correlation degree.

[0071] In some embodiments, the electronic device may perform regression processing on multiple first fusion features to obtain the temporal distortion degree between the video frames of the video to be detected, and perform regression processing on the output feature to obtain the content correlation degree between the video frames of the video to be detected. Exemplarily, the electronic device may use the first MLP network to transform the first fusion feature into a one-dimensional score value, so as to achieve the regression processing on the first fusion feature, and use a second MLP network to transform the output feature into a one-dimensional score value, so as to achieve the regression processing on the output feature. The second MLP network has the same network structure as the first MLP network.

[0072] In some embodiments, as shown in FIG. 7, S202 may be implemented by

S2021, which will be illustrated in combination with the steps shown in FIG. 7.

- [0073] At S2021, the sum of the temporal distortion degree and the content correlation degree is taken as the quality parameter for reflecting the video quality of the video to be detected.
- [0074] After two values, namely the temporal distortion degree and the content correlation degree, of the video to be detected are obtained, the electronic device may directly take the sum of the temporal distortion degree and the content correlation degree as the quality parameter of the corresponding video to be detected.
- [0075] In the embodiments of the present disclosure, for each video to be detected, a corresponding quality parameter is obtained, and the electronic device may select the video with high or low quality by comparing the quality parameters of different videos to be detected.
- [0076] In some embodiments, as shown in FIG. 8, S1031 may be implemented by S301 to S304, which will be illustrated in combination with the steps shown in FIG. 8.
- [0077] At S301, content correlation encoding processing is performed on multiple first fusion features based on the encoding network to obtain multiple encoded features. Each encoded feature represents the correlation between a corresponding first fusion feature and other first fusion features except this first fusion feature in the multiple first fusion features. The transformation network includes: an encoding network and a decoding network.
- [0078] The electronic device may input the obtained multiple first fusion features into the encoding network of the transformation network. The content correlation encoding processing is performed on each first fusion feature through the encoding network, and then multiple encoded features are obtained correspondingly. Moreover, each encoded feature represents the correlation between one first fusion feature corresponding to the encoded feature and other first fusion features. Exemplarily, the encoding network may be an Encoder network in the Transformer network.
- [0079] At S302, the multiple encoded features and the multiple first fusion features are fused to obtain multiple second fusion features.
- [0080] When the multiple encoded features corresponding to the multiple first fusion features are obtained, the electronic device may fuse each encoded feature with the corresponding first fusion feature, so as to obtain multiple second fusion features.
- [0081] Exemplarily, for one first fusion feature, the electronic device may use the Encoder network (which may be a four-layer network including a multi-head attention layer and multiple feedforward layers) to perform the content correlation encoding processing to the first fusion feature, so as to obtain the corresponding encoded feature, and fuse the encoded feature with the first fusion feature to obtain one second fusion feature.
- [0082] At S303, temporal pooling processing is performed on the multiple first fusion features to obtain an average feature.
- [0083] When encoding processing is performed on the multiple first fusion features by using the encoding network, the electronic device may also perform temporal pooling processing on the multiple first fusion features to obtain the average feature corresponding to the multiple first fusion features.
- [0084] In some embodiments, as shown in FIG. 9, before S301, S401 may also be executed, and S301 may be implemented by S402, S302 may be implemented by S403, and S303 may be implemented by S404, which will be illustrated in combination with the steps shown in FIG. 9.

[0085] At S401, the dimensions of each first fusion feature of the multiple first fusion features is reduced to preset dimensions to obtain multiple pre-encoded features.

[0086] At S402, based on the encoding network, the content correlation encoding processing is performed on the multiple pre-encoded features to obtain multiple encoded features. Each encoded feature represents the correlation between a corresponding pre-encoded feature and other pre-encoded features except this pre-encoded feature in the multiple pre-encoded features. The transformation network includes: an encoding network and a decoding network.

[0087] At S403, the multiple encoded features are fused with the multiple pre-encoded features to obtain multiple second fusion features.

[0088] Exemplarily, for one pre-encoded feature $T_{pe} \in \mathbb{R}^{B \times T \times C}$, the electronic device may use the Encoder network to perform the content correlation encoding processing on the pre-encoded feature $T_{pe} \in \mathbb{R}^{B \times T \times C}$ to obtain the corresponding encoded feature $\Phi(T_{pe})$, and fuse the encoded feature $\Phi(T_{pe})$ with the pre-encoded feature $T_{pe} \in \mathbb{R}^{B \times T \times C}$ to obtain a second fusion feature $T_{en} \in \mathbb{R}^{B \times T \times C}$. The second fusion feature $T_{en} \in \mathbb{R}^{B \times T \times C}$ may be expressed as the following formula (3):

$$T_{en} = \Phi(T_{pe}) + T_{pe} \quad (3),$$

where $\Phi(\cdot)$ represents the content correlation encoding processing.

[0089] At S404, temporal pooling processing is performed on the multiple pre-encoded features to obtain an average feature.

[0090] In some embodiments, before the content correlation encoding processing is performed on the obtained multiple first fusion features, the electronic device may reduce the dimensions of each first fusion feature to the preset dimensions, so as to obtain multiple pre-encoded features; thereafter, the electronic device inputs the multiple pre-encoded features into the encoding network to perform the content correlation encoding processing to each pre-encoded feature, so as to obtain multiple encoded features, and fuses each encoded feature with one corresponding pre-encoded feature to obtain one corresponding second fusion feature. The electronic device may also perform the temporal pooling processing on the multiple pre-encoded features, so as to obtain an average feature corresponding to the multiple pre-encoded features. Exemplarily, for multiple pre-encoded features $T_{pe} \in \mathbb{R}^{B \times T \times C}$, the electronic device may perform the temporal pooling processing on the multiple pre-encoded features, so as to obtain a corresponding average feature value $T_{avg} = \rho_t(T_{pe}) \in \mathbb{R}^{B \times 1 \times C}$, where $\rho_t(\cdot)$ represents the temporal pooling processing.

[0091] Exemplarily, the electronic device may use a third MLP network to reduce the dimensions of each first fusion feature to the preset dimensions. The network structure of the third MLP network is different from the second MLP network and the first MLP network. For example, for each first fusion feature $\hat{T}_i$, the electronic device may use the third MLP network to reduce the dimensions of the first fusion feature $\hat{T}_i$ to C dimensions, and obtain the corresponding pre-encoded feature $T_{pe} \in \mathbb{R}^{B \times T \times C}$.

[0092] In the embodiments of the present disclosure, the dimensions of each first

fusion feature is reduced to the preset dimensions, so as to obtain multiple pre-encoded features; in this way, unimportant features can be removed and important features are retained, which reduces the complexity of the features input into the transformation network, thus reducing the computation amount of the transformation network.

- [0093] At S304, content correlation decoding processing is performed on the multiple second fusion features and the average feature based on the decoding network to obtain the output feature.
- [0094] The electronic device may input the obtained multiple second fusion features and the average feature into the decoding network, and perform the content correlation decoding processing on the average feature and each second fusion feature through the decoding network, so as to obtain the output feature. The average feature represents the feature value of the whole content of the video to be detected, which may be taken as query of a decoder. By referring to the feature value, the decoder finally finds the features that are inconsistent in content. Exemplarily, the decoding network is a Decoder network in the Transformer network.
- [0095] In some embodiments, as shown in FIG. 10, S304 may be implemented by S3041 to S3044, which will be illustrated in combination with the steps shown in FIG. 10.
- [0096] At S3041, first content correlation decoding processing is performed on multiple second fusion features and the average feature based on a first decoding network to obtain a first query matrix, a first key matrix and a first value matrix, and a first key-value matrix is obtained according to the first query matrix and the first key matrix. The decoding network includes: a first decoding network and a second decoding network.
- [0097] The electronic device may take the average feature as query of the first decoding network to perform the first content correlation decoding processing on the average feature and each second fusion feature to obtain the first query matrix, the first key matrix and the first value matrix, and determine the first key-value matrix according to the obtained first query matrix and first key matrix.
- [0098] In some embodiments, S3041 may be implemented in the following mode: respective products of multiple second fusion features and both a first mapping matrix and a second mapping matrix corresponding to the first decoding network are determined according to the first decoding network to obtain the first key matrix and the first value matrix; the product of the average feature and a third mapping matrix corresponding to the first decoding network is determined to obtain the first query matrix; and the product of the first query matrix and the first key matrix is determined, the first key-value matrix is obtained based on the product of the first query matrix and the first key matrix.
- [0099] Here, the first decoding network may be the first layer of Decoder network, the first mapping matrix corresponding to the first decoding network may be one of the mapping matrix P_V or the mapping matrix P_K , and the second mapping matrix may be the other of the mapping matrix P_V or the mapping matrix P_K . The third mapping matrix corresponding to the first decoding network may be the mapping matrix P_Q . The electronic device may obtain by calculation the product of multiple second fusion features T_{en} and the mapping matrix P_K , and take the product as the first key matrix K_{de} ; the electronic device may obtain by calculation the product of multiple second fusion features T_{en} and the mapping matrix P_V , and take the product as the first value

matrix V_{de} ; and the electronic device may obtain by calculation the product of the average feature T_{avg} and the mapping matrix P_Q , and take the product as the first query matrix Q_{de} ; then calculate the product of the first query matrix Q_{de} and the first key matrix K_{de} , and determine the first key-value matrix M_{QK} according to the product. The first key matrix K_{de} may be expressed by the following formula (4), the first value matrix V_{de} may be expressed by the following formula (5), the first query matrix Q_{de} may be expressed by the following formula (6), and the first key-value matrix M_{QK} may be expressed by the following formula (7):

$$K_{de} = T_{en} P_K \quad (4),$$

$$V_{de} = T_{en} P_V \quad (5),$$

$$Q_{de} = T_{avg} P_Q \quad (6),$$

$$M_{QK} = \text{Soft max} \left(\frac{Q_{de} K_{de}^T}{\sqrt{C}} \right) \quad (7),$$

where C represents the preset dimensions, and $\text{Soft max}(\cdot)$ represents normalization processing.

[00100] At S3042, the product of the first key-value matrix and the first value matrix is determined to obtain an intermediate feature.

[00101] When the first key-value matrix M_{QK} and the first value matrix V_{de} are obtained, the electronic device may calculate the product of the first key-value matrix M_{QK} and the first value matrix V_{de} , and take the product as the intermediate feature.

[00102] At S3043, second content correlation decoding processing is performed on the multiple second fusion features and the intermediate feature based on the second decoding network to obtain a second query matrix, a second key matrix and a second value matrix, and a second key-value matrix is obtained according to the second query matrix and the second key matrix.

[00103] The electronic device may take the intermediate feature output by the first decoding network as query of the second decoding network to perform the second content correlation decoding processing on the intermediate feature and each second fusion feature, so as to obtain the second query matrix, the second key matrix and the first second matrix, and determine the second key-value matrix according to the obtained second query matrix and second key matrix.

[00104] Here, the second decoding network may be the second layer of Decoder network, and the two layers of Decoder network have the same network structure. Likewise, the second decoding network correspondingly has a fourth mapping matrix, a fifth mapping matrix and a sixth mapping matrix. The fourth mapping matrix may be one of the mapping matrix P_V or the mapping matrix P_K , and the fifth mapping matrix may be the other of the mapping matrix P_V or the mapping matrix P_K . The sixth

mapping matrix may be the mapping matrix P_Q . Moreover, the principle of calculating the second query matrix Q'_{de} , the second key matrix K'_{de} and the second value matrix V'_{de} according to multiple second fusion features T_{en} and the intermediate feature T_z is the same as the principle of the above formulas (4), (5) and (6), the principle of calculating the second key-value matrix M'_{QK} according to the second query matrix Q'_{de} and the second key matrix K'_{de} is the same as the principle of the above formula (7); however, the difference is that the second query matrix Q'_{de} is calculated using the intermediate feature T_z and the sixth mapping matrix P_Q . It is to be noted that the values of P_V , P_K and P_Q corresponding to the second decoding network are different from or the same as the values of P_V , P_K and P_Q corresponding to the first decoding network.

[00105] At S3044, the product of the second key-value matrix and the second value matrix is determined to obtain the output feature.

[00106] When the second key-value matrix M'_{QK} and the second value matrix V'_{de} are obtained, the electronic device may calculate the product of the second key-value matrix M'_{QK} and the second value matrix V'_{de} , and take the product as the output feature.

[00107] FIG. 11 illustrates another flowchart of an exemplary video quality detection method provided by an embodiment of the present disclosure. As shown in FIG. 11, the video quality detection method may be divided into two parts, namely Temporal Distortion Encoder (TDE) and Temporal Content Transformer (TCT). For the video I to be detected, in the TDE part, the multilevel features of the video I to be detected are extracted through the backbone network including five network layers, and the GAP processing is performed on each extracted multilevel feature to obtain the pooled feature; then, the difference feature between each two adjacent pooled features is calculated to obtain the temporal difference feature, the obtained temporal difference feature is fused with the pooled feature to obtain the first fusion feature, and the regression processing is performed on first fusion features through the first MLP network to obtain the temporal distortion degree, thus completing the TDE part and obtaining the temporal distortion degree of the TDE part and the first fusion feature. Thereafter, in the TCT part, the third MLP network is used to reduce the dimensions of the first fusion feature obtained from the TDE part to obtain the pre-encoded feature, and the temporal pooling processing is performed on the pre-encoded feature to obtain the average feature; at the same time, the pre-encoded feature is also taken as input of the Encoder network Φ in the Transformer network, the content correlation encoding processing is performed on the pre-encoded feature through the Encoder network to obtain the encoded feature, and the encoded feature is fused with the corresponding pre-encoded feature to obtain the second fusion feature; then, the second fusion feature is taken as input of the Decoder network in the Transformer network, the key matrix (K in FIG. 11), the query matrix (Q) and the value matrix (V) corresponding to the first Decoder layer are determined through the first Decoder layer in the Decoder network, the key-value matrix (QK) corresponding to the

first Decoder layer is obtained according to the key matrix (K) and the query matrix (Q), the output feature corresponding to the first Decoder layer is calculated according to the value matrix (V) and the key-value matrix (QK), the output feature and the encoded feature corresponding to the first Decoder layer are also taken as input of a second Decoder layer, and the second Decoder layer outputs its output feature according to the same decoding principle as the first Decoder layer; then, the regression processing is performed on the output feature of the second Decoder layer through the second MLP network, thus completing the TCT part and obtaining the content correlation degree of the TCT part; finally, the sum of the temporal distortion degree obtained from the TDE part and the content correlation degree obtained from the TCT part is taken as the quality parameter of the video I to be detected.

[00108] In the embodiments of the present disclosure, because the feature which is used for representing the temporal distortion content in the video to be detected can be obtained, and the feature which is used for representing the content whose correlation with the whole content of the video to be detected is low can be obtained, the video quality of the video to be detected in terms of temporal distortion and content inconsistency between video frames can be determined based on this. Compared with video quality detection based on individual video frames, the detection dimensions of video quality detection are further increased, thus further improving the accuracy of the detected video quality.

[00109] In some embodiments, as shown in FIG. 12, before S1031, S501 to S505 may also be executed, which will be illustrated in combination with the steps shown in FIG. 12.

[00110] At S501, feature extraction of different levels is performed on each video in a training video set to obtain multiple multilevel training features corresponding to each video.

[00111] In the embodiments of the present disclosure, the training video set is a training set composed of multiple videos, and the multiple videos may be of different types or the same type, which is not limited by the embodiments of the present disclosure. The electronic device may use the video in the training video set to train an initial transformation network. For each video, the electronic device may use a pre-trained backbone network to perform feature extraction of multiple levels on each video frame in each video, so as to obtain multiple features of different levels of each video frame in each video, and fuse the multiple features of different levels to obtain the multilevel feature of each video frame in each video, so as to obtain multiple multilevel features for training (hereinafter called multiple multilevel training features) corresponding to each video used for training an initial third MLP network, the initial transformation network and an initial second MLP network.

[00112] Exemplarily, the electronic device may use the pre-trained VSwin-T as the backbone network to perform feature extraction of five levels on each video frame in each video, so as to obtain five features of different levels, and fuse the five features of different levels to obtain multiple multilevel training features corresponding to each video.

[00113] At S502, the temporal difference feature between adjacent multilevel training features of the multiple multilevel training features is determined, and multiple fusion training features are obtained based on the temporal difference feature and the multiple multilevel training features.

[00114] In the embodiments of the present disclosure, the electronic device may first

perform GAP processing on each multilevel training feature M_i^l to pool M_i^l into one-dimensional features, combine the one-dimensional features into a multilevel feature vector to obtain the corresponding pooled feature $T_i' = \bigoplus_{l=0}^4 \text{GAP}(M_i^l)$, and take the difference feature $T_i' - T_{i+1}'$ between each group of adjacent pooled features T_i' and T_{i+1}' as an obtained temporal difference feature. Moreover, when the temporal difference feature $T_i' - T_{i+1}'$ between the group of adjacent pooled features T_i' and T_{i+1}' is obtained, the electronic device fuses the temporal difference feature $T_i' - T_{i+1}'$ and the pooled feature T_i' to obtain the corresponding fusion training feature $\hat{T}_i'$.

[00115] At S503, a first probability value and a second probability value are determined from a first preset value range.

[00116] At S504, feature sampling is performed on the multiple fusion training features to obtain a sampling feature based on a magnitude relationship between the first probability value and a first preset threshold, as well as a magnitude relationship between the second probability value and a second preset threshold.

[00117] In the embodiments of the present disclosure, when multiple fusion training features corresponding to a video in the training video set are obtained, the electronic device may first determine two probability values from the first preset value range, judge the magnitude relationship between the two probability values and the corresponding two preset thresholds to obtain a judgment result, determine according to the judgment result a sampling mode corresponding to the judgment result, and use the sampling mode to perform feature sampling on multiple fusion training features corresponding to the video to obtain the sampling feature.

[00118] Here, the first preset value range, the first preset threshold, and the second preset threshold may be set according to actual needs. The first probability value and the second probability value may be the same or different. Exemplarily, the first preset value range may be [0, 1], the first preset threshold may be 0.75, and the second preset threshold may be 0.25.

[00119] Here, the electronic device may randomly select the first probability value p_1 and the second probability value p_2 from the first preset value range, judge the magnitude relationship between the first probability value p_1 and the first preset threshold $p_{t,1}$, and judge the magnitude relationship between the second probability value p_2 and the second preset threshold $p_{t,2}$ to obtain the judgment result, determine according to the judgment result the sampling mode corresponding to the judgment result, and use the sampling mode to perform feature sampling to multiple fusion training features corresponding to the video to obtain the sampling feature.

[00120] In some embodiments, as shown in FIG. 13, S504 may be implemented by S5041, S5042 or S5044, which will be illustrated in combination with the steps shown in FIG. 13.

[00121] At S5041, when the first probability value is less than or equal to the first preset threshold, and the second probability value is less than or equal to the second preset threshold, feature sampling is performed on the multiple fusion training features in

a first sampling mode to obtain an intermediate sampling feature, and feature sampling is performed on the intermediate sampling feature in a second sampling mode to obtain the sampling feature.

[00122] In the present disclosure, for multiple fusion training features corresponding to a video, when the first probability value is less than or equal to the first preset threshold, and the second probability value is less than or equal to the second preset threshold, the electronic device may use the first sampling mode to sample multiple fusion training features corresponding to the video to obtain the intermediate sampling feature, and on the basis of this, use the second sampling mode to continue to sample the intermediate sampling feature, and finally obtain the sampling feature, thus completing the sampling of multiple fusion training features corresponding to the video.

[00123] In some embodiments, the first sampling mode is to collect features except the first features, the number of which is a first value, in the features; and the second sampling mode is to divide the features into some parts, the number of which is a second value, and to sample the features from each and every part. FIG. 14A and FIG. 14B illustrate principle diagrams of a sampling mode provided by an embodiment of the present disclosure. FIG. 14A illustrates a sampling principle of the first sampling mode. As shown in FIG. 14A, in a general situation that the features to be sampled are N fusion

training features ($\hat{T}_0' \sim \hat{T}_{N-1}'$) corresponding to a video, the fusion training features except the first f fusion training features in the N fusion training features may be taken as the sampling features. FIG. 14B illustrates a sampling principle of the second sampling mode. As shown in FIG. 14B, the features to be sampled may be divided into S_0 parts, and the features are sampled from each and every part to obtain the sampling features.

[00124] In some embodiments, S5041 may be implemented in the following mode: the first value is determined from a second preset value range, the first fusion training features, the number of which is the first value, in multiple fusion training features are discarded, and the remaining fusion training features are taken as the intermediate sampling features; a second value is determined from a third preset value range, the intermediate sampling features are divided into the second number of parts, and the features in each and every part are sampled to obtain the sampling feature.

[00125] Here, the second preset value range may be set according to actual needs, and the first value may be an integer f which is randomly determined from the second preset value range. Exemplarily, the second preset value range may be [0, F], and F is 4.

[00126] Here, the third preset value range may be set according to actual needs, and the second value may be an integer S_0 randomly determined from the third preset value range. Exemplarily, the third preset value range may be [$S_{\min}$, N/2], where N refers to all fusion training features corresponding to a video used for training, and $S_{\min}$ may be determined according to actual needs, for example, it may be 1.

[00127] In the embodiments of the present disclosure, for all fusion training features (for example, N fusion training features) corresponding to a video, when the first

probability value P_1 is less than or equal to the first preset threshold $P_{t,1}$, the electronic device may randomly determine an integer f from the second preset value range [0, F], discard the first f fusion training features in the N fusion training features, and take the fusion training features from the (f+1)-th fusion training feature to the N-th fusion training feature as the intermediate sampling features, so as to obtain the intermediate

sampling features $\tau_A = \{\hat{T}_i \mid_{i=0}^{M-1}\}$, where $M=N-f$. Moreover, based on this, when the

second probability value p_2 is also less than or equal to the second preset threshold $p_{t,2}$, the electronic device may randomly determine an integer S_0 from the third preset value range $[S_{\min}, N/2]$, divide the intermediate sampling features $\tau_A = \{\hat{T}_i \mid_{i=0}^{M-1}\}$ into S_0 parts, and sample one or more features from each part, thus obtaining the sampling features τ_B . The sampling features τ_B may be expressed by the following formula (8):

$$\tau_B = \left\{ \hat{T} \mid_{\mu_{\text{int}}(\frac{j*M}{S_0}, \frac{(j+1)*M}{S_0})} \mid_{j=0}^{S_0-1} \right\}, p_1 \leq p_{t,1}, p_2 \leq p_{t,2} \quad (8);$$

where $M = N - f$, $S_0 \sim \mu_{\text{int}}(S_{\min}, N/2)$ represents that S_0 is a randomly determined

integer in $[S_{\min}, N/2]$, and $\left\{ \hat{T} \mid_{\mu_{\text{int}}(\frac{j*M}{S_0}, \frac{(j+1)*M}{S_0})} \mid_{j=0}^{S_0-1} \right\}$ represents all the features sampled from S_0 parts of features.

[00128] At S5042, when the first probability value is less than or equal to the first preset threshold, and the second probability value is greater than the second preset threshold, feature sampling is performed on the multiple fusion training features in the first sampling mode to obtain a sampling feature.

[00129] In the embodiments of the present disclosure, for multiple fusion training features corresponding to a video, when the first probability value is less than or equal to the first preset threshold, and the second probability value is greater than the second preset threshold, the electronic device may use the first sampling mode to sample multiple fusion training features corresponding to the video to obtain the sampling feature, thus completing the sampling of multiple fusion training features corresponding to the video.

[00130] In the embodiments of the present disclosure, for N fusion training features

corresponding to a video, when the first probability value p_1 is less than or equal to the

first preset threshold $p_{t,1}$, the electronic device may randomly determine an integer f from the second preset value range $[0, F]$, discard the first f fusion training features in the N fusion training features, and take the fusion training features from the $(f+1)$ -th fusion training feature to the N -th fusion training feature as sampling features to obtain the sampling features $\tau_A = \{\hat{T}_i \mid_{i=0}^{M-1}\}$. Moreover, based on this, when the second

probability value p_2 is greater than the second preset threshold $p_{t,2}$, the electronic device may directly take the sampling features $\tau_A = \{\hat{T}_i \mid_{i=0}^{M-1}\}$ as the final sample feature τ_B . The sampling features τ_B may be expressed by the following formula (9):

$$\tau_B = \tau_A = \{\hat{T}_i \mid_{i=0}^{M-1}\}, p_1 \leq p_{t,1}, p_2 > p_{t,2} \quad (9);$$

where, $M = N - f$.

[00131] At S5043, when the second probability value is greater than the first preset threshold, and the second probability value is less than or equal to the second preset threshold, feature sampling is performed on the multiple fusion training features in the second sampling mode to obtain the sampling feature.

[00132] In the embodiments of the present disclosure, for multiple fusion training features corresponding to a video, when the first probability value is greater than the first preset threshold, and the second probability value is less than or equal to the second preset threshold, the electronic device may use the second sampling mode to sample multiple fusion training features corresponding to the video to obtain the sampling feature, thus completing the sampling of multiple fusion training features corresponding to the video.

[00133] In the embodiments of the present disclosure, for N fusion training features corresponding to a video, when the first probability value P_1 is greater than the first preset threshold $P_{t,1}$, and the second probability value P_2 is less than or equal to the second preset threshold $P_{t,2}$, the electronic device may randomly determine an integer S_0 from the third preset value range $[S_{\min}, N/2]$, equally divide the N fusion training features into S_0 parts, and sample one or more features in each and every part, thus obtaining sampling features τ_B . The sampling features τ_B may be expressed by the following formula (10):

$$\tau_B = \left\{ \hat{T}''_{\mu_{int}(\frac{j*M}{S_0}, \frac{(j+1)*M}{S_0})} \Big|_{j=0}^{S_0-1} \right\}, p_1 > P_{t,1}, p_2 \leq P_{t,2}, S_0 \sim \mu_{int}(S_{\min}, N/2) \quad (10),$$

where τ_B represents the sampling features, $\left\{ \hat{T}''_{\mu_{int}(\frac{j*M}{S_0}, \frac{(j+1)*M}{S_0})} \Big|_{j=0}^{S_0-1} \right\}$ represents all features sampled from S_0 parts, and $M=N$.

[00134] At S5044, when the second probability value is greater than the first preset threshold, and the second probability value is greater than the second preset threshold, all features in multiple fusion training features are collected to obtain the sampling feature.

[00135] In the embodiments of the present disclosure, for multiple fusion training features corresponding to a video, when the first probability value is greater than the first preset threshold, and the second probability value is greater than the second preset threshold, the electronic device may take all the fusion training features corresponding to the video as sampling features, thus completing the sampling of multiple fusion training features corresponding to the video.

[00136] In the embodiments of the present disclosure, for N fusion training features corresponding to a video, when the first probability value P_1 is greater than the first preset threshold $P_{t,1}$, and the second probability value P_2 is greater than the second preset threshold $P_{t,2}$, the electronic device may take all the N fusion training features as

sampling features.

- [00137] In the embodiments of the present disclosure, through the above sampling modes, the content learned by the transformation network may be more diverse, thus improving the stability of the trained transformation network.
- [00138] At S505, an initial third MLP network, an initial transformation network and an initial second MLP network are trained by using the sampling feature and the loss function, so as to obtain the third MLP network, the transformation network and the second MLP network.
- [00139] When the sampling feature is obtained, the electronic device may use the sampling feature and the loss function to train the initial third MLP network, the initial transformation network and the initial second MLP network, until the trained third MLP network, transformation network and second MLP network are obtained. Exemplarily, the initial transformation network may be an untrained Transformer network, the initial second MLP network may be an untrained second MLP network, and the initial third MLP network may be an untrained third MLP network.
- [00140] In some embodiments, the electronic device may use the initial third MLP network to reduce the dimensions of each sampling feature to the preset dimensions (C dimensions) to obtain an initial pre-encoded feature, use the initial transformation network to perform the content correlation processing on each initial pre-encoded feature to obtain an initial output feature, and use the initial second MLP network to perform regression processing on the initial output feature to obtain an initial content correlation degree. The electronic device trains and adjusts the initial third MLP network, the initial transformation network and the initial second MLP network based on the initial content correlation degree, the preset content correlation degree and the loss function, until the third MLP network, the transformation network and the second MLP network are obtained with final loss being less than a preset loss threshold.
- [00141] Here, the electronic device may use the initial third MLP network and the initial transformation network to perform dimension reduction, pooling, encoding and decoding on each sampling feature through the methods described in S401 to S403 and S304, and S3041 to S3044, and finally obtain the feature after decoding, that is, an initial output feature; then, the electronic device uses the initial second MLP network to regress the initial output feature to an initial content correlation degree, and then uses the loss function to calculate current training loss according to the initial content correlation degree and the preset content correlation degree, and adjusts network parameters of the initial third MLP network, the initial transformation network and the initial second MLP network based on the current training loss, thus completing the current training process. After that, the electronic device continues to use the sampling feature and the adjusted network parameters to perform the next iterative training to the initial third MLP network, the initial transformation network and the initial second MLP network, until the preset training condition is met, for example, when the final loss is less than the preset loss threshold, the third MLP network, the transformation network and the second MLP network are obtained.
- [00142] In some other embodiments, the initial third MLP network, the initial transformation network and the initial second MLP network may also be separately trained to obtain the third MLP network, the transformation network and the second MLP network.
- [00143] The present disclosure also provides a virtual label display apparatus. FIG. 15 illustrates a structural diagram of a video quality detection apparatus provided by an

embodiment of the present disclosure. As shown in FIG. 15, the video quality detection apparatus 1 may include: an extraction unit 11 configured to perform feature extraction of different levels on the video to be detected including multiple video frames to obtain multiple multilevel features corresponding to the multiple video frames; and a determination unit 12 configured to determine the temporal difference feature between the adjacent multilevel features of multiple multilevel features, and obtain multiple first fusion features based on the temporal difference feature and multiple multilevel features, the temporal difference feature being used for representing the temporal distortion content in the video to be detected, and obtain the quality parameter for reflecting the video quality of the video to be detected based on the multiple first fusion features.

[00144] In some embodiments of the present disclosure, the determination unit 12 is further configured to determine the temporal distortion degree among the multiple video frames based on the multiple first fusion features, and take the temporal distortion degree as the quality parameter for reflecting the video quality of the video to be detected.

[00145] In some embodiments of the present disclosure, the determination unit 12 is further configured to perform the content correlation processing on the multiple first fusion features based on the transformation network to obtain an output feature, and obtain the quality parameter for reflecting the video quality of the video to be detected based on the multiple first fusion features and the output feature. The output feature is used for representing the screened content, whose correlation with the whole content is lower than the remaining content in the whole content, in the whole content of the video to be detected, and the remaining content is the content in the whole content except the screened content.

[00146] In some embodiments of the present disclosure, the determination unit 12 is further configured to determine the temporal distortion degree among the multiple video frames based on the multiple first fusion features, and determine the content correlation degree among the multiple video frames based on the output feature; and obtain the quality parameter for reflecting the video quality of the video to be detected based on the temporal distortion degree and the content correlation degree.

[00147] In some embodiments of the present disclosure, the determination unit 12 is further configured to take the sum of the temporal distortion degree and the content correlation degree as the quality parameter for reflecting the video quality of the video to be detected.

[00148] In some embodiments of the present disclosure, the determination unit 12 is further configured to perform regression processing to the multiple first fusion features to obtain the temporal distortion degree among the multiple video frames; and perform regression processing on the output feature to obtain the content correlation degree among the multiple video frames.

[00149] In some embodiments of the present disclosure, the determination unit 12 is further configured to perform the content correlation encoding processing on the multiple first fusion features based on the encoding network to obtain multiple encoded features, each encoded feature representing the correlation between a corresponding first fusion feature and other first fusion features except this first fusion feature in the multiple first fusion features; fuse the multiple encoded features with the multiple first fusion features to obtain multiple second fusion features; perform the temporal pooling processing on the multiple first fusion features to obtain the average feature; and based on the decoding network, perform content correlation decoding processing on the multiple second fusion features and the average feature to obtain the output feature.

- [00150] In some embodiments of the present disclosure, the determination unit 12 is further configured to, before performing content correlation encoding processing on the multiple first fusion features based on the encoding network to obtain multiple encoded features, reduce the dimensions of each first fusion feature of the multiple first fusion features to the preset dimensions to obtain multiple pre-encoded features; and perform content correlation encoding processing on the multiple pre-encoded features based on the encoding network to obtain the multiple encoded features.
- [00151] In some embodiments of the present disclosure, the decoding network may include: a first decoding network and a second decoding network. The determination unit 12 is further configured to perform the first content correlation decoding processing on the multiple second fusion features and the average feature based on the first decoding network to obtain the first query matrix, the first key matrix and the first value matrix, and obtain the first key-value matrix according to the first query matrix and the first key matrix; determine the product of the first key-value matrix and the first value matrix to obtain the intermediate feature; perform the second content correlation decoding processing on the multiple second fusion features and the intermediate feature based on the second decoding network to obtain the second query matrix, the second key matrix and the second value matrix, and obtain the second key-value matrix according to the second query matrix and the second key matrix; and determine the product of the second key-value matrix and the second value matrix to obtain the output feature.
- [00152] In some embodiments of the present disclosure, the determination unit 12 is further configured to determine, through the first decoding network, respective products of the multiple second fusion features and both a first mapping matrix and a second mapping matrix that correspond to the first decoding network to obtain the first key matrix and the first value matrix; determine the product of the average feature and a third mapping matrix corresponding to the first decoding network to obtain the first query matrix; and determine the product of the first query matrix and the first key matrix, and obtain the first key-value matrix based on the product of the first query matrix and the first key matrix.
- [00153] In some embodiments of the present disclosure, the determination unit 12 is further configured to perform GAP processing on each multilevel feature of multiple multilevel features to obtain multiple pooled features; and determine the difference feature between the adjacent pooled features of the multiple pooled features to obtain the temporal difference feature.
- [00154] In some embodiments of the present disclosure, the extraction unit 11 is further configured to perform feature extraction of different levels on the video to be detected including multiple video frames by using the pre-trained backbone network, so as to obtain multiple features of different levels of each video frame at multiple levels; and perform feature fusion on the multiple features of different levels of each video frame to obtain multiple multilevel features corresponding to the multiple video frames.
- [00155] In some embodiments of the present disclosure, the determination unit 12 is further configured to perform regression processing on the multiple first fusion features by using a first MLP network to obtain the temporal distortion degree among the multiple video frames; and perform regression processing on the output feature by using a second MLP network to obtain the content correlation degree among the multiple video frames.
- [00156] In some embodiments of the present disclosure, the determination unit 12 is further configured to reduce the dimensions of each first fusion feature of the multiple first fusion features by using the third MLP network to the preset dimensions, so as to

obtain the multiple pre-encoded features.

[00157] In some embodiments of the present disclosure, the extraction unit 11 is further configured to, before performing content correlation processing on the multiple first fusion features based on the transformation network to obtain the output feature, perform feature extraction of different levels on each video in the training video set to obtain multiple multilevel training features corresponding to each video. The determination unit 12 is further configured to determine a temporal difference feature between adjacent multilevel training features of the multiple multilevel training features, and obtain multiple fusion training features based on the temporal difference feature and the multiple multilevel training features. The apparatus 1 may further include a sampling unit, which is configured to determine a first probability value and a second probability value from the first preset value range; and perform feature sampling on the multiple fusion training features to obtain the sampling feature based on the magnitude relationship between the first probability value and the first preset threshold, as well as the magnitude relationship between the second probability value and the second preset threshold. The apparatus 1 may further include a training unit, which is configured to train the initial third MLP network, the initial transformation network and the initial second MLP network by using the sampling feature and the loss function, so as to obtain the third MLP network, the transformation network and the second MLP network.

[00158] In some embodiments of the present disclosure, the sampling unit is further configured to: when the first probability value is less than or equal to the first preset threshold, and the second probability value is less than or equal to the second preset threshold, perform feature sampling on the multiple fusion training features in a first sampling mode to obtain an intermediate sampling feature, and perform feature sampling on the intermediate sampling feature in a second sampling mode to obtain the sampling feature; and/or, when the first probability value is less than or equal to the first preset threshold, and the second probability value is greater than the second preset threshold, perform feature sampling on the multiple fusion training features in the first sampling mode to obtain the sampling feature; and/or, when the second probability value is greater than the first preset threshold, and the second probability value is less than or equal to the second preset threshold, perform feature sampling on the multiple fusion training features in the second sampling mode to obtain the sampling feature; and/or, when the second probability value is greater than the first preset threshold, and the second probability value is greater than the second preset threshold, collect all features in the multiple fusion training features to obtain the sampling feature.

[00159] In some embodiments of the present disclosure, the first sampling mode includes to collect features except the first features, the number of which is a first value, in the features; and the second sampling mode includes to divide the features into a second number of parts and sample the features in each part. The sampling unit is further configured to: determine the first value from a second preset value range, discard first fusion training features, the number of which is the first value, in the multiple fusion training features, and take the remaining fusion training features as intermediate sampling features; and determine a second value from the third preset value range, divide the intermediate sampling features into a second number of parts, and sample features in each part to obtain the sampling feature.

[00160] In some embodiments of the present disclosure, the training unit is further configured to: perform the content correlation processing on the sampling feature by using the initial transformation network to obtain the initial output feature; reduce the

dimensions of each sampling feature to the preset dimensions by using the initial third MLP network to obtain an initial pre-encoded feature; perform content correlation processing on the initial pre-encoded feature by using the initial transformation network to obtain an initial output feature; perform regression processing on the initial output feature by using the initial second MLP network to obtain the initial content correlation degree; train and adjust the initial third MLP network, the initial transformation network and the initial second MLP network based on the initial content correlation degree, the preset content correlation degree and the loss function, until the third MLP network, the transformation network and the second MLP network are obtained with final loss being less than a preset loss threshold.

[00161] The embodiments of the present disclosure further provide an electronic device. FIG. 16 illustrates a structural diagram of an electronic device provided by an embodiment of the present disclosure. As shown in FIG. 16, the electronic device may include: a memory 21 and a processor 22. The memory 21 and the processor 22 are connected via a bus 23. The memory 21 is configured to store an executable computer program. The processor 22 is configured to implement, when executing the executable computer program stored in the memory 21, the video quality detection method provided by the embodiments of the present disclosure.

[00162] The embodiments of the present disclosure provide a computer-readable storage medium, in which a computer program is stored. The computer program is configured to implement, when executed by the processor 22, the video quality detection method provided by the embodiments of the present disclosure.

[00163] In some embodiments of the present disclosure, the storage medium may be a Ferroelectric Random Access Memory (FRAM), a Read-Only Memory (ROM), a Programmable Read-Only Memory (PROM), an Erasable Programmable Read-Only Memory (EPROM), an Electrically Erasable Programmable Read-Only Memory (EEPROM), a flash memory, a magnetic surface storage, compact disc, a Compact Disc Read-Only Memory (CD-ROM) and other memories; the storage medium may also be various devices including one or any combination of the above memories.

[00164] In some embodiments of the present disclosure, the executable instruction may be compiled according to a programming language of any form (including a compiling or interpretive language, or a declarative or procedural language) in form of a program, software, a software module, a script, or a code, and may be deployed according to any form, including deployed as an independent program or deployed as a module, a component, a subroutine or another unit suitable to be used in a computing environment.

[00165] As an example, the executable instruction may but not always correspond to a file in a file system, and may be stored in a part of a file that stores another program or data, for example, stored in one or more scripts in a Hyper Text Markup Language (HTML) document, stored in a single file dedicated to a discussed program, or stored in multiple collaborative files (for example, files storing one or more modules, subprograms or code parts).

[00166] As an example, the executable instruction may be deployed in a computing device for execution, or executed in multiple computing devices at the same place, or executed in multiple computing devices that are interconnected through a communication network at multiple places.

[00167] To sum up, by using the technical solution, because the feature which is used for representing the temporal distortion content in the video to be detected can be obtained, and the feature which is used for representing the content whose correlation

with the whole content of the video to be detected is low can be obtained, the video quality of the video to be detected in terms of temporal distortion between the video frames and content inconsistency can be determined based on this. Compared with video quality detection based on individual video frames, the detection dimension of video quality detection is further increased, thus further improving the accuracy of the detected video quality.

[00168] The above is only the embodiments of the disclosure and not intended to limit the protection scope of the disclosure. Any modifications, equivalent replacements, improvements and the like within the spirit and scope of the present disclosure should fall within the protection scope of the claims of the present disclosure.

CLAIMS

1. A video quality detection method, comprising:
 - performing feature extraction of different levels on a video to be detected comprising a plurality of video frames to obtain a plurality of multilevel features corresponding to the plurality of video frames;
 - determining a temporal difference feature between adjacent multilevel features of the plurality of multilevel features, and obtaining a plurality of first fusion features based on the temporal difference feature and the plurality of multilevel features, wherein the temporal difference feature is used for representing temporal distortion content in the video to be detected; and
 - obtaining a quality parameter for reflecting video quality of the video to be detected based on the plurality of first fusion features.
2. The method of claim 1, wherein obtaining the quality parameter for reflecting the video quality of the video to be detected based on the plurality of first fusion features comprises:
 - determining a temporal distortion degree among the plurality of video frames based on the plurality of first fusion features, and taking the temporal distortion degree as the quality parameter for reflecting the video quality of the video to be detected.
3. The method of claim 1, wherein obtaining the quality parameter for reflecting the video quality of the video to be detected based on the plurality of first fusion features comprises:
 - performing content correlation processing on the plurality of first fusion features based on a transformation network to obtain an output feature, wherein the output feature is used for representing screened content, whose correlation with whole content is lower than remaining content in the whole content, in the whole content of the video to be detected, and wherein the remaining content is content in the whole content except the screened content; and
 - obtaining the quality parameter for reflecting the video quality of the video to be detected based on the plurality of first fusion features and the output feature.
4. The method of claim 3, wherein obtaining the quality parameter for reflecting the video quality of the video to be detected based on the plurality of first fusion features and the output feature comprises:
 - determining the temporal distortion degree among the plurality of video frames based on the plurality of first fusion features, and determining content correlation degree among the plurality of video frames based on the output feature; and
 - obtaining the quality parameter for reflecting the video quality of the video to be detected based on the temporal distortion degree and the content correlation degree.
5. The method of claim 4, wherein obtaining the quality parameter for reflecting the video quality of the video to be detected based on the temporal distortion degree and the content correlation degree comprises:
 - taking sum of the temporal distortion degree and the content correlation degree as the quality parameter for reflecting the video quality of the video to be detected.
6. The method of claim 4, wherein determining the temporal distortion degree among the plurality of video frames based on the plurality of first fusion features, and determining the content correlation degree among the plurality of video frames based on the output feature

comprises:

- performing regression processing on the plurality of first fusion features to obtain the temporal distortion degree among the plurality of video frames; and

- performing regression processing on the output feature to obtain the content correlation degree among the plurality of video frames.

7. The method of any one of claims 3 to 5, wherein the transformation network comprises: an encoding network and a decoding network; wherein performing the content correlation processing to the plurality of first fusion features based on the transformation network to obtain the output feature comprises:

- performing content correlation encoding processing on the plurality of first fusion features based on the encoding network to obtain a plurality of encoded features, wherein each encoded feature of the plurality of encoded features represents correlation between a first fusion feature corresponding to the each encoded feature and other first fusion features except this first fusion feature in the plurality of first fusion features;

- fusing the plurality of encoded features with the plurality of first fusion features to obtain a plurality of second fusion features;

- performing temporal pooling processing on the plurality first fusion features to obtain an average feature; and

- performing content correlation decoding processing on the plurality of second fusion features and the average feature based on the decoding network to obtain the output feature.

8. The method of claim 7, wherein before performing content correlation encoding processing on the plurality of first fusion features based on the encoding network to obtain the plurality of encoded features, the method further comprises:

- reducing dimensions of each first fusion feature of the plurality of first fusion features to preset dimensions to obtain a plurality of pre-encoded features;

wherein performing content correlation encoding processing on the plurality of first fusion features based on the encoding network to obtain the plurality of encoded features comprises:

- performing content correlation encoding processing on the plurality of pre-encoded features based on the encoding network to obtain the plurality of encoded features.

9. The method of claim 7, wherein the decoding network comprises: a first decoding network and a second decoding network; wherein performing content correlation decoding processing on the plurality of second fusion features and the average feature based on the decoding network to obtain the output feature comprises:

- performing first content correlation decoding processing on the plurality of second fusion features and the average feature based on the first decoding network to obtain a first query matrix, a first key matrix and a first value matrix, and obtaining a first key-value matrix according to the first query matrix and the first key matrix;

- determining a product of the first key-value matrix and the first value matrix to obtain an intermediate feature;

- performing second content correlation decoding processing on the plurality of second fusion features and the intermediate feature based on the second decoding network to obtain a second query matrix, a second key matrix and a second value matrix, and obtaining a second key-value matrix according to the second query matrix and the second key matrix; and

- determining a product of the second key-value matrix and the second value matrix to obtain the output feature.

10. The method of claim 9, wherein performing the first content correlation decoding processing on the plurality of second fusion features and the average feature based on the first decoding network to obtain the first query matrix, the first key matrix and the first value matrix, and obtaining the first key-value matrix according to the first query matrix and the first key matrix comprise:

determining, through the first decoding network, a product of the plurality of second fusion features and a first mapping matrix corresponding to the first decoding network to obtain the first key matrix, and determining, through the first decoding network, a product of the plurality of second fusion features and a second mapping matrix corresponding to the first decoding network to obtain the first value matrix;

determining a product of the average feature and a third mapping matrix corresponding to the first decoding network to obtain the first query matrix;

determining a product of the first query matrix and the first key matrix, and obtaining the first key-value matrix based on the product of the first query matrix and the first key matrix.

11. The method of any one of claims 1 to 10, wherein determining the temporal difference feature between the adjacent multilevel features of the plurality of multilevel features comprises:

performing global average pooling processing on each multilevel feature of the plurality of multilevel features to obtain a plurality of pooled features; and

determining a difference feature between adjacent pooled features of the plurality of pooled features to obtain the temporal difference feature.

12. The method of any one of claims 1 to 11, wherein performing feature extraction of different levels on the video to be detected comprising the plurality of video frames to obtain the plurality of multilevel features corresponding to the plurality of video frames comprises:

performing feature extraction of different levels on the video to be detected comprising the plurality of video frames by using a pre-trained backbone network to obtain a plurality of features of different levels of each video frame at a plurality of levels; and

performing feature fusion on the plurality of features of different levels of the each video frame to obtain the plurality of multilevel features corresponding to the plurality of video frames.

13. The method of claim 4 or 6, wherein determining the temporal distortion degree among the plurality of video frames based on the plurality of first fusion features, and determining the content correlation degree among the plurality of video frames based on the output feature comprise:

performing regression processing on the plurality of first fusion features by using a first multilayer perceptron network to obtain the temporal distortion degree among the plurality of video frames; and

performing regression processing on the output feature by using a second multilayer perceptron network to obtain the content correlation degree among the plurality of video frames.

14. The method of claim 8, wherein reducing the dimensions of the each first fusion feature of the plurality of first fusion features to the preset dimensions to obtain the plurality of pre-

encoded features comprises:

reducing the dimensions of the each first fusion feature of the plurality of first fusion features by using a third multilayer perceptron network to the preset dimensions to obtain the plurality of pre-encoded features.

15. The method of any one of claims 3 to 5, 7 to 10, and 14, wherein before performing content correlation processing on the plurality of first fusion features based on the transformation network to obtain the output feature, the method further comprises:

performing feature extraction of different levels on each video in a training video set to obtain a plurality of multilevel training features corresponding to the each video;

determining a temporal difference feature between adjacent multilevel training features of the plurality of multilevel training features, and obtaining a plurality of fusion training features based on the temporal difference feature and the plurality of multilevel training features;

determining a first probability value and a second probability value from a first preset value range;

performing feature sampling on the plurality of fusion training features to obtain a sampling feature based on a magnitude relationship between the first probability value and a first preset threshold, as well as a magnitude relationship between the second probability value and a second preset threshold; and

training an initial third multilayer perceptron network, an initial transformation network and an initial second multilayer perceptron network by using the sampling feature and a loss function to obtain the third multilayer perceptron network, the transformation network and the second multilayer perceptron network.

16. The method of claim 15, wherein performing feature sampling on the plurality of fusion training features to obtain the sampling feature based on the magnitude relationship between the first probability value and the first preset threshold, as well as the magnitude relationship between the second probability value and the second preset threshold comprises at least one of:

in a case where the first probability value is less than or equal to the first preset threshold, and the second probability value is less than or equal to the second preset threshold, performing feature sampling on the plurality of fusion training features in a first sampling mode to obtain an intermediate sampling feature, and performing feature sampling on the intermediate sampling feature in a second sampling mode to obtain the sampling feature;

in a case where the first probability value is less than or equal to the first preset threshold, and the second probability value is greater than the second preset threshold, performing feature sampling on the plurality of fusion training features in the first sampling mode to obtain the sampling feature;

in a case where the second probability value is greater than the first preset threshold, and the second probability value is less than or equal to the second preset threshold, performing feature sampling on the plurality of fusion training features in the second sampling mode to obtain the sampling feature; or

in a case where the second probability value is greater than the first preset threshold, and the second probability value is greater than the second preset threshold, collecting all features in the plurality of fusion training features to obtain the sampling feature.

17. The method of claim 16, wherein the first sampling mode comprises: collecting features

except first features, a number of which is a first value, in the plurality of fusion training features; the second sampling mode comprises: dividing features into a second number of parts, and sampling a feature in each part of the second number of parts; wherein performing feature sampling on the plurality of fusion training features in the first sampling mode to obtain the intermediate sampling feature, and performing feature sampling on the intermediate sampling feature in the second sampling mode to obtain the sampling feature comprise:

determining the first value from a second preset value range, discarding first fusion training features, a number of which is the first value, in the plurality of fusion training features, and taking remaining fusion training features as intermediate sampling features; and

determining a second value of the second number from a third preset value range, dividing the intermediate sampling features into the second number of parts, and sampling the feature in the each part to obtain the sampling feature.

18. The method of any one of claims 15 to 17, wherein training the initial third multilayer perceptron network, the initial transformation network and the initial second multilayer perceptron network by using the sampling feature and the loss function to obtain the third multilayer perceptron network, the transformation network and the second multilayer perceptron network comprises:

reducing dimensions of each sampling feature to preset dimensions by using the initial third multilayer perceptron network to obtain an initial pre-encoded feature;

performing content correlation processing on the initial pre-encoded feature by using the transformation network to obtain an initial output feature;

performing regression processing on the initial output feature by using the initial second multilayer perceptron network to obtain an initial content correlation degree; and

training and adjusting the third multilayer perceptron network, the initial transformation network and the initial second multilayer perceptron network based on the initial content correlation degree, a preset content correlation degree and the loss function, until the third multilayer perceptron network, the transformation network and the second multilayer perceptron network are obtained with final loss being less than a preset loss threshold.

19. A video quality detection apparatus, comprising:

an extraction unit, configured to perform feature extraction of different levels on a video to be detected comprising a plurality of video frames to obtain a plurality of multilevel features corresponding to the plurality of video frames; and

a determination unit, configured to determine a temporal difference feature between adjacent multilevel features of the plurality of multilevel features, and obtain a plurality of first fusion features based on the temporal difference feature and the plurality of multilevel features, and obtain a quality parameter for reflecting the video quality of the video to be detected based on the plurality of first fusion features, wherein the temporal difference feature is used for representing temporal distortion content in the video to be detected.

20. An electronic device, comprising:

a memory, configured to store an executable computer program; and

a processor, configured to implement, when executing the executable computer program stored in the memory, the method of any one of claims 1 to 18.