(19) Europäisches Patentamt
European Patent Office
Office européen des brevets

(11) **EP 4 390 725 A1**

(12) **EUROPEAN PATENT APPLICATION**
published in accordance with Art. 153(4) EPC

(43) Date of publication:
26.06.2024 Bulletin 2024/26

(21) Application number: **22860095.3**

(22) Date of filing: **15.07.2022**

(51) International Patent Classification (IPC):
G06F 16/783 *(2019.01)*

(52) Cooperative Patent Classification (CPC):
G06F 16/7857; G06N 3/04; G06N 3/08; G06V 10/54; G06V 10/778; G06V 20/41; G06V 20/47; G06V 20/48

(86) International application number:
PCT/CN2022/105871

(87) International publication number:
WO 2023/024749 (02.03.2023 Gazette 2023/09)

(84) Designated Contracting States:
AL AT BE BG CH CY CZ DE DK EE ES FI FR GB GR HR HU IE IS IT LI LT LU LV MC MK MT NL NO PL PT RO RS SE SI SK SM TR
Designated Extension States:
BA ME
Designated Validation States:
KH MA MD TN

(30) Priority: **24.08.2021 CN 202110973390**

(71) Applicant: **Tencent Technology (Shenzhen) Company Limited**
Shenzhen, Guangdong, 518057 (CN)

(72) Inventor: **GUO, Hui**
Shenzhen, Guangdong 518057 (CN)

(74) Representative: **Gunzelmann, Rainer et al**
Wuesthoff & Wuesthoff
Patentanwälte und Rechtsanwalt PartG mbB
Schweigerstraße 2
81541 München (DE)

(54) VIDEO RETRIEVAL METHOD AND APPARATUS, DEVICE, AND STORAGE MEDIUM

(57) This application relates to the field of computers, and in particular, to the field of artificial intelligence, and provides a video retrieval method and apparatus, a device, and a storage medium to solve the problems of low quantization efficiency and low accuracy. The method includes: extracting a first quantization feature from an image feature of an input video, identifying a second candidate video with a high category similarity to the input video based on the first quantization feature, and finally determining second candidate video with the high content similarity to the input video as a target video. Since quantization control parameters are adjusted according to a texture feature loss value corresponding to each training sample, a quantization processing sub-model can learn the ranking ability of a texture feature sub-model, which ensures that the ranking effect of two sub-models tend to be consistent. An end-to-end model architecture enables the quantization processing sub-model to obtain the quantization feature based on the image feature, which improves the accuracy of the quantization feature and a recall performance of a video retrieval.

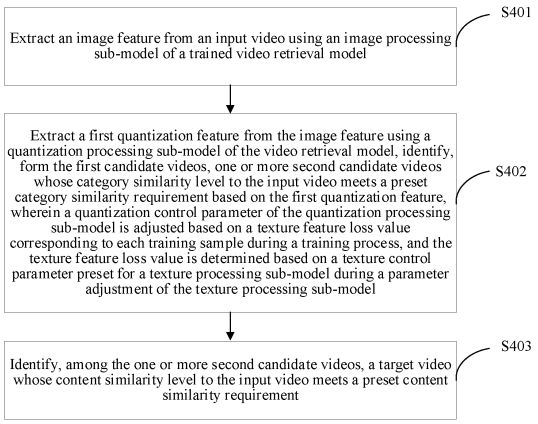

FIG. 4a

EP 4 390 725 A1

Description

CROSS-REFERENCE TO RELATED APPLICATIONS

[0001] This application claims priority to China Patent Application No. 202110973390.0, filed on August 24, 2021, and entitled "VIDEO RETRIEVAL METHOD AND APPARATUS, DEVICE, AND STORAGE MEDIUM", the entirety of which is incorporated herein by reference.

FIELD OF THE TECHNOLOGY

[0002] This application relates to the field of computers, and in particular, to the field of artificial intelligence, and provides a video retrieval method and apparatus, a device, and a storage medium.

BACKGROUND OF THE DISCLOSURE

[0003] In the related art, a quantization feature is usually used as an index label of a video to retrieve and obtain the video. Generally, any one of the following methods is used to obtain a quantization feature:

a first method, the quantization feature is obtained based on K-means clustering algorithm, but for a large-scale sample data clustering, in order to ensure the accuracy of index retrieval, it needs a lot of resources to obtain enough quantization features;

a second method, the quantization feature is obtained based on product quantization (PQ), but the quantization feature obtained by the method will reduce the generation accuracy of quantization features due to the loss in the generation process, thus affecting a match performance of a video retrieval; and

a third method, the quantization feature is obtained based on an in-depth learning neural network, but the neural network extracts an embedding feature of a video image, and then performs feature extraction on the embedding features to obtain the quantization feature, which will reduce the generation accuracy of the quantization features due to the loss in the generation process, thus affecting the match performance of the video retrieval.

SUMMARY

[0004] The embodiments of this application provide a video retrieval method and apparatus, a device, and a storage medium to solve the problems of low quantization efficiency and low accuracy.

[0005] In a first aspect, the embodiments of this application provide a video retrieval method, including:

extracting an image feature from an input video using an image processing sub-model of a trained video retrieval model;

extracting a first quantization feature from the image feature using a quantization processing sub-model of the video retrieval model, wherein the quantization processing sub-model comprises a quantization control parameter adjusted based on a texture feature loss value corresponding to each training sample during a training process, and the texture feature loss value is determined based on a texture control parameter preset for a texture processing sub-model during a parameter adjustment of the texture processing sub-model;

identifying, from first candidate videos, one or more second candidate videos whose category similarity level to the input video meets a preset category similarity requirement based on the first quantization feature; and

identifying, among the one or more second candidate videos, a target video whose content similarity level to the input video meets a preset content similarity requirement.

[0006] In a second aspect, the embodiments of this application also provide a video retrieval apparatus, including:

an image processing unit, configured to extract an image feature from an input video using an image processing sub-model of a trained video retrieval model;

a quantization processing unit, configured to extract a first quantization feature from the image feature using a

quantization processing sub-model of the video retrieval model, identify, from first candidate videos, one or more second candidate videos whose category similarity level to the input video meets a preset category similarity requirement based on the first quantization feature, wherein the quantization processing sub-model comprises a quantization control parameter adjusted based on a texture feature loss value corresponding to each training sample during a training process, and the texture feature loss value is determined based on a texture control parameter preset for a texture processing sub-model during a parameter adjustment of the texture processing sub-model; and

a retrieval unit, configured to identify, among the one or more second candidate videos, a target video whose content similarity level to the input video meets a preset content similarity requirement.

[0007] In a third aspect, the embodiments of this application also provide a computer device including a processor and a memory, the memory storing program codes which, when executed by the processor, cause the processor to perform the operations of any one of the above video retrieval methods.

[0008] In a fourth aspect, the embodiments of this application also provide a computer readable storage medium including program codes which, when executed by a computer device, cause the computer device to perform the operations of any one of the above video retrieval methods.

[0009] In a fifth aspect, the embodiments of this application also provide a computer program product or computer program including computer instructions stored in a computer readable storage medium, which, when read and executed by a processor of a computer device, cause the computer device to perform the operations of any one of the above video retrieval methods.

[0010] The beneficial effects of this application are as follows.

[0011] The embodiments of this application provide a video retrieval method and apparatus, a device, and a storage medium, and the method may include: extracting a first quantization feature from an image feature of an input video, identifying one or more second candidate videos with a high category similarity level to the input video based on the first quantization feature, and finally determining a second candidate video with the high content similarity level to the input video as a target video. Since the quantization control parameter of the quantization processing sub-model is adjusted according to the texture feature loss value corresponding to each training sample, the quantization processing sub-model can learn the ranking ability of the texture feature sub-model, which ensures that ranking effects of two sub-models tend to be consistent, and avoids a random ranking of the quantization processing sub-model due to fixed quantization control parameters. Due to an end-to-end model architecture, the quantization processing sub-model trained by the above-mentioned manner can obtain a quantization feature based on the image feature, which reduces the loss in the process of generating the quantization feature and improves the accuracy of the quantization feature. In addition, the embodiments of this application also optimize the ranking ability of the quantization processing sub-model and further improves a match performance of a video retrieval.

[0012] Other features and advantages of this application will be set forth in the description which follows, and in part will be obvious from the description, or may be learned by practice of this application. The objects and other advantages of this application may be realized and obtained by the structure particularly pointed out in the written specification, claims, and the appended drawings.

BRIEF DESCRIPTION OF THE DRAWINGS

[0013] The accompanying drawings described herein are used to provide a further understanding of this application and constitute a part of this application. The illustrative embodiments of this application and descriptions are used to explain this application and do not constitute undue limitations on this application. In the drawings:

FIG. 1a is a schematic diagram of an application scenario in the embodiments of this application.

FIG. 1b is a schematic diagram of a first display interface according to the embodiments of this application.

FIG. 1c is a schematic diagram of a second display interface according to the embodiments of this application.

FIG. 1d is a schematic architecture diagram of a video retrieval model according to the embodiments of this application.

FIG. 1e is a schematic architecture diagram of a quantization processing model used in the related art.

FIG. 2a is a schematic flow diagram of a training video retrieval model according to the embodiments of this application.

FIG. 2b is a schematic flow diagram of mining a plurality of sample triplets according to the embodiments of this application.

FIG. 2c is a schematic flow diagram of a first generation of a quantization feature loss value according to the embodiments of this application.

FIG. 2d is a schematic flow diagram of a second generation of a quantization feature loss value according to the embodiments of this application.

FIG. 3a is a schematic flow diagram of establishing an index table and a mapping table according to the embodiments of this application.

FIG. 3b is a logical schematic diagram of establishing an index table and a mapping table according to the embodiments of this application.

FIG. 4a is a schematic flow diagram of a video retrieval method according to the embodiments of this application.

FIG. 4b is a logic schematic diagram of a video retrieval method according to the embodiments of this application.

FIG. 5 is a schematic structural diagram of a video retrieval apparatus according to the embodiments of this application.

FIG. 6 is a schematic diagram of a composition structure of a computer device according to the embodiments of this application.

FIG. 7 is a schematic structural diagram of one computing apparatus according to the embodiments of this application.

DESCRIPTION OF EMBODIMENTS

[0014] In order to make the objects, technical solutions and advantages of embodiments of this application clearer, the technical solutions of this application will be described clearly and completely with reference to the accompanying drawings in the embodiments of this application. Obviously, the described embodiments are a part of the technical solutions of this application, not the whole embodiments. Based on the embodiments recorded in this application, all the other embodiments obtained by ordinary technicians in the art without making any inventive effort fall within the scope of protection of the technical solution of this application.

[0015] Some terms in the embodiments of this application including the embodiments of both the claims and the specification (hereinafter referred to as "all embodiments of this application) are explained below to facilitate understanding by those skilled in the art.

[0016] All embodiments of this application relate to the field of artificial intelligence (AI), and are designed based on machine learning (ML) and computer vision (CV) techniques. The solutions provided by all embodiments of this application relate to technologies such as in-depth learning and augmented reality of artificial intelligence, and are further illustrated by the following embodiments.

[0017] All embodiments of this application may be briefly described below.

[0018] All embodiments of this application may provide a video retrieval method and apparatus, a device, and a storage medium to solve the problems of low quantization efficiency and low accuracy. All embodiments of this application can be applied to various types of video retrieval scenes, for example, in video infringement scenarios, and a batch of videos with high content similarity level to the input video are matched by using the video retrieval method provided by all embodiments of the application, and the matched videos are determined as infringing videos.

[0019] The method may include: extracting a first quantization feature from an image feature of an input video, identifying one or more second candidate videos with a high category similarity level to the input video based on the first quantization feature, and finally determining a second candidate video with the high content similarity level to the input video as a target video. Since the quantization control parameter of the quantization processing sub-model is adjusted according to the texture feature loss value corresponding to each training sample, the quantization processing sub-model can learn the ranking ability of the texture feature sub-model, which ensures that ranking effects of two sub-models tend to be consistent, and avoids a random ranking of the quantization processing sub-model due to fixed quantization control parameters. Due to an end-to-end model architecture, the quantization processing sub-model trained by the above-mentioned manner can obtain a quantization feature based on the image feature, which reduces the loss in the process of generating the quantization feature and improves the accuracy of the quantization feature. In addition, all embodiments

of this application may also optimize the ranking ability of the quantization processing sub-model and further improves a match performance of a video retrieval.

[0020] All embodiments of this application will be described below with reference to the drawings in the specification. All embodiments described here is only used to illustrate and explain this application, and are not used to limit this application, and all embodiments of this application and the features in all embodiments can be combined with each other without conflict.

[0021] With reference to the schematic diagrams shown in FIGS 1a and 1b, in an application scenario of all embodiments of this application, two physical terminal devices 110 and one server 130 may be included.

[0022] An object (for example, a user) can log into a video retrieval client through a physical terminal device 110, and a retrieval interface is presented on a display screen 120 of the physical terminal device 110; then, the object inputs a video in a search interface, to cause a video retrieval model running on a server 130 obtains a target video with a higher content similarity level to the input video from a huge video library connected to a background port; after receiving all the target videos returned by the server 130, the physical terminal device 110 presents each of target videos on the display interface of the display screen 120, and at the same time, the user can also view the video details of the selected target video by clicking a page or other gesture operations, and a similar segment or a repeated segment of the target video and the input video will be marked on a progress bar.

[0023] The display interface shown in FIG. 1b presents an excerpt segment of a certain television series, and the color of a corresponding progress bar is white for the played segment; the color of a corresponding progress bar is black for an unplayed segment; in addition, the color of a corresponding progress bar is grey for a similar segment or a repeated segment, in this way, a user can roughly estimate the similarity level between a target video and an input video through the color of the progress bar, to facilitate the user to determine the infringement of video creation.

[0024] The display interface shown in FIG. 1c presents an excerpt segment of a certain television series, and the color of a corresponding progress bar is white for the played segment; the color of a corresponding progress bar is black for an unplayed segment; a triangle mark point or mark points with other shapes are marked on the progress bar for a similar segment or a repeated segment to mark a start point and an end point of these segments, in this way, a user can directly jump to the corresponding scenario by clicking the mark point; likewise, the user can also roughly estimate the similarity level between the target video and the input video through the number of mark points on the progress bar.

[0025] In all embodiments of this application, the physical terminal device 110 can be an electronic device used by a user, and the electronic device can be a computer device such as a personal computer, a mobile phone, a tablet computer, a notebook computer, an electronic book reader, and a smart home.

[0026] Each physical terminal device 110 communicates with the server 130 through a communication network. In all embodiments, the communication network may be a wired or wireless network, and therefore, each physical terminal device 110 may directly or indirectly establish a communication connection with the server 130 through the wired or wireless network, and this application is not limited here.

[0027] The server 130 can be an independent physical server, and can also be a server cluster or distributed system composed of a plurality of physical servers, and can also be a cloud server providing basic cloud computing services, such as a cloud service, a cloud database, a cloud computing, a cloud function, a cloud storage, a network service, a cloud communication, a middle ware service, a domain name service, a security service, a content delivery network (CDN), large data, an artificial intelligence platform, etc., and this application is not limited here.

[0028] A video retrieval model is deployed on the server 130, and as shown in FIG. 1d, the video retrieval model includes an image processing sub-model, a texture processing sub-model and a quantization processing sub-model.

[0029] The image processing sub-model and the texture processing sub-model are in-depth learning network models constructed by using the network architecture of ResNet_101, and are pre-trained based on imageNet. ImageNet is an open-source data set for large-scale general object recognition. There are a lot of pre-marked image data in imageNet, and imageNet probably contains 1000 kinds of image data. Therefore, the in-depth learning network model based on imageNet pre-training has better stability of model parameters and universality of the whole model.

[0030] In addition, a network architecture other than ResNet_101 can be used to build an in-depth learning network model, which can be pre-trained based on other large-scale data sets, such as an in-depth learning network model obtained based on open image pre-training.

[0031] In the video/image processing field, quantization refers to a process of converting the continuous variation interval of a video signal into a series of discrete specified values, that is, a process of converting the continuous variation interval of brightness corresponding to image pixels into discrete specified values. The obtained discrete specified values are called quantization features. In all embodiment of this application, the complex high-dimensional image features are processed by binary quantization using the quantization processing sub-model, and the high-dimensional image feature is compressed into binary codes with a specified number of bits (namely, a quantization feature). In the process of video retrieval, the target video is matched by taking the quantization feature as an index, which greatly reduces the calculation time and complexity, and is more conducive to calculation. Therefore, it is very beneficial to the retrieval of massive data.

[0032] In addition, for binary coding, each bit has a value of 0 or 1, such as a binary code 0100, which compresses

128-dimensional image features into 4 bits.

[0033] Different from the traditional quantization processing model shown in FIG. 1e, the texture processing sub-model and the quantization processing sub-model in all embodiments of this application may be two sub-models which are placed in parallel. The benefit of such deployment is that in the training phase, the quantization control parameter of the quantization processing sub-model may be adjusted according to the texture feature loss value corresponding to each training sample, the quantization processing sub-model learns the ranking ability of the texture feature sub-model, which ensures that the ranking effects of the two sub-models tend to be consistent, and avoids the random ranking of the quantization processing sub-model due to the fixed quantization control parameters.

[0034] In the application stage, compared with the non-end-to-end model architecture shown in FIG. 1e, all embodiments of this application may adopt the end-to-end model architecture, which can obtain a quantization feature based on the image feature, reduce the loss in the process of generating the quantization feature and improve the generation accuracy of the quantization feature. In addition, all embodiments of this application may also optimize the ranking ability of the quantization processing sub-model and further improves the match performance of the video retrieval.

[0035] Moreover, all embodiments of this application may build a video retrieval model based on artificial intelligence technology, and compared with the traditional K-means clustering algorithm, it has better processing speed and match performance and consumes less resources when dealing with large-scale retrieval videos.

[0036] In the training process of the video retrieval model, it is specifically divided into a pre-training stage and a fine-tuning joint learning stage; however, the training data used in the two training stages are the same, and the difference is that the network parameters needing to be learned and the generated loss values in the two training stages are different.

[0037] In the pre-training stage, the parameters of an image processing sub-model to be trained and a texture processing sub-model to be trained are adjusted using an image processing texture feature loss value to obtain a candidate image processing sub-model and a candidate texture processing sub-model; In the fine-tuning joint learning stage, the parameters of the candidate texture processing sub-model are adjusted by using the texture feature loss value, and the parameters of the candidate image processing sub-model and the quantization processing sub-model to be trained are adjusted by using the quantization feature loss value, to obtain the image processing sub-model, the texture processing sub-model, and the quantization processing sub-model.

[0038] To facilitate understanding, the training process of the video retrieval model is introduced with reference to the schematic flow diagram shown in FIG. 2a.

[0039] S201: Obtain a plurality of sample triplets, each sample triplet containing a sample video, a positive label and a negative label associated with the sample video.

[0040] Compared with the traditional quantization processing method, all embodiments of this application uses the labeled training data, the quantization processing sub-model to be trained can learn the positive label and the negative label at the same time, thus improving the match effect of the quantization processing sub-model.

[0041] The positive label refers to a sample video with a higher content similarity level to the sample video, and the negative label refers to a sample video with only a small amount of the same or similar content as the sample video.

[0042] With reference to the schematic flow diagram shown in FIG. 2b, all embodiments of this application can obtain a plurality of sample triplets by executing the following operations:

S2011: Obtain one similar sample set containing a plurality of similar sample pairs, and each similar sample pair contains a sample video and a positive label associated with the sample video.

S2012: Sequentially input each of similar sample pairs in the similar sample set into an image processing sub-model to be trained and a texture processing sub-model to be trained to obtain a texture feature group.

[0043] The image processing sub-model to be trained and the texture processing sub-model to be trained are in-depth learning network models which are pre-trained based on the large-scale general object recognition open-source data set ImageNet. The texture feature groups of each of similar sample pairs can be obtained by sequentially inputting each of similar sample pairs into the above two sub-models, and each texture feature group includes the texture feature of the sample video and the texture feature of the positive label.

S2013: Read a sample video c of one similar sample pair;

S2014: Perform the following operations for each of other similar sample pairs in the similar sample set: obtaining a texture feature distance based on the texture feature of the sample video c and the texture feature of any one of other sample videos in one other similar sample pair.

[0044] In all embodiments, a Euclidean distance between two texture features may be taken as the texture feature distance. The smaller the value of Euclidean distance is, the higher the content similarity level between the two sample

videos is characterized; on the contrary, the larger the value of Euclidean distance is, the lower the content similarity level between the two sample videos is characterized.

[0045] S2015: Arrange each of the other sample videos in order of distance from the texture feature.

[0046] S2016: Determine the other sample videos arranged in a first m as negative labels of the sample videos after removing a first k% of the other sample videos.

[0047] Each of other sample videos is arranged according to the order of the texture feature distance from near to far, and it can be seen from the above-mentioned introduction of the texture feature distance that the content similarity level between the other sample video of the first k% and the sample video is very high. However, the negative label that needs to be mined in all embodiments of this application may be that there are only a few sample videos with the same or similar contents as the sample videos. Obviously, the other sample videos of the first k% does not meet the definition of the negative label and are removed as interference noise. k is a controllable value, and the greater the interference noise is, the larger the value of k is.

[0048] However, other sample videos ranked extremely low do not conform to the definition of negative label because they have almost the same or similar content with the sample videos. Therefore, in all embodiments of this application, other sample videos ranked in the first m after removing the first k% of other sample videos may be determined as the negative labels of the sample videos.

[0049] For example, assuming that similar sample pairs are (sample video 1, sample video 2), Table 1 shows texture feature distances between other sample videos and sample video 1, removing the first k% of other sample videos, such as the sample video 3 and the sample video 6, and removing the sample video 8 and the sample video 9 with extremely low ranking, and the following sample triplets are finally obtained by screening: (sample video 1, sample video 2, other sample video 7), (sample video 1, sample video 2, other sample video 4), and (sample video 1, sample video 2, other sample video 5).

Table 1

Name of other sample videos	Texture feature distance
Other sample video 3	0.5
Other sample video 6	0.55
Other sample video 7	0.8
Other sample video 4	0.83
Other sample video 5	0.9
Other sample video 8	1.2
Other sample video 9	1.5

[0050] S2017: Determine whether all the similar sample pairs in the similar sample sets have been read, and if so, execute operation S2018; otherwise, return operation S2013.

[0051] S2018: Determine whether all the similar sample sets have been read, and if so, output all the sample triplets; otherwise, return operation S2011.

[0052] S202: Read one sample triplet d, take the same as training data, and sequentially input the same into an image processing sub-model to be trained and a texture processing sub-model to be processed to obtain a first texture set.

[0053] S203: Generate an image processing texture feature loss value based on a plurality of first sample texture features contained in a first texture set, and adjust the parameters of the image processing sub-model to be trained and the texture processing sub-model to be trained based on the image processing texture feature loss value.

[0054] An image processing texture feature loss value is generated using the following Formula 1; then, the parameters of the image processing sub-model to be trained and the texture processing sub-model to be trained are adjusted by using the stochastic gradient descent (SGD) method based on the loss value of image processing texture features.

[0055] In Formula 1, L_{em} is an image processing texture feature loss value, x_a is a first sample texture feature of a sample video, x_p is a first sample texture feature of a positive label, and then x_n is a first sample texture feature of a negative label; $\|x_a - x_p\|$ characterizes a texture feature distance between positive sample pairs, then $\|x_a - x_n\|$ characterizes a texture feature distance between negative sample pairs; and margin_em characterizes a texture control parameter, which is a preset threshold used to control the distance between positive and negative samples.

$$L_{\text{em}} = \max(\|x_a - x_p\| - \|x_a - x_n\| + \text{margin\_em}) \quad \text{Formula 1:}$$

S204: Determine whether the image processing texture feature loss value is higher than a preset image processing texture feature loss threshold value, and if so, return operation S202; otherwise, execute operation S205.

S205: Stop iteration training of the sub-model and output the candidate image processing sub-model and the candidate texture processing sub-model obtained in the last iteration.

S206: Read one sample triplet e, take the same as training data, and sequentially input the same into the candidate image processing sub-model, the candidate texture processing sub-model and a quantization processing sub-model to be trained to obtain a second texture set and a quantization feature group.

S207: Generate a texture feature loss value based on a plurality of second sample texture features contained in a second texture set, and adjust a parameter of the candidate texture processing sub-model based on the texture feature loss value.

[0056] A texture feature loss value is generated using the following Formula 2; then, the parameters of the candidate texture processing sub-model are adjusted by using the SGD method based on the texture feature loss value.

[0057] In Formula 2, L_{em}' is an image processing texture feature loss value, x_a' is a second sample texture feature of a sample video, x_p' is a second sample texture feature of a positive label, and then x_n' is a second sample texture feature of a negative label; $\|x_a' - x_p'\|$ characterizes a texture feature distance between positive sample pairs, then $\|x_a' - x_n'\|$ characterizes a texture feature distance between negative sample pairs; and margin_em characterizes a texture control parameter.

$$L_{\text{em}}' = \max(\left\|x_a' - x_p'\right\| - \|x_a' - x_n'\| + \text{margin\_em}) \quad \text{Formula 2:}$$

S208: Generate a quantization feature loss value based on a plurality of sample quantization features contained in a quantization feature group and the texture feature loss values, and adjust the parameters of the candidate image processing sub-model and the quantization processing sub-model to be trained based on the quantization feature loss value.

[0058] A first generation of a quantization feature loss value is introduced with reference to the schematic flow diagram shown in FIG. 2c.

[0059] S2081: Adjust a quantization control parameter of the quantization processing sub-model to be trained based on the texture feature loss value.

[0060] The quantization control parameters of an i-th sample triplet are calculated by using the following Formula 3. Specifically, margin i characterizes the quantization control parameter of the i-th sample triplet, margin0 is a preset Hamming distance, and Mem is the ratio between the texture feature distance and the Hamming distance, and L_{em}_i is the texture feature loss value of the i-th sample triplet.

$$\text{margin\_i} = \text{margin0} * L_{em}\_i/Mem \quad \text{Formula 3:}$$

S2082: Determine a training sample loss value and a symbol quantization loss value of the quantization processing sub-model to be trained based on a plurality of sample quantization features and the quantization control parameter contained in one quantization feature group;

The following Formula 4 is the formula for calculating the loss value of training samples. L_{triplet} is the training sample loss value, x_a^q is a sample quantization feature of a sample video, x_p^q is a sample quantization feature of a positive label, x_n^q is a sample quantization feature of a negative label, $\left\|x_a^q - x_p^q\right\|$ characterizes a quantization feature distance between positive sample pairs, $\left\|x_a^q - x_n^q\right\|$ characterizes a quantization feature distance between negative sample pairs, and margin i characterizes a quantization control parameter of the i-th sample triplet.

$$L_{\text{triplet}} = \max(\left\|x_a^q - x_p^q\right\| - \left\|x_a^q - x_n^q\right\| + \text{margin\_i}) \quad \text{Formula 4:}$$

Each bit in the sample quantization feature is symbol quantized using the following Formula 5 to obtain the symbol

quantization feature. A symbol quantization loss value is then generated based on the sample quantization feature and the symbol quantization feature using the following Formula 6.

[0061] L_{coding} is a symbol quantization loss value, ui characterizes an i-th bit in a sample quantization feature, and bi characterizes an i-th bit of the symbol quantization feature; if ui is a negative number, then the value of bi is -1, otherwise, the value of bi is 1.

$$\mathrm{bi} = \mathrm{sgn}(\mathrm{ui}) = \begin{Bmatrix} -1, \text{if ui} < 0 \\ 1, \quad \text{else} \end{Bmatrix} \text{Formula 5:}$$

$$L_{\mathrm{coding}} = \sum_{i=1}^{128} (\mathrm{bi} - \mathrm{ui})^2 \text{ Formula 6:}$$

S2083: Generate a quantization feature loss value based on a training sample loss value and a symbol quantization loss value of a quantization processing sub-model to be trained.

[0062] A quantization feature loss value is generated using the following Formula 7; then, the parameters of the candidate image processing sub-model and the quantization processing sub-model to be trained are adjusted by using the SGD method based on the quantization feature loss value.

[0063] In Formula 7, L_q is the quantization feature loss value, $L_{triplet}$ is the training sample loss value of the quantization processing sub-model to be trained, w_{21} is the weight allocated to the training sample loss value, L_{coding} is the symbol quantization loss value of the quantization processing sub-model to be trained, and w_{22} is the weight allocated to the symbol quantization loss value.

$$L_q = w_{21} L_{\mathrm{triplet}} + w_{22} L_{\mathrm{coding}} \text{ Formula 7:}$$

With reference to the schematic flow diagram shown in FIG. 2d, embodiments of this application also provide a second way to generate a quantization feature loss value.

[0064] S2081': Determine a training sample loss value and a symbol quantization loss value of the quantization processing sub-model to be trained based on a plurality of sample quantization features contained in one quantization feature group;

The quantization features of a plurality of samples are substituted into Formula 4, and a loss value of training samples is generated; however, at this time, the value of margin i in Formula 4 is the same as that of margin_em.

[0065] A plurality of sample quantization features are sequentially substituted into Formulas 5 to 6 to obtain symbol quantization loss values. The related formulas have been introduced in the previous article, and will not be described here.

[0066] S2082': Generate a quantization feature loss value based on a training sample loss value, a symbol quantization loss value, and the texture feature loss value of the quantization processing sub-model to be trained.

[0067] A quantization feature loss value is generated using the following Formula 8; L_q is the quantization feature loss value, $L_{triplet}$ is the training sample loss value of the quantization processing sub-model to be trained, w_{21} and L_{em}' are the weights allocated to the training sample loss value, L_{coding} is the symbol quantization loss value of the quantization processing sub-model to be trained, and w_{22} is the weight allocated to the symbol quantization loss value.

$$L_q = w_{21}(L_{\mathrm{em}}' * L_{\mathrm{triplet}}) + w_{22} L_{\mathrm{coding}} \text{ Formula 8:}$$

S209: Determine whether the texture feature loss value and the quantization feature loss value are higher than the preset image processing texture feature loss threshold value, and if so, return operation S206; otherwise, execute operation S210.

[0068] S210: Stop iteration training of the sub-model and output the image processing sub-model, the texture processing sub-model, and the quantization processing sub-model obtained in the last iteration.

[0069] Next, with reference to the schematic flow diagram shown in FIG. 3a and the logic diagram shown in FIG. 3b, an index table and a mapping table of a video database are established by using a trained video retrieval model.

[0070] S301: Read one first candidate video s.

[0071] S302: Input the first candidate video s into a video retrieval model to obtain an initial quantization feature and a second texture feature.

[0072] S303: Add the second texture feature to the mapping table and determine a quantization feature distance between the initial quantization feature and each of second quantization features recorded in the index table.

[0073] S304: Add the first candidate video s to a second quantization feature corresponding to a minimum quantization

feature distance.

[0074] S305: Determine whether all the first candidate videos in the video database have been read, and if so, execute operation S306; otherwise, return operation S301.

[0075] S306: Output the mapping table and the index table obtained in the last iteration.

[0076] When operation S303 is executed, if the index table has a null value, the initial quantization feature of the first candidate video s is added to the index table as the second quantization feature. The index table is shown in Lindex: [q1:[img1,img2,img6],q2:[img3],q3:[img4]], the table includes a plurality of second quantization features, each quantization feature corresponds to one or more first candidate videos, and therefore each second quantization feature characterizes a video category to which the first candidate video corresponding to the second quantization feature belongs. The mapping table is shown as T: [[img1, embedding1], [img2, embedding2], ... [img6, embedding6]], and the table includes a plurality of first candidate videos, and second texture features.

[0077] In addition, for the first candidate video newly added into the video database, the flow as shown in FIG. 3a can also be executed to establish an indexing relationship and mapping relationship.

[0078] Next, with reference to the schematic flow diagram shown in FIG. 4a, the video retrieval method provided by all embodiments of this application may be applied on the trained video retrieval model.

[0079] S401: Extract an image feature from an input video using an image processing sub-model of a trained video retrieval model.

[0080] When operation S401 is executed, the complete video of the input video can be input into the image processing sub-model to obtain one image feature; it is also possible to extract key frames from the input video, and then input the obtained plurality of key frames into an image processing sub-model to obtain a plurality of image features.

[0081] S402: Extract a first quantization feature from the image feature using a quantization processing sub-model of the video retrieval model, identify, form the first candidate videos, one or more second candidate videos whose category similarity level to the input video meets a preset category similarity requirement based on the first quantization feature, wherein a quantization control parameter of the quantization processing sub-model is adjusted based on a texture feature loss value corresponding to each training sample during a training process, and the texture feature loss value is determined based on a texture control parameter preset for a texture processing sub-model during a parameter adjustment of the texture processing sub-model.

[0082] The quantization control parameter of the quantization processing sub-model is adjusted according to the texture feature loss value corresponding to each training sample, the quantization processing sub-model learns the ranking ability of the texture feature sub-model, which ensures that the ranking effects of the two sub-models tend to be consistent, and avoids the random ranking of the quantization processing sub-model due to the fixed quantization control parameters. The end-to-end model architecture enables the quantization processing sub-model to obtain a quantization feature based on the image feature, which reduces the loss in the process of generating the quantization feature and improves the generation accuracy of the quantization feature. In addition, all embodiments of this application may also optimize the ranking ability of the quantization processing sub-model and further improves the match performance of the video retrieval.

[0083] According to the introduction above, the index table contains a plurality of second quantization features, and each quantization feature corresponds to one or more first candidate videos. Therefore, when operation S402 is executed, a quantization feature distance between the first quantization feature and the second quantization feature of each of first candidate videos is determined, and then the first candidate video of which the quantization feature distance is lower than a preset quantization feature distance threshold value is determined as a second candidate video.

[0084] S403: Identify, among the one or more second candidate videos, a target video whose content similarity level to the input video meets a preset content similarity requirement.

[0085] In all embodiments of this application, a complete video of an input video can be used as a model input, or a plurality of obtained key frames can be used as a model input; therefore, the following several methods for obtaining a target video are provided for different model inputs.

[0086] Mode 1. For the above two models, the input is suitable, and the target video is identified according to the texture feature distance.

[0087] In all embodiments, a first texture feature may be extracted from the image feature using the texture processing sub-model. Then, the following operations may be performed for one or more second candidate videos: determining a texture feature distance between the first texture feature and a second texture feature of each second candidate video, determining that a content similarity level between the input video and the second candidate video meets the preset content similarity requirement when the texture feature distance is lower than a preset texture feature distance threshold value, and identifying the second candidate video as the target video, the second texture feature characterizing texture information of the second candidate video. In all embodiments of this application, the content similarity level between two videos may be determined by the texture feature distance between the texture features of the two videos. Accordingly, whether the content similarity level between the two videos meets the preset content similarity requirement can be determined as whether the texture feature distance reaches the preset texture feature distance threshold value. In all

embodiments of this application, a variety of distance calculation methods, such as a Euclidean distance and a Hamming distance, can be used to calculate a quantization feature distance and a texture feature distance; no matter which distance calculation method is used, if the value of the distance is small, it represents that the content similarity level to two videos is high; on the contrary, if the value of the distance is larger, it represents that the content similarity level between two videos is low, which will not be repeated in the future.

[0088] Mode 2. Taking the complete video as the model input, the target video is identified according to the content repetition degree.

[0089] In all embodiments, the following operations may be performed for the one or more second candidate videos:

determining a ratio between a total matching duration and a comparison duration as a content repetition degree between the input video and each second candidate video, wherein the total matching duration is determined as a matching duration between the one or more second candidate videos and the input video, and the comparison duration is determined as a value of a shorter video duration in the input video and the second candidate video; and

determining that a content similarity level between the input video and the second candidate video meets the preset content similarity requirement when the content repetition degree exceeds a preset content repetition degree threshold value, and identifying the second candidate video as the target video.

[0090] For example, assuming that the video duration of the input video is 30 s, and matching duration between each second candidate video and the input video is shown in Table 2, the content repetition degree between the input video and each of the or more second candidate videos is obtained by using Mode 2, and finally the second candidate videos 1 to 3 are returned to the user as target videos.

Table 2

The name of second candidate video	Video duration	Matching duration	Content repetition degree
Second candidate video 1	15 s	5 s	6
Second candidate video 2	20 s	10 s	4.5
Second candidate video 3	25 s	20 s	3.6
Second candidate video 4	60 s	35 s	3
Second candidate video 5	120 s	20 s	3

[0091] Mode 3. Taking a plurality of key frames as model input, the target video is identified according to the content repetition degree.

[0092] Each key frame corresponds to one first quantization feature, and each first quantization feature can match a second candidate video with the same feature; therefore, the ratio between the number of same quantization features and the comparison duration can be determined as the content repetition degree between the input video and the second candidate video.

[0093] In all embodiments, the following operations may be performed for the one or more second candidate videos:

determining the number of same quantization features between the input video and each second candidate video;

determining a ratio between the number of same quantization features and a comparison duration as a content repetition degree between the input video and the second candidate video, wherein the comparison duration is determined as a value of a shorter video duration in the input video and the second candidate video; and

determining that a content similarity level between the input video and the second candidate video meets the preset content similarity requirement when the content repetition degree exceeds a preset content repetition degree threshold value, and identifying the second candidate video as the target video.

[0094] For example, assuming that the video duration of the input video is 30s, a total of 10 key frames are extracted. The number of same quantization features between each second candidate video and the input video is shown in Table 3, the content repetition degree between the input video and each of the or more second candidate videos is obtained by using Mode 2, and finally the second candidate videos 1-2 are returned to the user as target videos.

Table 3

The name of second candidate video	Video duration	The number of same quantization features	Content repetition degree
Second candidate video 1	15 s	5	0.33
Second candidate video 2	20 s	8	0.4
Second candidate video 3	25 s	2	0.08
Second candidate video 4	60 s	3	0.1
Second candidate video 5	120 s	1	0.03

[0095] To facilitate understanding, the process of applying the video retrieval method in all embodiments may be described with reference to the logic diagram shown in FIG. 4b.

[0096] A complete video of an input video is inputted into a trained video retrieval model to obtain a first texture feature and a first quantization feature; a plurality of candidate videos with a higher category similarity level to the input video are obtained according to the quantization feature distance between the first quantization feature and each of second quantization features in the index table; and then the candidate videos ranked in the first N are taken as target videos with a higher content similarity level to the input video according to the texture feature distance between the first texture feature and the second texture feature of each of candidate videos matched in the last round, and returned to the user. In all embodiments of this application, the category similarity level between two videos may be determined by the quantization feature distance between the quantization features of the two videos. Accordingly, whether the category similarity level between the two videos meets the preset category similarity requirement can be determined as whether the quantization feature distance reaches the preset quantization feature distance threshold value. The smaller the quantization feature distance, the higher the category similarity level, and the larger the quantization feature distance, the lower the category similarity level.

[0097] Based on the same inventive concept as the above method embodiments, all embodiments of this application may also provide a video retrieval apparatus. As shown in FIG. 5, the apparatus 500 may include:

an image processing unit 501, configured to extract an image feature from an input video using an image processing sub-model of a trained video retrieval model;

a quantization processing unit 502, configured to extract a first quantization feature from the image feature using a quantization processing sub-model of the video retrieval model to obtain a first quantization feature, identify, from first candidate videos, one or more second candidate videos whose category similarity level to the input video meets a preset category similarity requirement based on the first quantization feature, wherein a quantization control parameter of the quantization processing sub-model is adjusted based on a texture feature loss value corresponding to each training sample during a training process, and the texture feature loss value is determined based on a texture control parameter preset for a texture processing sub-model during a parameter adjustment of the texture processing sub-model; and

a retrieval unit 503, configured to identify, among the one or more second candidate videos, a target video whose content similarity level to the input video meets a preset content similarity requirement.

[0098] In all embodiments, the video retrieval model may further include a texture processing sub-model, and the retrieval unit 503 may be configured to:

extract a first texture feature from the image feature using the texture processing sub-model; and

perform the following operations for the one or more second candidate videos: determining a texture feature distance between the first texture feature and a second texture feature of each second candidate video, determining that a content similarity level between the input video and the second candidate video meets the preset content similarity requirement when the texture feature distance is lower than a preset texture feature distance threshold value, and identifying the second candidate video as the target video, the second texture feature characterizing texture information of the second candidate video.

[0099] In all embodiments, the retrieval unit 503 may be configured to:

perform the following operations for the one or more second candidate videos:

determining a ratio between a total matching duration and a comparison duration as a content repetition degree between the input video and each second candidate video, wherein the total matching duration is determined as a matching duration between the one or more second candidate videos and the input video, and the comparison duration is determined as a value of a shorter video duration in the input video and the second candidate video; and

determining that a content similarity level between the input video and the second candidate video meets the preset content similarity requirement when the content repetition degree exceeds a preset content repetition degree threshold value, and identifying the second candidate video as the target video.

[0100] In all embodiments, the retrieval unit 503 may be configured to:

perform the following operations for the one or more second candidate videos:

determining the number of same quantization features between the input video and each second candidate video;

determining a ratio between the number of same quantization features and a comparison duration as a content repetition degree between the input video and the second candidate video, wherein the comparison duration is determined as a value of a shorter video duration in the input video and the second candidate video; and

determining that a content similarity level between the input video and the second candidate video meets the preset content similarity requirement when the content repetition degree exceeds a preset content repetition degree threshold value, and identifying the second candidate video as the target video.

[0101] In all embodiments, the quantization processing unit 502 may be configured to:

determine a quantization feature distance between the first quantization feature and a second quantization feature of each of the first candidate videos; and

determine a first candidate video with a quantization feature distance lower than a preset quantization feature distance threshold value as a second candidate video, each second quantization feature characterizing a video category to which a first candidate video corresponding to the second quantization feature belongs.

[0102] In all embodiments, the apparatus 500 may further include a model training unit 504, the model training unit 504 may obtain a trained video retrieval model by executing following operations:

obtaining a plurality of sample triplets, each sample triplet containing a sample video, a positive label and a negative label associated with the sample video;

sequentially inputting each sample triplet into an image processing sub-model and a texture processing sub-model as training data, to obtain a first texture set, generating an image processing texture feature loss value based on a plurality of first sample texture features contained in the first texture set, adjusting parameters of the image processing sub-model and the texture processing sub-model based on the image processing texture feature loss value, until the image processing texture feature loss value is not higher than a preset image processing texture feature loss threshold value, and determining the image processing sub-model and the texture processing sub-model as a candidate image processing sub-model and a candidate texture processing sub-model; and

sequentially inputting each sample triplet into the candidate image processing sub-model, the candidate texture processing sub-model and a quantization processing sub-model, to obtain a second texture set and a quantization feature group, generating a texture feature loss value based on a plurality of second sample texture features contained in the second texture set, adjusting a parameter of the candidate texture processing sub-model based on the texture feature loss value, generating a quantization feature loss value based on a plurality of sample quantization features contained in the quantization feature group and the texture feature loss value, adjusting parameters of the candidate image processing sub-model and the quantization processing sub-model based on the quantization feature loss value, until the texture feature loss value and the quantization feature loss value is not higher than a preset feature

loss threshold value, and determining the candidate image processing sub-model, the candidate texture processing sub-model and the quantization processing sub-model as the image processing sub-model, the texture processing sub-model and the quantization processing sub-model.

[0103] In all embodiments, the model training unit 504 may be configured to:

adjust a quantization control parameter of the quantization processing sub-model to be trained based on the texture feature loss value.

determine a training sample loss value and a symbol quantization loss value of the quantization processing sub-model to be trained based on the plurality of sample quantization features and the quantization control parameter contained in the quantization feature group; and

generate the quantization feature loss value based on the training sample loss value and the symbol quantization loss value of a quantization processing sub-model to be trained.

[0104] In all embodiments, the model training unit 504 may be configured to:

determine a training sample loss value and a symbol quantization loss value of the quantization processing sub-model to be trained based on the plurality of sample quantization features contained in one quantization feature group; and
generate the quantization feature loss value based on the training sample loss value, the symbol quantization loss value, and the texture feature loss value of the quantization processing sub-model to be trained.

[0105] For the convenience of description, the above parts are divided into modules (or units) and described separately according to the functions. Of course, the functions of each module (or unit) may be implemented in the same or in one or more pieces of software or hardware when this application is implemented.

[0106] After introducing the access method and apparatus of the service platform according to all embodiments of this application, next, the computer device according to all embodiments of this application may be introduced.

[0107] Those skilled in the art can understand that various aspects of this application can be implemented as a system, a method or a program product. Therefore, various aspects of this application can be embodied in the following forms, namely: hardware only implementations, software only implementations (including firmware, micro code, etc.), or implementations with a combination of software and hardware, which are collectively referred to as "circuit", "module", or "system" herein.

[0108] Based on the same inventive concept as the method embodiments described above, a computer device may be also provided in all embodiments of this application. With reference to FIG. 6, a computer device 600 may include at least a processor 601 and a memory 602. The memory 602 stores program codes which, when executed by the processor 601, cause the processor 601 to perform the operations of any one of the above video retrieval methods.

[0109] In all embodiments, a computing apparatus according to this application may include at least one processor, and at least one memory. The memory stores program codes which, when executed by the processor, cause the processor to perform the operations in the above-described video retrieval method according to all embodiments of this application. For example, the processor may execute the operations shown in FIG. 4.

[0110] A computing apparatus 700 according to all embodiments of this application may be described below with reference to FIG. 7. The computing apparatus 700 of FIG. 7 is merely one example and to be not pose any limitation on the function and application scope of embodiments of this application.

[0111] As shown in FIG. 7, the computing apparatus 700 is represented in the form of a general computing apparatus. Components of computing apparatus 700 may include, but are not limited to the at least one processing unit 701, the at least one memory unit 702, and a bus 703 connecting the different system components (including the memory unit 702 and the processing unit 701).

[0112] Bus 703 represents one or more of several types of bus structures, including a memory bus or memory controller, a peripheral bus, a processor, or a local bus using any of a variety of bus architectures.

[0113] The memory unit 702 may include computer readable storage medium in the form of volatile or non-volatile memory, such as a random access memory (RAM) 7021 and/or a cache memory unit 7022, and may further include a read-only memory (ROM) 7023. The computer readable storage medium includes program codes which, when run on a computer device, are configured to cause the computer device to execute the operations of any one of the above video retrieval methods.

[0114] The memory unit 702 may also include a program/utility 7025 having a set (at least one) of program modules 7024, such program modules 7024 including, but not limited to an operating system, one or more application programs,

other program modules, and program data, and each or a combination of these examples may include implementation of a network environment.

[0115] Computing apparatus 700 can also communicate with one or more external devices 704 (e.g., a keyboard, a pointing device, etc.), one or more devices that enable a user to interact with computing apparatus 700, and/or any device (e.g., a router, a modem, etc.) that enables computing apparatus 700 to communicate with one or more other computing apparatuses. Such communication may occur through an input/output (I/O) interface 705. Moreover, the computing apparatus 700 can also communicate with one or more networks (such as a local area network (LAN), a wide area network (WAN) and/or a public network, such as the Internet) through a network adapter 706. As shown, the network adapter 706 communicates with other modules for the computing apparatus 700 through the bus 703. Although not shown in the figures, other hardware and/or software modules may be used in conjunction with computing apparatus 700 including, but not limited to: Microcode, device drivers, redundant processors, external disk drive arrays, RAID systems, tape drives, and data backup storage systems and the like.

[0116] Based on the same inventive concept as the above method embodiments, various aspects of the video retrieval method provided by this application can also be realized in the form of a program product, which includes program codes. When the program product is run on a computer device, the program code is used to make the computer device execute the operations in the video retrieval method according to all embodiments of this application described above in this specification. For example, the electronic device can perform the operations as shown in FIG. 4.

[0117] The program product can use any combination of one or more readable media. The readable medium may be a computer-readable signal medium or a computer readable storage medium. The readable storage medium can be, for example but not limited to, electronic, magnetic, optical, electromagnetic, infrared, or a semiconductor system, an apparatus, or a device, or a combination of any of the above. More specific examples (a non-exhaustive list) of readable storage media include: an electrical connection having one or more wires, a portable disk, a hard disk, a random-access memory (RAM), a read-only memory (ROM), an erasable programmable read-only memory (EPROM or flash memory), an optical fiber, a portable compact disk read-only memory (CD-ROM), an optical storage device, a magnetic storage device, or any suitable combination of the above.

[0118] Specifically, all embodiments of this application may provide a computer program product or computer program including computer instructions stored in a computer readable storage medium. A processor of a computer device reads the computer instructions from a computer readable storage medium, and the processor executes the computer instructions to cause the computer device to perform the operations of any one of the above video retrieval methods.

[0119] Although all embodiments of this application have been described, once persons skilled in the art know a basic creative concept, they can make other changes and modifications to these embodiments. Therefore, the following claims are intended to cover all embodiments and all changes and modifications falling within the scope of this application.

[0120] Obviously, a person skilled in the art can make various modifications and variations to this application without departing from the scope of this application. In this case, if the modifications and variations made to this application fall within the scope of the claims of this application and their equivalent technologies, this application is intended to contain these modifications and variations.

Claims

1. A video retrieval method, executable by a computer device, and comprising:

   extracting an image feature from an input video using an image processing sub-model of a trained video retrieval model;

   extracting a first quantization feature from the image feature using a quantization processing sub-model of the video retrieval model, wherein the quantization processing sub-model comprises a quantization control parameter adjusted based on a texture feature loss value corresponding to each training sample during a training process, and the texture feature loss value is determined based on a texture control parameter preset for a texture processing sub-model during a parameter adjustment of the texture processing sub-model;

   identifying, from first candidate videos, one or more second candidate videos whose category similarity level to the input video meets a preset category similarity requirement based on the first quantization feature; and

   identifying, among the one or more second candidate videos, a target video whose content similarity level to the input video meets a preset content similarity requirement.

2. The method according to claim 1, wherein the video retrieval model further comprises a texture processing sub-model; and

   the identifying, among the one or more second candidate videos, a target video whose content similarity level to the input video meets a preset content similarity requirement comprises:

extracting a first texture feature from the image feature using the texture processing sub-model;
determining a texture feature distance between the first texture feature and a second texture feature of each second candidate video;
determining that a content similarity level between the input video and the second candidate video meets the preset content similarity requirement when the texture feature distance is lower than a preset texture feature distance threshold value; and
identifying the second candidate video as the target video.

3. The method according to claim 1, wherein the identifying, among the one or more second candidate videos, a target video whose content similarity level to the input video meets a preset content similarity requirement comprises:

determining a ratio between a total matching duration and a comparison duration as a content repetition degree between the input video and each second candidate video, wherein the total matching duration is determined as a matching duration between the one or more second candidate videos and the input video, and the comparison duration is determined as a value of a shorter video duration in the input video and the second candidate video;
determining that a content similarity level between the input video and the second candidate video meets the preset content similarity requirement when the content repetition degree exceeds a preset content repetition degree threshold value; and
identifying the second candidate video as the target video.

4. The method according to claim 1, wherein the identifying, among the one or more second candidate videos, a target video whose content similarity level to the input video meets a preset content similarity requirement comprises:

determining a number of same quantization features between the input video and each second candidate video;
determining a ratio between the number of same quantization features and a comparison duration as a content repetition degree between the input video and the second candidate video, wherein the comparison duration is determined as a value of a shorter video duration in the input video and the second candidate video;
determining that a content similarity level between the input video and the second candidate video meets the preset content similarity requirement when the content repetition degree exceeds a preset content repetition degree threshold value; and
identifying the second candidate video as the target video.

5. The method according to any one of claims 1 to 4, wherein the identifying, from first candidate videos, one or more second candidate videos whose category similarity level to the input video meets a preset category similarity requirement based on the first quantization feature comprises:

determining a quantization feature distance between the first quantization feature and a second quantization feature of each of the first candidate videos; and
determining a first candidate video with a quantization feature distance lower than a preset quantization feature distance threshold value as a second candidate video, the second quantization feature characterizing a video category to which a first candidate video corresponding to the second quantization feature belongs.

6. The method according to any one of claims 1 to 4, wherein the video retrieval model is obtained by performing following operations:

obtaining a plurality of sample triplets, each sample triplet containing a sample video, a positive label and a negative label associated with the sample video;
sequentially inputting each sample triplet into an image processing sub-model and a texture processing sub-model as training data, to obtain a first texture set, generating an image processing texture feature loss value based on a plurality of first sample texture features contained in the first texture set, adjusting parameters of the image processing sub-model and the texture processing sub-model based on the image processing texture feature loss value, until the image processing texture feature loss value is not higher than a preset image processing texture feature loss threshold value, and determining the image processing sub-model and the texture processing sub-model as a candidate image processing sub-model and a candidate texture processing sub-model; and
sequentially inputting each sample triplet into the candidate image processing sub-model, the candidate texture processing sub-model and a quantization processing sub-model, to obtain a second texture set and a quanti-

zation feature group, generating a texture feature loss value based on a plurality of second sample texture features contained in the second texture set, adjusting a parameter of the candidate texture processing sub-model based on the texture feature loss value, generating a quantization feature loss value based on a plurality of sample quantization features contained in the quantization feature group and the texture feature loss value, adjusting parameters of the candidate image processing sub-model and the quantization processing sub-model based on the quantization feature loss value, until the texture feature loss value and the quantization feature loss value is not higher than a preset feature loss threshold value, and determining the candidate image processing sub-model, the candidate texture processing sub-model and the quantization processing sub-model as the image processing sub-model, the texture processing sub-model and the quantization processing sub-model.

7. The method according to claim 6, wherein the generating a quantization feature loss value based on a plurality of sample quantization features contained in the quantization feature group and the texture feature loss value comprises:

adjusting a quantization control parameter of the quantization processing sub-model based on the texture feature loss value;
determining a training sample loss value and a symbol quantization loss value of the quantization processing sub-model based on the plurality of sample quantization features and the quantization control parameter contained in the quantization feature group; and
generating the quantization feature loss value based on the training sample loss value and the symbol quantization loss value of the quantization processing sub-model.

8. The method according to claim 6, wherein the generating a quantization feature loss value based on a plurality of sample quantization features contained in the quantization feature group and the texture feature loss value comprises:

determining a training sample loss value and a symbol quantization loss value of the quantization processing sub-model based on the plurality of sample quantization features contained in the quantization feature group; and generating the quantization feature loss value based on the training sample loss value, the symbol quantization loss value, and the texture feature loss value of the quantization processing sub-model.

9. A video retrieval apparatus, comprising:

an image processing unit, configured to extract an image feature from an input video using an image processing sub-model of a trained video retrieval model;
a quantization processing unit, configured to extract a first quantization feature from the image feature using a quantization processing sub-model of the video retrieval model, identify, from first candidate videos, one or more second candidate videos whose category similarity level to the input video meets a preset category similarity requirement based on the first quantization feature, wherein the quantization processing sub-model comprises a quantization control parameter adjusted based on a texture feature loss value corresponding to each training sample during a training process, and the texture feature loss value is determined based on a texture control parameter preset for a texture processing sub-model during a parameter adjustment of the texture processing sub-model; and
a retrieval unit, configured to identify, among the one or more second candidate videos, a target video whose content similarity level to the input video meets a preset content similarity requirement.

10. The apparatus according to claim 9, wherein the video retrieval model further comprises a texture processing sub-model, and the retrieval unit is configured to:

extract a first texture feature from the image feature using the texture processing sub-model;
determine a texture feature distance between the first texture feature and a second texture feature of each second candidate video;
determine that a content similarity level between the input video and the second candidate video meets the preset content similarity requirement when the texture feature distance is lower than a preset texture feature distance threshold value; and
identify the second candidate video as the target video.

11. The apparatus according to claim 9, wherein the retrieval unit is configured to:

determine a ratio between a total matching duration and a comparison duration as a content repetition degree

between the input video and each second candidate video, wherein the total matching duration is determined as a matching duration between the one or more second candidate videos and the input video, and the comparison duration is determined as a value of a shorter video duration in the input video and the second candidate video; and

determine that a content similarity level between the input video and the second candidate video meets the preset content similarity requirement when the content repetition degree exceeds a preset content repetition degree threshold value; and

identify the second candidate video as the target video.

12. The apparatus according to claim 9, wherein the retrieval unit is configured to:

determine a number of same quantization features between the input video and each second candidate video;

determine a ratio between the number of same quantization features and a comparison duration as a content repetition degree between the input video and the second candidate video, wherein the comparison duration is determined as a value of a shorter video duration in the input video and the second candidate video; and

determine that a content similarity level between the input video and the second candidate video meets the preset content similarity requirement when the content repetition degree exceeds a preset content repetition degree threshold value; and

identify the second candidate video as the target video.

13. The apparatus according to any of claims 9 to 12, wherein the quantization processing unit is configured to:

determine a quantization feature distance between the first quantization feature and a second quantization feature of each of the first candidate videos; and

determine a first candidate video with a quantization feature distance lower than a preset quantization feature distance threshold value as a second candidate video, the second quantization feature characterizing a video category to which a first candidate video corresponding to the second quantization feature belongs.

14. A computer device comprising a processor and a memory, the memory storing program codes which, when executed by the processor, cause the processor to perform the method according to any one of claims 1 to 8.

15. A computer readable storage medium comprising program codes which, when executed by a computer device, cause the computer device to perform the method according to any one of claims 1 to 8.

16. A computer program product or computer program comprising computer instructions stored in a computer readable storage medium, which, when read and executed by a processor of a computer device, cause the computer device to perform the method according to any one of claims 1 to 8.

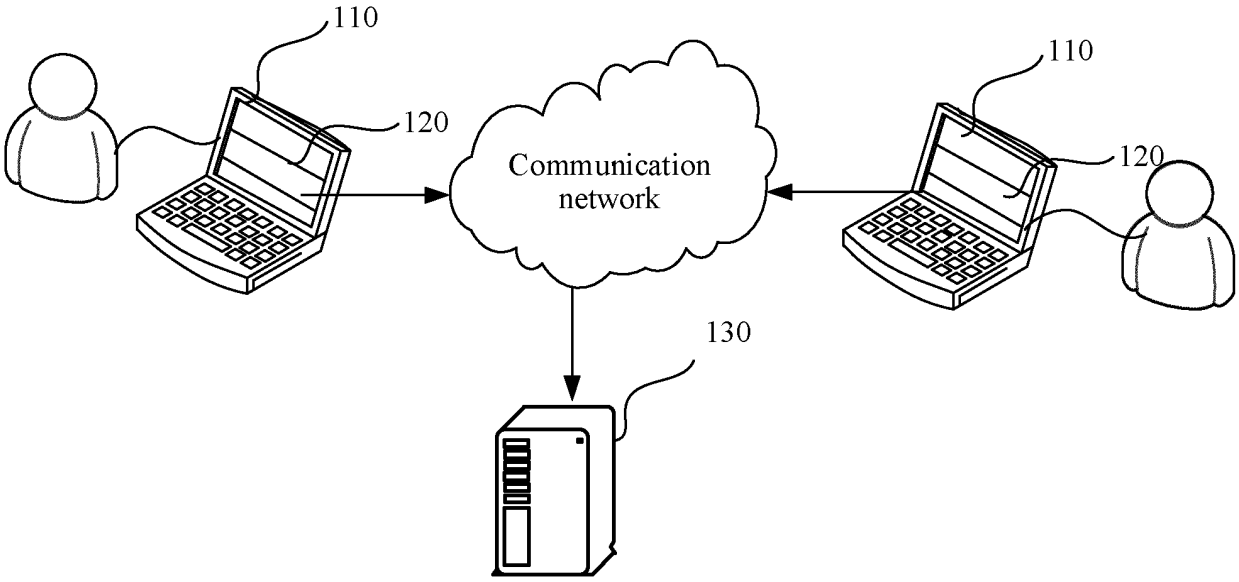

FIG. 1a

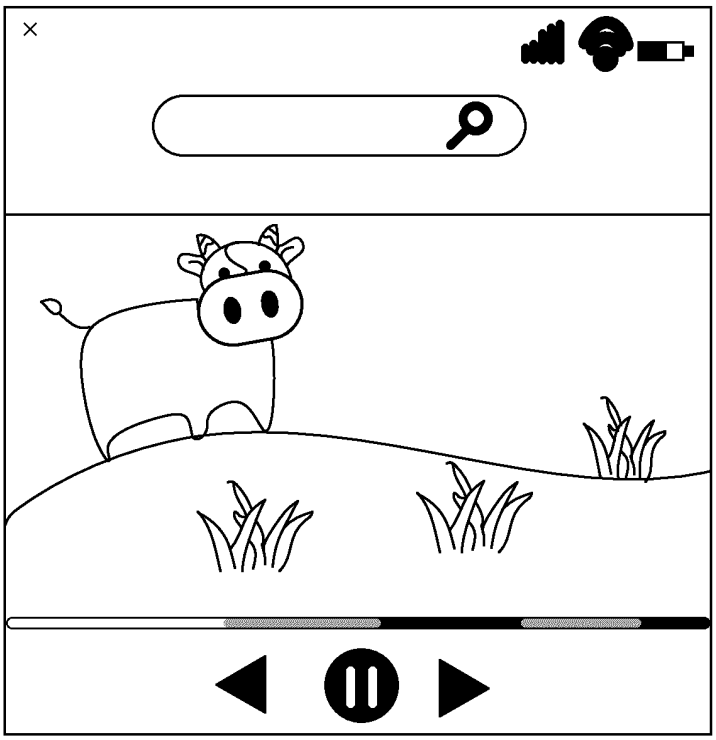

FIG. 1b

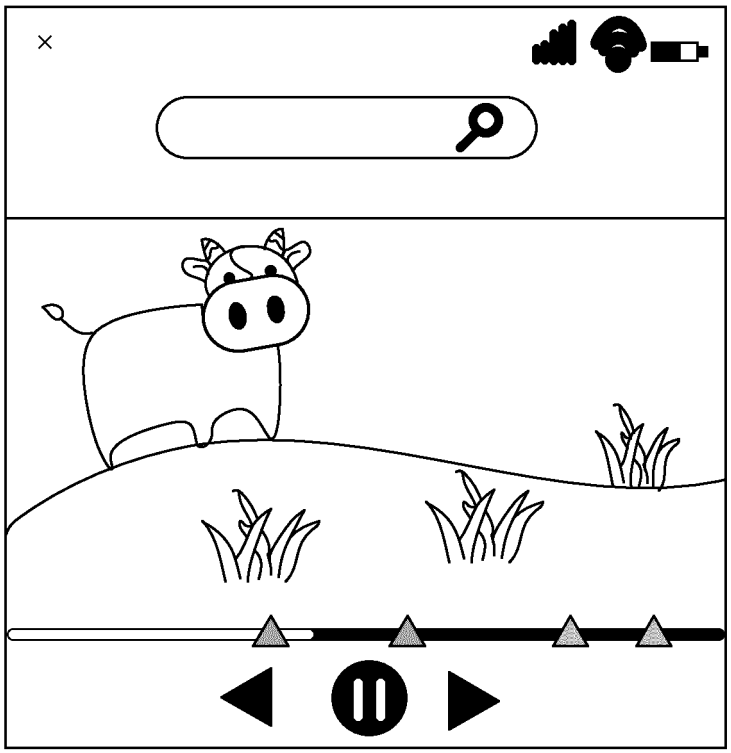

FIG. 1c

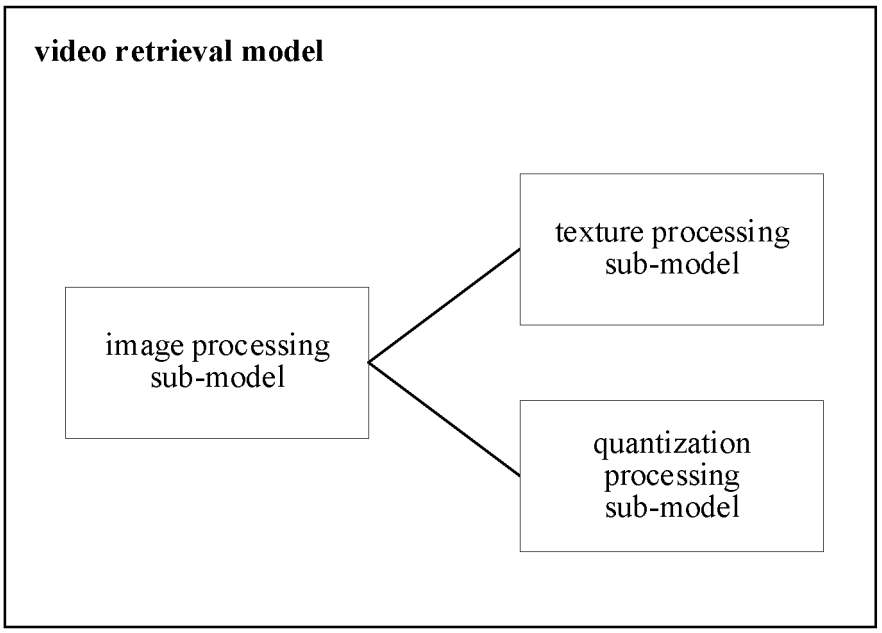
video retrieval model
image processing sub-model
texture processing sub-model
quantization processing sub-model

FIG. 1d

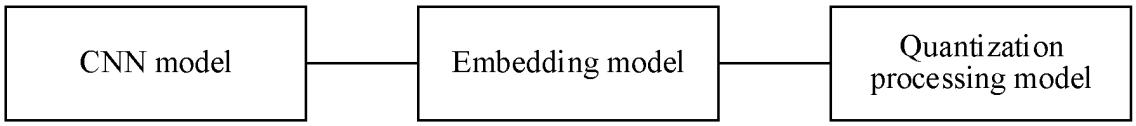
CNN model
Embedding model
Quantization processing model

FIG. 1e

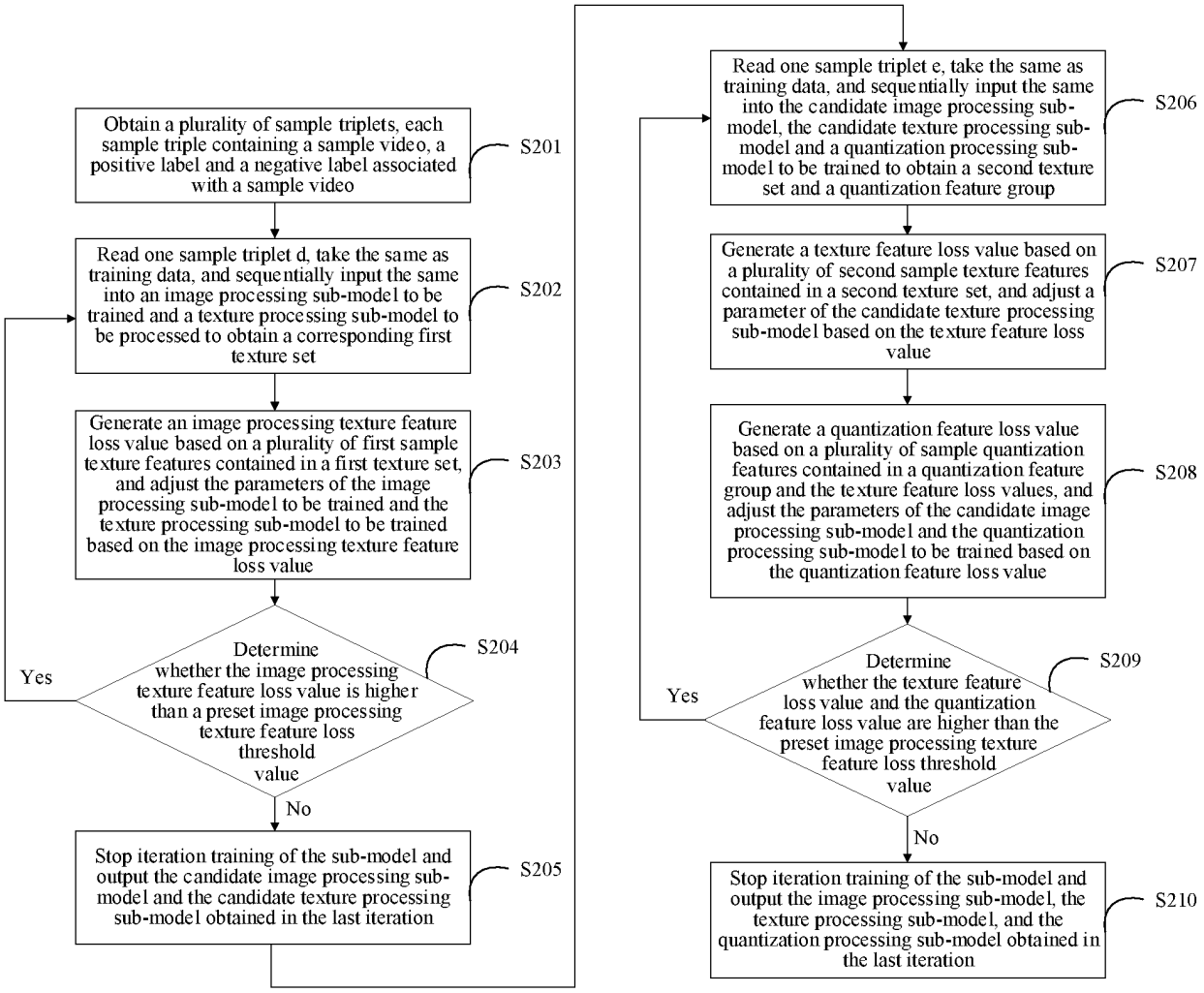
Obtain a plurality of sample triplets, each sample triple containing a sample video, a positive label and a negative label associated with a sample video
S201
Read one sample triplet d, take the same as training data, and sequentially input the same into an image processing sub-model to be trained and a texture processing sub-model to be processed to obtain a corresponding first texture set
S202
Generate an image processing texture feature loss value based on a plurality of first sample texture features contained in a first texture set, and adjust the parameters of the image processing sub-model to be trained and the texture processing sub-model to be trained based on the image processing texture feature loss value
S203
Determine whether the image processing texture feature loss value is higher than a preset image processing texture feature loss threshold value
S204
Yes
No
Stop iteration training of the sub-model and output the candidate image processing sub-model and the candidate texture processing sub-model obtained in the last iteration
S205
Read one sample triplet e, take the same as training data, and sequentially input the same into the candidate image processing sub-model, the candidate texture processing sub-model and a quantization processing sub-model to be trained to obtain a second texture set and a quantization feature group
S206
Generate a texture feature loss value based on a plurality of second sample texture features contained in a second texture set, and adjust a parameter of the candidate texture processing sub-model based on the texture feature loss value
S207
Generate a quantization feature loss value based on a plurality of sample quantization features contained in a quantization feature group and the texture feature loss values, and adjust the parameters of the candidate image processing sub-model and the quantization processing sub-model to be trained based on the quantization feature loss value
S208
Determine whether the texture feature loss value and the quantization feature loss value are higher than the preset image processing texture feature loss threshold value
S209
Yes
No
Stop iteration training of the sub-model and output the image processing sub-model, the texture processing sub-model, and the quantization processing sub-model obtained in the last iteration
S210

FIG. 2a

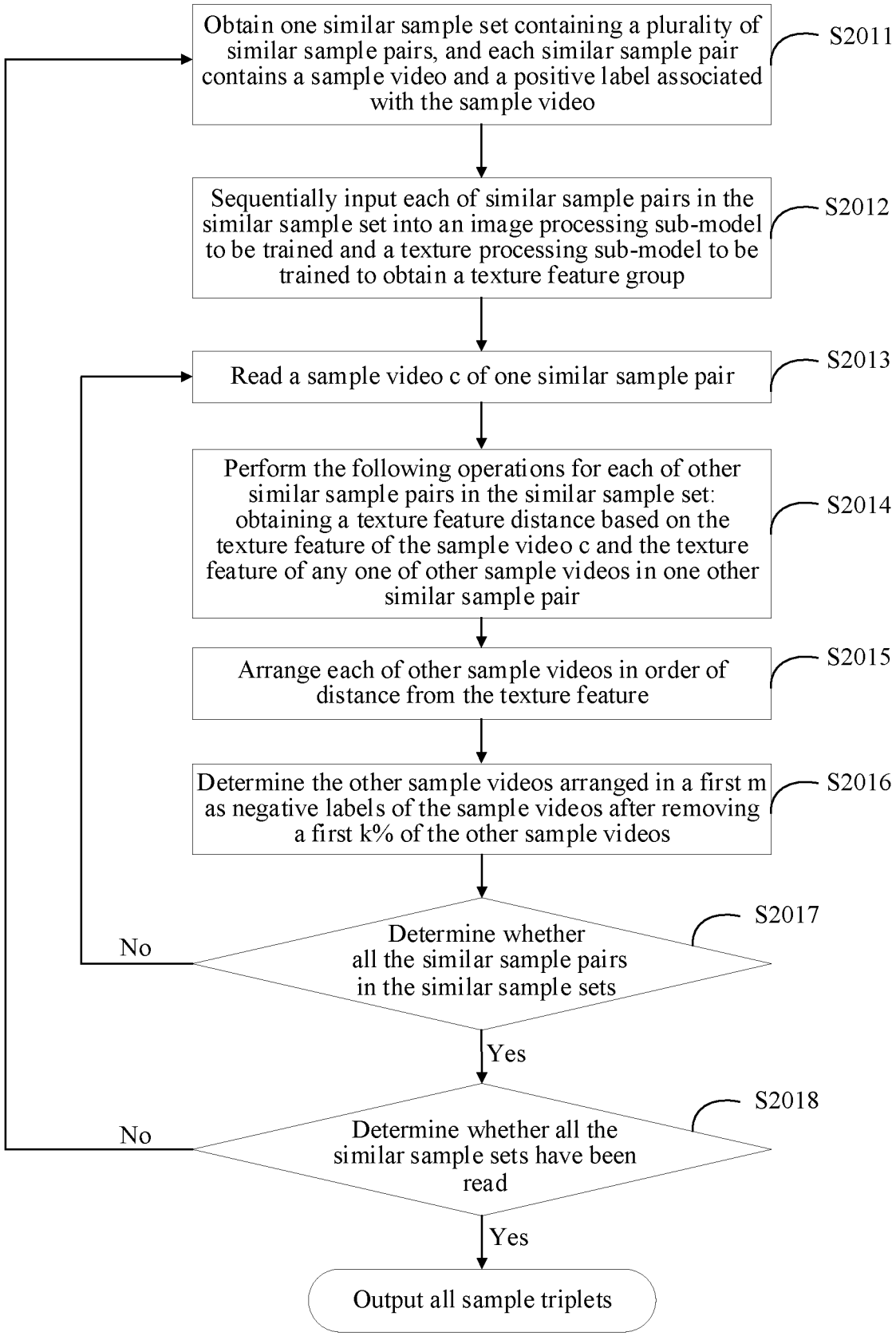
Obtain one similar sample set containing a plurality of similar sample pairs, and each similar sample pair contains a sample video and a positive label associated with the sample video
S2011
Sequentially input each of similar sample pairs in the similar sample set into an image processing sub-model to be trained and a texture processing sub-model to be trained to obtain a texture feature group
S2012
Read a sample video c of one similar sample pair
S2013
Perform the following operations for each of other similar sample pairs in the similar sample set: obtaining a texture feature distance based on the texture feature of the sample video c and the texture feature of any one of other sample videos in one other similar sample pair
S2014
Arrange each of other sample videos in order of distance from the texture feature
S2015
Determine the other sample videos arranged in a first m as negative labels of the sample videos after removing a first k% of the other sample videos
S2016
Determine whether all the similar sample pairs in the similar sample sets
S2017
No
Yes
Determine whether all the similar sample sets have been read
S2018
No
Yes
Output all sample triplets

FIG. 2b

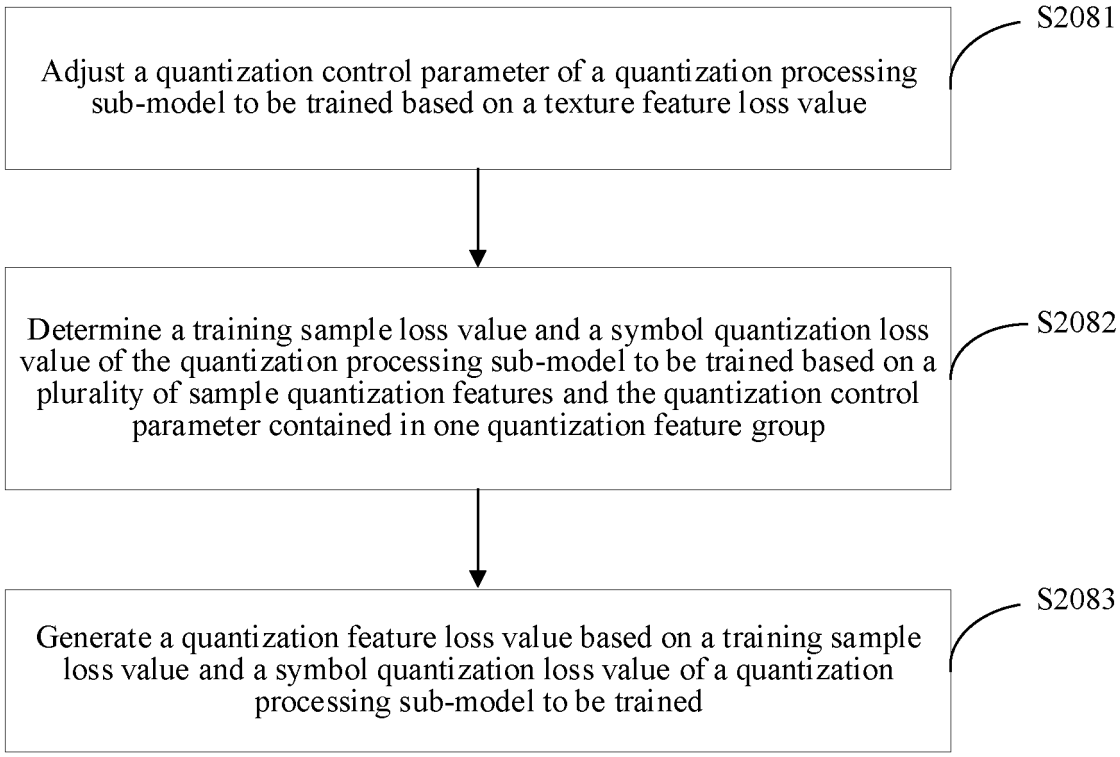
Adjust a quantization control parameter of a quantization processing sub-model to be trained based on a texture feature loss value
S2081
Determine a training sample loss value and a symbol quantization loss value of the quantization processing sub-model to be trained based on a plurality of sample quantization features and the quantization control parameter contained in one quantization feature group
S2082
Generate a quantization feature loss value based on a training sample loss value and a symbol quantization loss value of a quantization processing sub-model to be trained
S2083

FIG. 2c

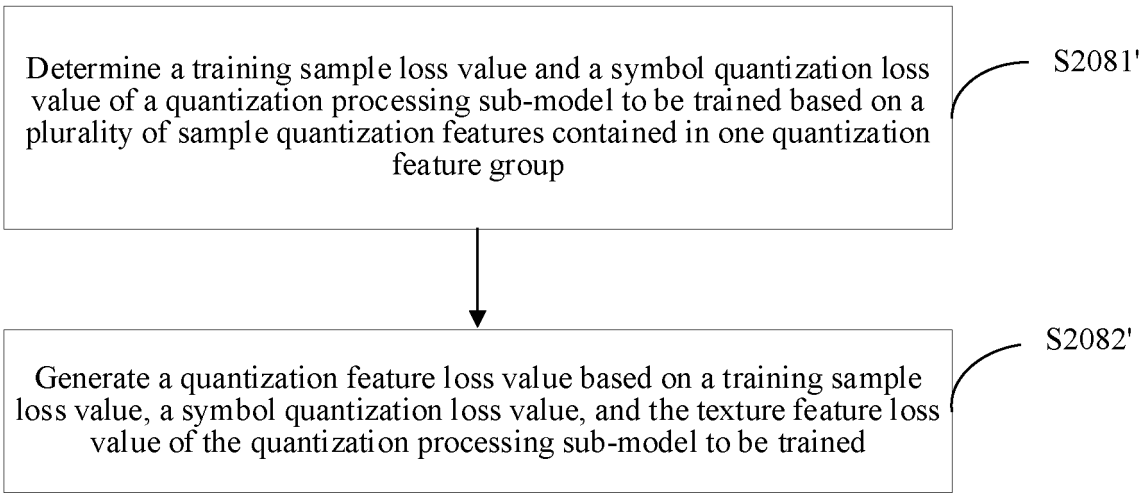
Determine a training sample loss value and a symbol quantization loss value of a quantization processing sub-model to be trained based on a plurality of sample quantization features contained in one quantization feature group
S2081'
Generate a quantization feature loss value based on a training sample loss value, a symbol quantization loss value, and the texture feature loss value of the quantization processing sub-model to be trained
S2082'

FIG. 2d

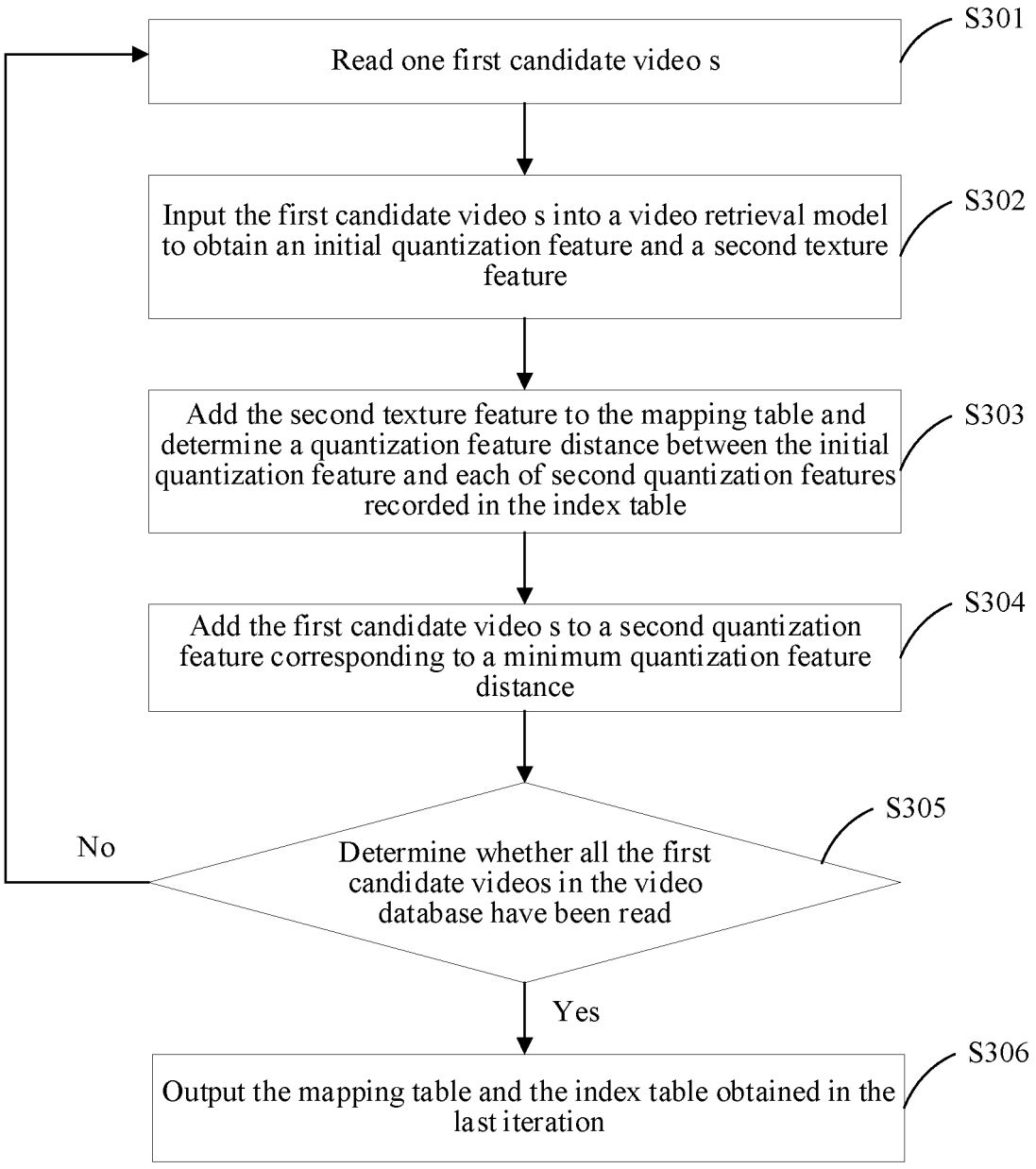
S301
Read one first candidate video s
S302
Input the first candidate video s into a video retrieval model to obtain an initial quantization feature and a second texture feature
S303
Add the second texture feature to the mapping table and determine a quantization feature distance between the initial quantization feature and each of second quantization features recorded in the index table
S304
Add the first candidate video s to a second quantization feature corresponding to a minimum quantization feature distance
S305
Determine whether all the first candidate videos in the video database have been read
No
Yes
S306
Output the mapping table and the index table obtained in the last iteration

FIG. 3a

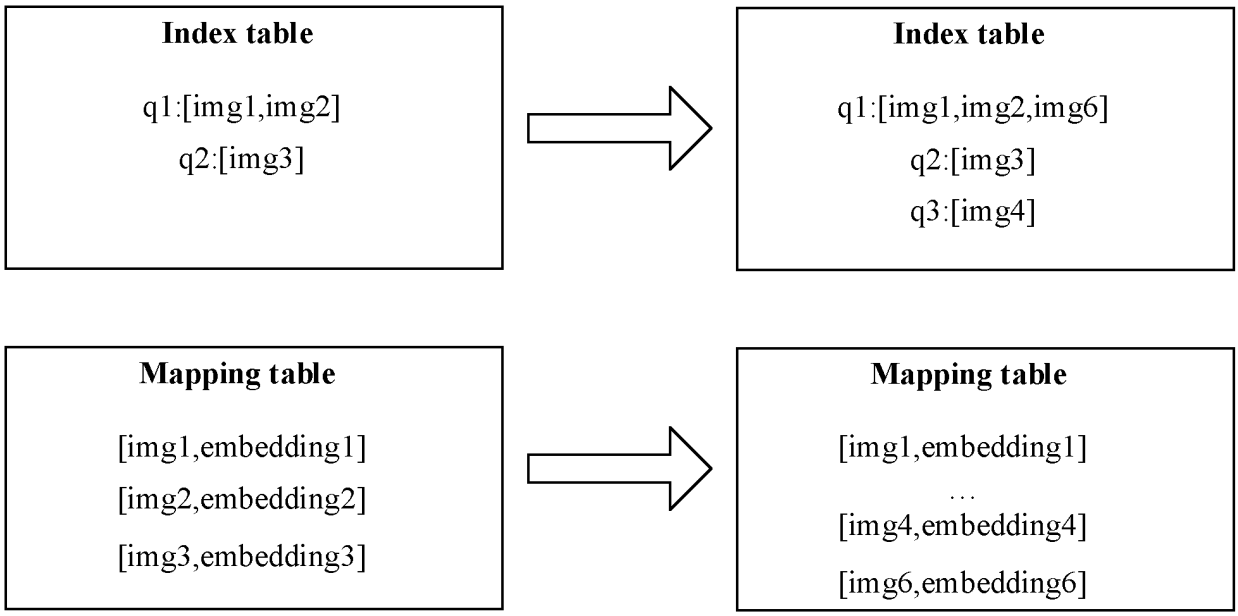
Index table
q1:[img1,img2]
q2:[img3]
Index table
q1:[img1,img2,img6]
q2:[img3]
q3:[img4]
Mapping table
[img1,embedding1]
[img2,embedding2]
[img3,embedding3]
Mapping table
[img1,embedding1]
...
[img4,embedding4]
[img6,embedding6]

FIG. 3b

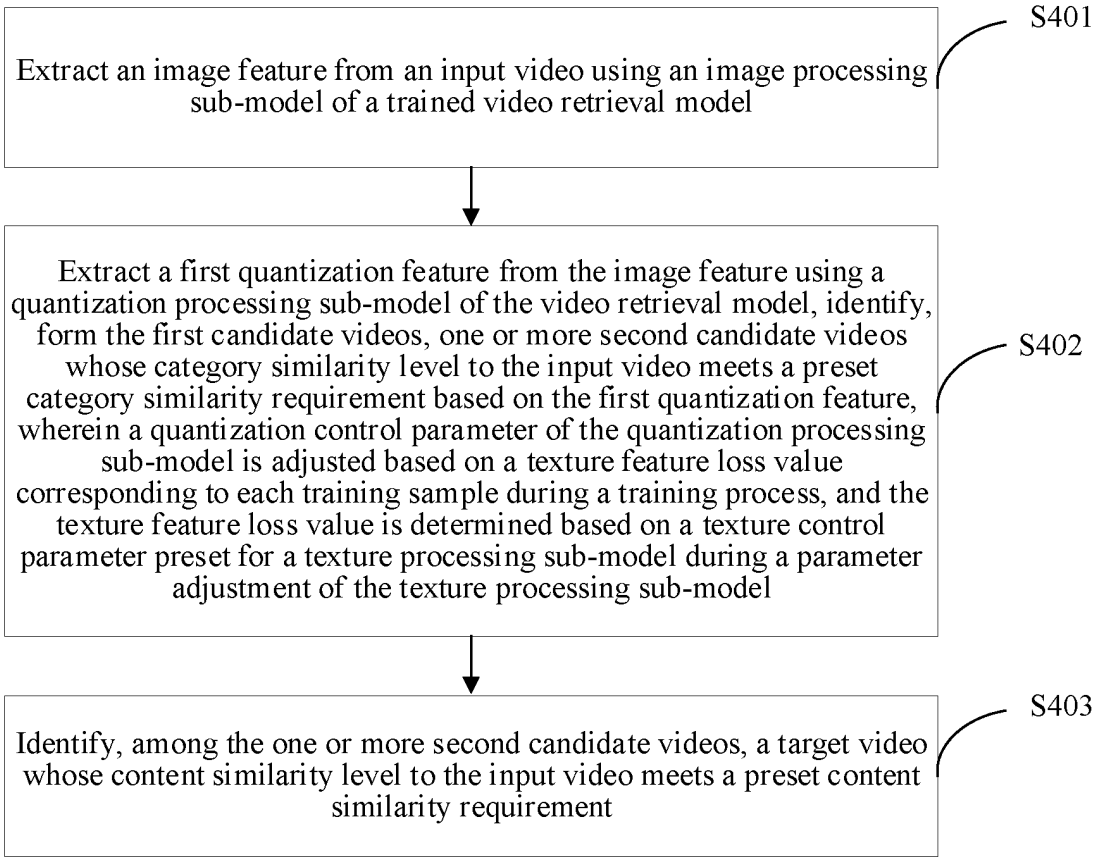
Extract an image feature from an input video using an image processing sub-model of a trained video retrieval model
S401
Extract a first quantization feature from the image feature using a quantization processing sub-model of the video retrieval model, identify, form the first candidate videos, one or more second candidate videos whose category similarity level to the input video meets a preset category similarity requirement based on the first quantization feature, wherein a quantization control parameter of the quantization processing sub-model is adjusted based on a texture feature loss value corresponding to each training sample during a training process, and the texture feature loss value is determined based on a texture control parameter preset for a texture processing sub-model during a parameter adjustment of the texture processing sub-model
S402
Identify, among the one or more second candidate videos, a target video whose content similarity level to the input video meets a preset content similarity requirement
S403

FIG. 4a

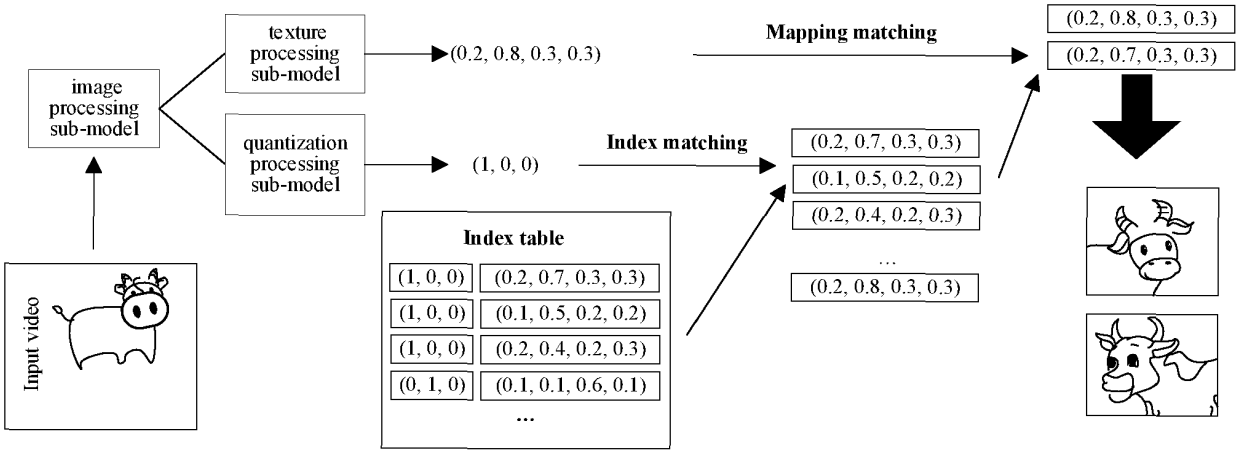
image processing sub-model
texture processing sub-model
quantization processing sub-model
(0.2, 0.8, 0.3, 0.3)
(1, 0, 0)
Mapping matching
Index matching
Index table
(1, 0, 0) (0.2, 0.7, 0.3, 0.3)
(1, 0, 0) (0.1, 0.5, 0.2, 0.2)
(1, 0, 0) (0.2, 0.4, 0.2, 0.3)
(0, 1, 0) (0.1, 0.1, 0.6, 0.1)
(0.2, 0.7, 0.3, 0.3)
(0.1, 0.5, 0.2, 0.2)
(0.2, 0.4, 0.2, 0.3)
(0.2, 0.8, 0.3, 0.3)
(0.2, 0.8, 0.3, 0.3)
(0.2, 0.7, 0.3, 0.3)
Input video

FIG. 4b

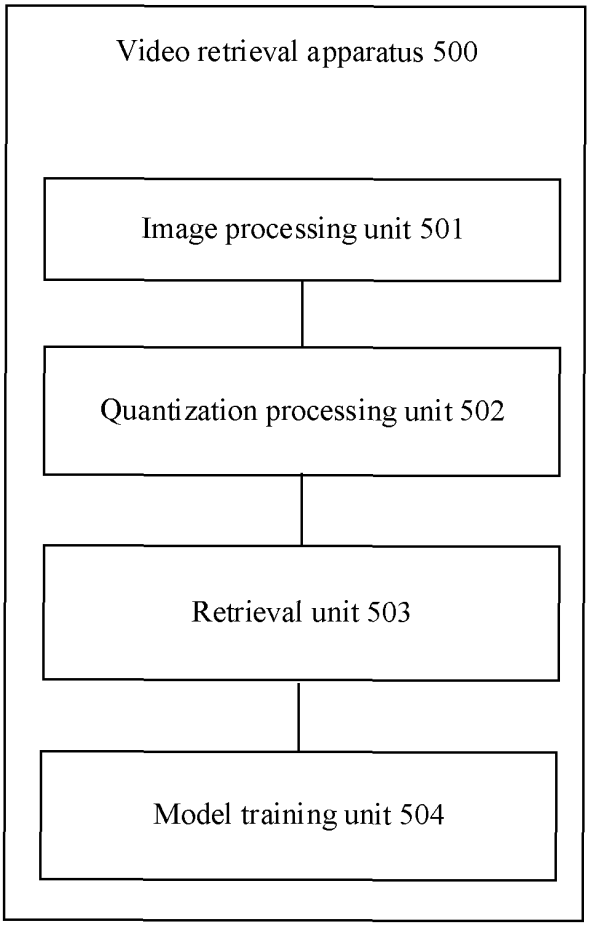
Video retrieval apparatus 500
Image processing unit 501
Quantization processing unit 502
Retrieval unit 503
Model training unit 504

FIG. 5

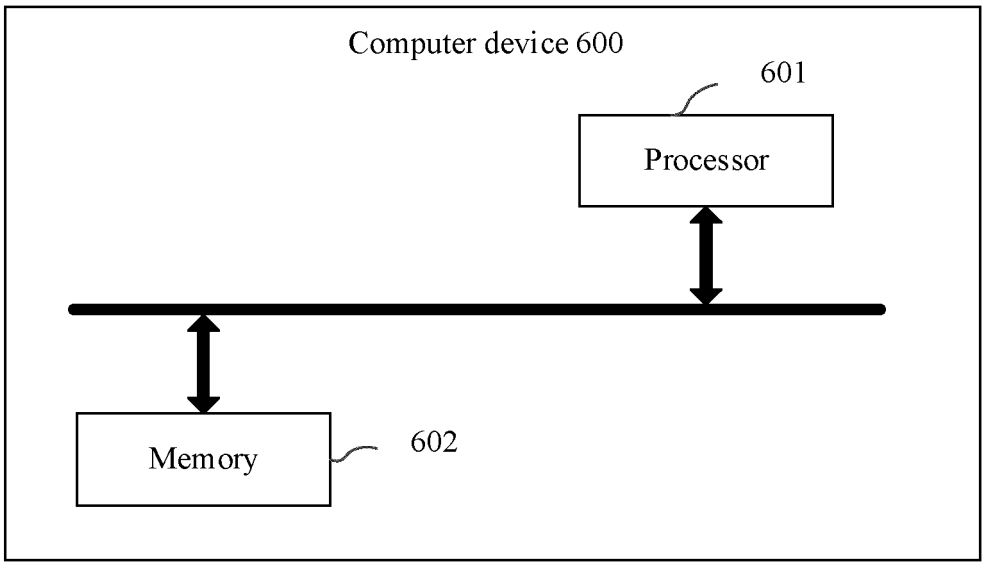
Computer device 600
601
Processor
Memory
602

FIG. 6

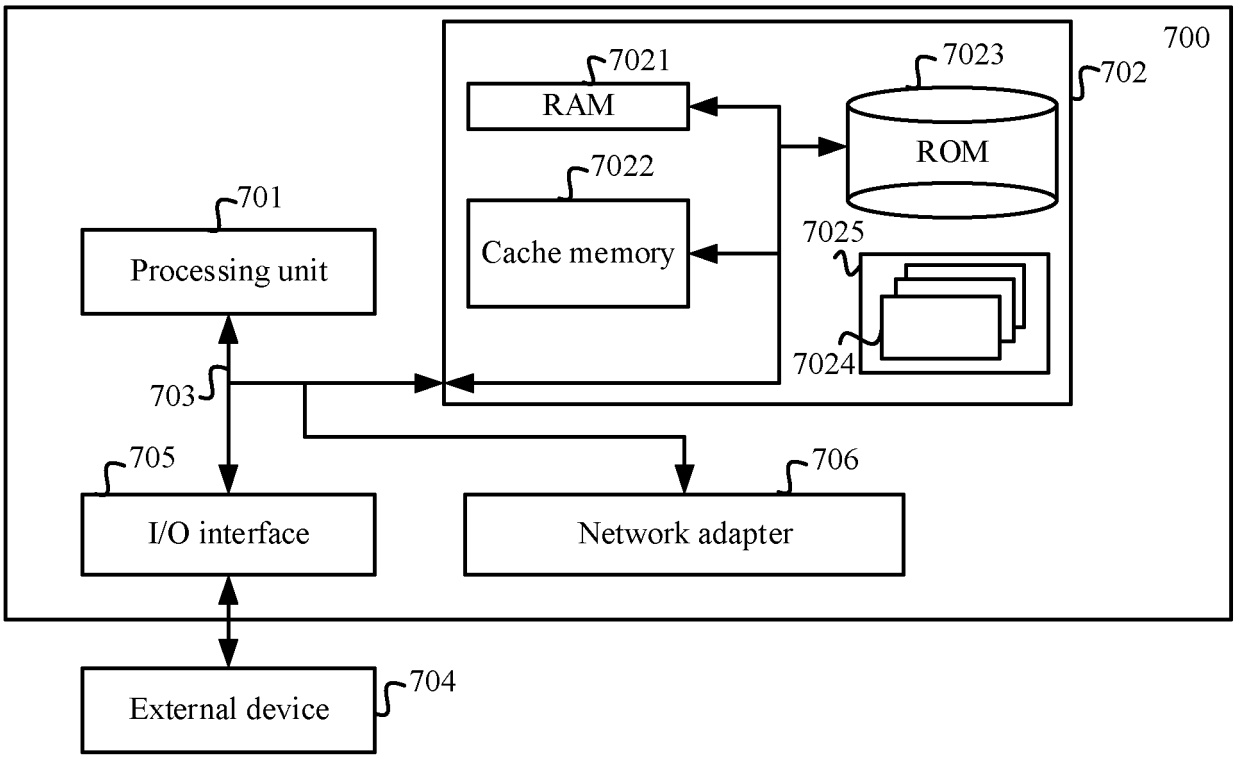
700
7021
RAM
7023
ROM
702
7022
Cache memory
7025
7024
701
Processing unit
703
705
706
I/O interface
Network adapter
External device
704

FIG. 7

INTERNATIONAL SEARCH REPORT

International application No.
PCT/CN2022/105871

A. CLASSIFICATION OF SUBJECT MATTER

G06F 16/783(2019.01)i

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

G06F

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

CNPAT, WPI, EPODOC, CNKI, IEEE: 视频, 图像, 检索, 特征, 提取, 纹理, 相似度, 内容, 损失值, video, image, retrieval, feature, extraction, texture, similarity, content, loss value

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
PX	CN 114282059 A (TENCENT TECHNOLOGY SHENZHEN CO., LTD.) 05 April 2022 (2022-04-05) claims 1-15, and description, paragraphs [0217]-[0218]	1-16
A	CN 113254687 A (TENCENT TECHNOLOGY SHENZHEN CO., LTD.) 13 August 2021 (2021-08-13) abstract, and description, paragraphs [0004]-[0020]	1-16
A	CN 113255625 A (TENCENT TECHNOLOGY SHENZHEN CO., LTD.) 13 August 2021 (2021-08-13) entire document	1-16
A	US 2019332867 A1 (TENCENT TECHNOLOGY (SHENZHEN) CO., LTD.) 31 October 2019 (2019-10-31) entire document	1-16
A	WO 2021017289 A1 (PING AN TECHNOLOGY SHENZHEN CO., LTD.) 04 February 2021 (2021-02-04) entire document	1-16

☐ Further documents are listed in the continuation of Box C. ☑ See patent family annex.

* Special categories of cited documents:
"A" document defining the general state of the art which is not considered to be of particular relevance
"E" earlier application or patent but published on or after the international filing date
"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)
"O" document referring to an oral disclosure, use, exhibition or other means
"P" document published prior to the international filing date but later than the priority date claimed
"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
05 August 2022	**25 August 2022**
Name and mailing address of the ISA/CN **China National Intellectual Property Administration (ISA/CN)** **No. 6, Xitucheng Road, Jimenqiao, Haidian District, Beijing 100088, China** Facsimile No. **(86-10)62019451**	Authorized officer Telephone No.

Form PCT/ISA/210 (second sheet) (January 2015)

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.
PCT/CN2022/105871

Patent document cited in search report	Publication date (day/month/year)	Patent family member(s)	Publication date (day/month/year)
CN 114282059 A	05 April 2022	None	
CN 113254687 A	13 August 2021	None	
CN 113255625 A	13 August 2021	None	
US 2019332867 A1	31 October 2019	CN 107066621 A	18 August 2017
		WO 2018205838 A1	15 November 2018
WO 2021017289 A1	04 February 2021	CN 110633627 A	31 December 2019

Form PCT/ISA/210 (patent family annex) (January 2015)

REFERENCES CITED IN THE DESCRIPTION

This list of references cited by the applicant is for the reader's convenience only. It does not form part of the European patent document. Even though great care has been taken in compiling the references, errors or omissions cannot be excluded and the EPO disclaims all liability in this regard.

Patent documents cited in the description

- CN 202110973390 **[0001]**