CONTENT CHECK MODEL TRAINING METHOD AND APPARATUS, VIDEO CONTENT CHECK METHOD AND APPARATUS, COMPUTER DEVICE, AND STORAGE MEDIUM

CROSS-REFERENCE TO RELATED APPLICATION

[0001] This application claims priority to the Chinese patent application No. 201911051711.0, filed on October 31, 2019, the content of which is herein incorporated by reference in its entirety.

TECHNICAL FIELD

[0002] The present disclosure relates to the field of video moderation technologies, and in particular, relates to a method and apparatus for training a content moderation model, a method and apparatus for moderating video content, and a computer device and a storage medium.

BACKGROUND

[0003] With the development of the Internet technology, the traffic of internet video increases dramatically in recent years, and a new user generated content (UGC), for example, a short video, a live streaming, and the like, causes the internet video to become increasingly rich.

[0004] In this case, a video of a large quantity of offensive content, for example, the video of terrorism, violence, pornography, gambling, and the like, is also generated and intended to propagate in the Internet.

[0005] Thus, before or after publish of the video, the content of the video needs to be moderated, and the video with offensive content is filtered out.

[0006] In one fashion, the video resource uploaded to the Internet is manually moderated to determine whether contains offensive content. However, as increase of the amount of the video resources on the Internet, manual moderation for the video content is time consuming and less efficient. Thus, the fashion of moderating video content may be the fashion of machine learning. In this fashion, whether image data (image frame) in a training video is offensive and an offense category of the image data are first manually marked, and then the image data and corresponding offense category are input into a machine learning model for training, and the other video content is identified using the trained model.

[0007] However, when marking the image data in each video to be trained, the quantity of image data is large, the marking operation is cumbersome, causing a lower efficiency of training the model and moderating the video with the module, and a higher cost of training model due to manual marking.

SUMMARY

[0008] Embodiments of the present disclosure provide a method and apparatus for training a content moderation model, a method and apparatus for moderating video content, and a computer device and a storage medium thereof, so as to solve a problem of lower efficiency of training model and moderating the video by the model, and high cost of training the model caused by marking the image data manually.

[0009] A method for training a content moderation model is provided. The method includes:

[0010] extracting part of image data of a sample video file as sample image data;

[0011] positioning a time point of the sample image data in the sample video file in the case that the sample image data contains offensive content;

[0012] extracting salient image region data from the image data around the time point; and

[0013] training the content moderation model based on the image region data and the sample image data.

[0014] A method for moderating video content is also provided. The method includes:

[0015] extracting part of image data of a target video file as target image data;

[0016] positioning a time point of the target image data in the target video file in the case that the target image data contains offensive content;

[0017] extracting salient image region data from the image data around the time point; and

[0018] moderating the content of the target video file by inputting the image region data and the sample image data into a preset content moderation model.

[0019] An apparatus for training a content moderation model is also provided. The method includes:

[0020] a sample image data extracting module, configured to extract part of image data of a sample video file as sample image data;

[0021] a time point positioning module, configured to position a time point of the sample image data in the sample video file in the case that the sample image data contains offensive content;

[0022] an image region data extracting module, configured to extract salient image region data from the image data around the time point; and

[0023] a model training module, configured to train the content moderation model based on the image region data and the sample image data.

[0024] An apparatus for moderating video content is also provided. The method includes:

[0025] a target image data extracting module, configured to extract part of image data of a target video file as target image data;

[0026] a time point positioning module, configured to position a time point the target image data in the target video file in the case that the target image data contains offensive content;

[0027] an image region data extracting module, configured to extract salient image region data from the image data around the time point; and

[0028] a video moderating module, configured to moderate the content of the target video file by inputting the image region data and the sample image data into a preset content moderation model.

[0029] A computer device is also provided. The computer device includes:

[0030] one or more processors;

[0031] a memory configured to store one or more programs;

[0032] wherein the one or more processors, when running the one or more programs, is caused to perform the method for training the content moderation model described above or the method for moderating the video content described above.

[0033] A computer readable storage medium is also provided. The computer readable storage medium stores a computer program, wherein the computer program, when run by a processor of a computer device, causes the computer device to perform the method for training the content moderation model described above or the method for moderating the video content described above.

BRIEF DESCRIPTION OF THE DRAWINGS

[0034] FIG. 1 is a flowchart of a method for training a content moderation model according to a first embodiment of the present disclosure;

[0035] FIG. 2 is a schematic diagram of training a content moderation model according to the first embodiment of the present disclosure;

[0036] FIG. 3 is a flowchart of a method for training a content moderation model according to a second embodiment of the present disclosure;

[0037] FIG. 4 is a flowchart of a method for moderating video content according to a third embodiment of the present disclosure;

[0038] FIG. 5 is a schematic diagram of moderating a video content according to the third embodiment of the present disclosure;

[0039] FIG. 6 is a flowchart of a method for moderating video content according to a fourth embodiment of the present disclosure;

[0040] FIG. 7 is a schematic structural diagram of an apparatus for training a content moderation model according to a fifth embodiment of the present disclosure;

[0041] FIG. 8 is a schematic structural diagram of an apparatus for moderating video content according to a sixth embodiment of the present disclosure; and

[0042] FIG. 9 is a schematic structural diagram of a computer device according to a seventh embodiment of the present disclosure.

DETAILED DESCRIPTION

[0043] The following describes the present disclosure hereinafter in connection with the accompanying drawings and the embodiments. Only some, but not all, of the structures associated with the present disclosure are shown in the accompanying drawings. The embodiments of the present disclosure and the features in the embodiments may be combined with each other.

[0044] First embodiment

[0045] FIG. 1 is a flowchart of a method for training a content moderation model according to a first embodiment of the present disclosure. This embodiment is applicable to cases where video data is marked automatically over time, space. The method may be performed by an apparatus for training a content moderation model, and the apparatus for training the content moderation model may be implemented over software and/or hardware, and may be configured in a computer device, e.g., a server, a workstation, a personal computer, or the like. The method includes the following processes.

[0046] In S101, a sample video file is received.

[0047] The sample video file, as a video, includes a plurality of frames of consecutive image data. In the case that the consecutive image data changes are greater than 24 frames per second, according to the visual persistence principle, human eyes cannot discern a single static image, and thus, a visual effect is smooth and consecutive.

[0048] In this embodiment, a video file may be acquired as a sample for training the content moderation model by grabbing the video file from the network, accumulating the video file uploaded by a user, downloading the video file from a published database, and the like in advance, and the sample for training the content moderation model is also referred to as the sample video file.

[0049] For different business scenarios, the formats and forms of the sample video file are different, which are not limited in this embodiment.

[0050] In an example, the format of the sample video file may include a Moving Picture Experts Group (MPEG), a Real Media Variable Bitrate (RMVB), an Audio Video Interleaved (AVI), a Flash Video (FLV), and the like.

[0051] The form of the sample video file may include a short video, a live streaming video, a movie, a television series, or the like.

[0052] In S102, part of image data of a sample video file is extracted as sample image data.

[0053] In this embodiment, part of image data may be selected from all of the image data of the sample video file as the sample image data.

[0054] In S103, a time point of the sample image data in the sample video file is positioned in the case that the sample image data contains offensive content.

[0055] In this embodiment, the content of the sample image data may be identified to determine whether the content is offensive content. The content of the sample image data may be determined as offensive content in the case that the content relates to terrorism, violence, pornography, gambling, and the like, and the content of the sample image data may be determined to be legal in the case that the content relates to a landscape, a building, or the like.

[0056] For the sample image data with offensive content, a time point of the sample image data in the sample video file may be positioned.

[0057] In S104, salient image region data is extracted from the image data around the time point.

[0058] Salience, as visual features of the image, represents the attention of human eyes on some regions of the image.

[0059] For a frame of image, the user is interested in a partial region of the image, and the interesting partial region represents the intent of the user. The majority of the remaining regions are independent of the intent of the user, that is, the salient region is a region in the image that is most likely to cause the interest of the user and represent the image content.

[0060] In fact, the selection of salience is subjective, and for the same frame image, different users may select different regions as salient regions due to different tasks and knowledge bases of the users.

[0061] The attention mechanism of the person is used to calculate the salience of the region. The study of cognitive psychology has shown that some regions in the image can significantly attract the attention of the person, and these regions include a larger amount of information. Therefore, the attention mechanism of the person can be simulated based on a mathematical model, and the extracted salient regions are more compliant with the subjective evaluation of the person as the general rule in the image cognitive process is utilized.

[0062] On the timeline of the sample video file, a plurality of frames of the image data are present around the time point of the sample image data. In this embodiment, the salient regions may be extracted from the image data as the image region data.

[0063] For the sample video file, the scene is generally not transformed within a short period of time. That is, the other image data around the sample image data is substantially the same as the

content of the sample image data. In the case that the sample image data contains offensive content, it is highly probable that the image data content is offensive, and thus, the image data content is also considered to be offensive. Therefore, based on the sensitivity of the user to the offensive content related to terrorism, violence, pornography, gambling, and the like, the salient image region data in the image data is primarily focused on terrorism, violence, pornography, gambling, and the like.

[0064] In S105, the content moderation model is trained based on the image region data and the sample image data.

[0065] In this embodiment, the sample video file may be marked with an offense category of the offensive content in advance. By training over a preset network with a training sample of the image region data and the sample image data of different sample video files, and a tag of the offense category, the content moderation model is acquired in the case that the training is completed.

[0066] In an example, the network may include a machine model, for example, a support vector machine (SVM), a random forest, an Xgboost, and a neural network, for example, a convolutional neural network (CNN), a deep neural network (DNN), and a recurrent neural network (RNN), which is not limited in this embodiment.

[0067] In this embodiment, the DNN is illustrated as an example of the content moderation model.

[0068] In this embodiment, the offense category, for example, terrorism, violence, pornography, and the like, marked on the sample video file and representing the offensive content is determined.

[0069] The deep neural network and a pre-trained model are acquired. The pre-trained model is a deep learning architecture that has been trained to perform specific tasks (for example, identifying classification in a picture) with a large amount of data, and includes VGG, Inception, ResNet, MobileNet, NasNet, and the like.

[0070] The deep neural network is initialized using the pre-trained model. That is, by using the pre-trained model trained with a large dataset before use, the corresponding configuration and weight may be directly applied to the deep neural network to implement the migration learning.

[0071] By backpropagation, the deep neural network is trained as the content moderation model based on the image region data, the sample image data, and the offense category.

[0072] In an example, the image region data and the sample image data are input into the deep neural network, the original pixel information is fused by a non-linear mapping between neurons, and the scores of the different offense categories are acquired by a Softmax regression layer and

act as the offense score. A classification loss of the entire deep neural network is acquired by calculating a cross-entropy of the offense score and the tag of the training sample.

[0073] In an example, where the sample video file belongs to different offense categories, the content moderation model may be configured to identify different offense categories. Where the sample video file belongs to the same offense category, the content moderation model may be configured to identify the offense category.

[0074] The method for training the content moderation model in this embodiment is illustrated by the following examples.

[0075] For example, as shown in FIG. 2, for a sample video file 201 with the content being a boxing match, six frames of image data are extracted from the sample video file as a sample image data 202, four frames of the image data 202 containing offensive content are determined as a sample image data 203, and the sample image data 203 containing offensive content involves violence. A time point of the sample image data 203 containing offensive content is positioned on the timeline 204 of the sample video file 201, salient image region data is extracted from the image data 205 around the time point (block portion). The content moderation model 207 is trained with a training sample of the salient image region data in the sample image data 202 and the image data 205, and a tag of the offense category 206 of the sample video file 201, such that the content moderation model 207 can be configured to classify the image data, and the dimension of the classification is consistent with the offense category 206.

[0076] In this embodiment, the sample video file is received. The sample video file includes a plurality of frames of the image data, part of the image data is extracted as the sample image data, and the time point of the sample image data in the sample video file is positioned in the case that the sample image data contains offensive content. The salient image region data is extracted from the image data around the time point, and the content moderation model is trained based on the image region data and the sample image data. The sample image data containing offensive content is positioned in time, and the salient image region data is positioned on space, thereby achieving spatiotemporal positioning of the sample video file, i.e., self-positioning of the offensive content of the sample video file in time and on space dimensions. Therefore, the feature of the offensive content is quickly sampled from the sample video file for characterization, the quality of the feature is increased in the aspects of time and space for content moderation, and the performance of the content moderation model may be ensured by training the content moderation model in the fashion. In addition, automatic tagging of the sample image data is achieved without adding annotation effort by positioning the sample image data and the image region data with the offensive content, which is easy for operation and no

manually tagged, improves efficiency of training the content moderation model, and reduces the cost of training the content moderation model.

[0077] Second embodiment

[0078] FIG. 3 is a flowchart of a method for training a content moderation model according to a second embodiment of the present disclosure. This embodiment illustrates the operations of extracting sample image data, positioning time point, and extracting image region data based on above embodiment. The method includes the following processes.

[0079] In S301, a sample video file is received.

[0080] The sample video file includes a plurality of frames of image data.

[0081] In S302, the sample video file is partitioned into at least two sample video segments.

[0082] In S303, part of the image data is extracted from each sample video segment as the sample image data.

[0083] In an example, the sample video file may be segmented at a time interval, i.e., the sample video file is partitioned into at least two sample video segments.

[0084] n frames of the image data are extracted randomly from each sample video segment as the sample image data, and thus a sequence of video frames to be processed is formed.

[0085] t and n are both adjustable parameters.

[0086] In an example, except for the average partition of sample video file and the random extraction of the image, the sample image data may be extracted in other fashions according to the actual needs. For example, for the sample video file including barrages, the barrage may represent the interest of the user in the content of the sample video file to some extent, and the sample video file is partitioned with respect to the barrages, such that the density of the barrages in each sample video segment (the quantity within a unit time) is within a set range. Alternatively, the image data is extracted from each sample video segment, such that the time interval between every two frames of the image data is equal, and the like, which is not limited in this embodiment.

[0087] Furthermore, the sample image data may be scaled to a preset size, and the sample image data may further be ranked in time order, thereby facilitating training of the content moderation model.

[0088] In S304, an offense discrimination model is looked up.

[0089] In this embodiment, the image data containing offensive content may be pre-used as the training sample, and a network (e.g., CNN) is trained with a tag of the offense category. In the case that the training is completed, the offense discrimination model can be acquired. That is, the offense discrimination model can be configured to identify an image offense score of the content in the image data.

[0090] Optionally, different offense discrimination models can be trained for different offense categories. That is, an offense discrimination model can be configured to identify an image offense score in the image data of which the content belongs to the same offense category.

[0091] Generally, the sample video file is consistent with the offense category of the sample image data. Therefore, the offense category marked on the sample video file and representing the offensive content may be determined, the offense discrimination model corresponding to the offense category is looked up, and the offense discrimination model is configured to identify the image offense score in the image data of which the content belongs to the offense category.

[0092] A generic offense discrimination model may also be trained for different offense categories, i.e., one offense discrimination model can be configured to identify the image offense score in the image data of which the content belongs to different offense category, which are not limited in this embodiment.

[0093] In S305, an image offense score of the content of the sample image data is identified by inputting the sample image data into the offense discrimination model.

[0094] In the case that the offense discrimination model is determined, the sample image data of the sample video file may be sequentially input into the offense discrimination model for processing, and the offense discrimination model outputs the image offense score of the sample image data sequentially.

[0095] In S306, the sample image data with the image offense score meeting a preset offense condition is selected.

[0096] In this embodiment, the offense condition may be preset, and the offense condition used to determine the sample image data containing offensive content.

[0097] In the case that the image offense score of the sample image data is determined, the sample image data with the image offense score meeting the offense condition is determined.

[0098] In an example, the offense condition is that the image offense score is greater than an image score threshold, or the value of the image offense score is greatest.

[0099] In this embodiment, for the sample image data, the image offense score of the sample image data may be determined whether to be greater a preset image score threshold.

[00100] In the case that the image offense score of the sample image data is greater than the preset image score threshold, the image offense score is determined to meet the preset offense condition.

[00101] In the case that the image offense score of the sample image data is greater than the preset image score threshold, the image offense score is determined to meet the preset offense condition.

[00102] The above offense condition is merely an example, and when this embodiment is implemented, other offense condition may be set according to actual needs, for example, the top m of the value of the image offense score and the like, which is not limited in the embodiments of the present disclosure. In addition, other offense conditions, except for above offense condition, may be applied according to actual needs, which is not limited in the embodiments of the present disclosure.

[00103] In S307, the time point of the sample image data meeting the preset offense condition in the sample video file is determined.

[00104] In the case that a frame of sample image data is determined to meet the offense condition, the time point of the sample image data in the sample video file is determined.

[00105] In S308, a time range containing the time point is determined.

[00106] On the timeline of the sample video file, the time range containing the point is generated based on the time point.

[00107] In an example, assuming that the time point is T, then the time range of duration F[T-F/2, T+F/2] may be generated, wherein F is an adjustable parameter.

[00108] The fashion of generating the time range described above is merely an example, and when this embodiment is implemented, other fashions of generating the time range may be employed according to actual needs, for example, [T−F/3, T+2F/3], [T−3F/4, T+F/4], and the like, which is not limited in the embodiments of the present disclosure. In addition, except for the fashion of generating the time range, other fashions of generating the time range may be employed according to actual needs, which is not limited in the embodiments of the present disclosure.

[00109] In S309, a salient region detection model is looked up.

[00110] In this embodiment, the salient region detection model is also preset, and the salient region detection model can be configured to identify a salient image region in the image data.

[00111] In an example, the salient region detection model may be applied in three classes of algorithms as follows.

[00112] A first class is a salience analysis algorithm based on low-level vision feature, for example, a salience algorithm (Itti algorithm), which is a selective attention algorithm that simulates a visual attention mechanism of an organism, and is adapted to process natural images.

[00113] A second class is a complete mathematical calculation method that is not based on any biological visual principle, for example, a full-resolution algorithm (Aho-Corasick, AC algorithm), a spectral residual algorithm (SR algorithm) based on a spatial frequency domain.

[00114] A third class combines the above two classes of the algorithms, for example, a graph theory based algorithm (GBVS algorithm), which simulates the visual principle similar to the Itti algorithm in the feature extraction process, and introduces a Markov chain in the salient image generation process and acquires a salience value with a complete mathematical calculation method.

[00115] In S310, the salient image region data in the image data is identified by inputting the image data within the time range into the salient region detection model.

[00116] The image data within the time range is extracted from the sample video file, the image data is sequentially input into the salient region detection model for processing, and the salient region detection model outputs the salient image region data in the image data.

[00117] In S311, the content moderation model is trained based on the image region data and the sample image data.

[00118] In this embodiment, for the sample image data containing offensive content, the time range is determined based on the time point, and it is highly probable that the content of the image data within the time range is offensive. Thus, extracting the salient image region data may quickly expand the training sample with reliability, thereby improving the performance of the content moderation model by training the content moderation model.

[00119] Third embodiment

[00120] FIG. 4 is a flowchart of a method for moderating video content according to a third embodiment of the present disclosure. This embodiment is applicable to the case where video data is moderated in time and space. The method may be performed by an apparatus for moderating video content, and the apparatus for training a content moderation model may be implemented over software and/or hardware and may be configured in a computer device, e.g., a server, a workstation, a personal computer, and the like. The method includes the following processes.

[00121] In S401, a target video file is received.

[00122] The target video file, as a video, includes a plurality of frames of consecutive image data. In the case that the consecutive image data changes are greater than 24 frames per second, according to the visual persistence principle, human eyes cannot discern a single static image, and thus, a visual effect is smooth and consecutive.

[00123] For different business scenarios, the formats and forms of the target video file are different, which are not limited in this embodiment.

[00124] The format of the target video file may include MPEG, RMVB, AVI, FLV, and the like.

[00125] The form of the target video file may include a short video, a live streaming video, a movie, a television series, or the like.

[00126] The user uploads the target video file to the computer device, and is intended to publish the target video file to allow the public to review and browse.

[00127] The computer device may develop moderation criteria according to business, legality, and other factors. Prior to publishing the target video file, the content of the target video file is moderated based on the moderation criteria, some target video files that do not comply with the review criteria are filtered out, for example, the content of the target video file involving terrorism, violence, pornography, gambling, and the like, and some target video files that comply with the moderation criteria are published.

[00128] For the target video file with a higher requirement for timeliness, for example, short video, live streaming video, and the like, a streaming real-time system may be provided. The user uploads the target video file to the streaming real-time system over a client, and the streaming real-time system may transfer the target video file to the computer device for moderation.

[00129] For the target video file with a lower requirement for timeliness, for example, movies, television series, and the like, a database may be provided, for example, a distributed database and the like. The user uploads the computer device to the database through the client, and the computer device may read the target video file from the database for moderation.

[00130] In S402, part of the image data of the target video file is extracted as the target image data.

[00131] In this embodiment, part of the image data may be selected from all image data of the target video file as the target image data.

[00132] In S403, a time point of the target image data in the target video file is positioned in the case that the target image data contains offensive content.

[00133] In this embodiment, the content of the target image data may be identified to determine whether the content of the target image data is offensive content. The content of the target image data may be determined as offensive content in the case that the content relates to terrorism, violence, pornography, gambling, and the like, and the content of the target image data may be determined to be legal in the case that the content relates to a landscape, a building, or the like.

[00134] For the target image data containing offensive content, a time point of the target image data in the target video file may be positioned.

[00135] S404, salient image region data is extracted from the image data around the time point.

[00136] Salience, as visual features of the image, represents the attention of human eyes on some regions of the image.

[00137] For a frame of image, the user is interested in a partial region of the image, and the interesting partial region represents the intent of the user. The majority of the remaining regions are independent of the intent of the user, that is, the salient region is a region in the image that is most likely to cause the interest of the user and represent the image content.

[00138] In fact, the selection of salience is subjective, and for the same frame image, different users may select different regions as salient regions due to different tasks and knowledge bases of the users.

[00139] The attention mechanism of the person is used to calculate the salience of the region. The study of cognitive psychology has shown that some regions in the image can significantly attract the attention of the person, and these regions include a larger amount of information. Therefore, the attention mechanism of the person can be simulated based on a mathematical model, and the extracted salient regions are more complied with the subjective evaluation of the person as the general rule in the image cognitive process is utilized.

[00140] On the timeline of the target video file, a plurality of frames of the image data are present around the time point of the target image data. In this embodiment, the salient regions may be extracted from the image data as the image region data.

[00141] For the target video file, the scene is generally not transformed within a short period of time. That is, the other image data around the target image data is substantially the same as the content of the target image data. In the case that the target image data contains offensive content, it is highly probable that the image data content is offensive, and thus, the image data content is also considered as offensive content. Therefore, based on the sensitivity of the user to the offensive data related to terrorism, violence, pornography, gambling, and the like, the salient image region data in the image data is primarily focused on terrorism, violence, pornography, gambling, and offensive content.

[00142] In S405, the content of the target video file is moderated by inputting the image region data and the target image data into a preset content moderation model.

[00143] In this embodiment, the content moderation model may be pre-trained, and the content moderation model may be configured to identify a file offense score in the case that the content of the target video file belongs to a preset offense category.

[00144] As the training fashion of the content moderation model is substantially similar to the application of the first embodiment, the second embodiment described above, the description is briefly, and the reference may be made for parts of description of the first embodiment, the second embodiment, which is not described in this embodiment herein.

[00145] For the target video file, the image region data and the target image data may be input into the content moderation model for processing, and the content of the target video file may be moderated based on the result output from the content moderation model to determine whether the content is offensive content.

[00146] In an example, the image region data and the target image data are input into a preset content moderation model to determine the file offense score in the case that the content of the target video file belongs to the preset offense category.

[00147] A file score threshold is determined.

[00148] The file offense score is compared with the file score threshold.

[00149] In the case that the file offense score is less than or equal to the file score threshold, it is less probable that the target video file content is offensive , and the content of the target video file may be determined to be legal.

[00150] In the case that the file offense score is greater than the file score threshold, it is highly probable that the target video file content is offensive, and the target video file may be distributed to a designated client as a moderation task. The client is managed by a specialized moderator.

[00151] In the case that the client receives the moderation task, the moderator may browse the corresponding target video file to manually determine whether the content of the target video file is offensive.

[00152] The content of the target video file is determined to be legal in the case that first moderation information is received from the client.

[00153] The content of the target video file is determined to be offensive in the case that second moderation information is received from the client.

[00154] In a fashion of determining the file score threshold, a total quantity of the target video files with a previous time period (e.g., a previous day), and the file offense score of the target video file has been determined.

[00155] The file score threshold is generated, such that a ratio of a moderation quantity to the total quantity matches a preset push ratio, wherein the moderation quantity is a quantity of the target video files of which the file offense scores are greater than the file score threshold.

[00156] Generally, the ratio of the quantity of target video files with the offensive content to the quantity of all target video files is relatively low, for example, 1%. In this determination fashion, a push ratio higher than the ratio, for example, 10%, may be set to ensure that the target video file containing offensive content is moderated manually as possible.

[00157] Assuming that the quantity of all target video files within the previous time period is 100,000, and the push ratio is 10%, then the target video file may be ranked based on the file

offense score (from small to large), and the file offense score of the 10000th target video file is set as the score threshold.

[00158] The above fashion of determining the file score threshold is merely an example, and when this embodiment is implemented, other fashion of determining the file score threshold may be set according to actual needs, for example, setting a default value as the file score threshold, and the like, which is not limited in the embodiments of the present disclosure. In addition, other fashion of determining the file score threshold, except for above offense condition, may be applied according to actual needs, which is not limited in the embodiments of the present disclosure.

[00159] The method for moderating the video content in this embodiment is illustrated by the following embodiments.

[00160] For example, as shown in FIG. 5, for a sample video file 501 with the content being a boxing match, six frames of image data are extracted from the sample video file as a target image data 502, four frames of the image data 502 containing offensive content is determined as a target image data 503, and the target image data 503 containing offensive content involves violence. A time point of the target image data 503 containing offensive content is positioned on the timeline 504 of the target video file 501, salient image region data is extracted from the image data 505 around the time point (block portion). The salient image region data in the target image data 502 and the image data 505 are input into the content moderation model 506, and a tag 507, i.e., a legal or offensive class, of the target video file 501 is identified based on the result output by the content moderation model 506.

[00161] In this embodiment, the target video file is received. The target video file includes a plurality of frames of the image data, part of the image data is extracted as the target image data, and the time point of the target image data in the target video file is positioned in the case that the target image data contains offensive content. The salient image region data is extracted from the image data around the time point, and the image region data and the target image data are input into a preset content moderation model to moderate the content of the target video file. The target image data containing offensive content is positioned in time, and the salient image region data is positioned on space, thereby achieving spatiotemporal positioning of the target video file, i.e., self-positioning of the offense data of the target video file in time and on space dimensions. Therefore, the feature of the offensive content is quickly sampled from the sample video file for characterization, and the quality of the feature is increased in the aspects of time and space for content moderation, thereby ensuring the quality of the content moderation, reducing the rate of mis-moderation, and improving the efficiency of moderating the video content.

[00162] Fourth embodiment

[00163] FIG. 6 is a flowchart of a method for moderating video content according to a fourth embodiment of the present disclosure, and this embodiment illustrates the operations of extracting target image data, positioning time point, and extracting image region data based on above embodiments. The method includes the following processes.

[00164] In S601, a target video file is received.

[00165] The target video file includes a plurality of frames of image data.

[00166] In S602, the target video file is partitioned into at least two target video segments.

[00167] In S603, part of the image data is extracted from each target video segment as the target image data.

[00168] In an example, the target video file may be segmented at a time interval, i.e., the target video file is partitioned into at least two target video segments.

[00169] n frames of the image data are extracted randomly from each target video segment as the target image data, and thus a sequence of video frames to be processed is formed.

[00170] t and n are both adjustable parameters.

[00171] Except for the average partition of target video file and the random extraction of the image, the target image data may be extracted in other fashions according to the actual needs. For example, when the target video file is partitioned, the duration of the target video segments at both ends is longer, the duration of the middle target video segment is shorter. Alternatively, the image data is extracted from each target video segment, such that the time interval between every two frames of the image data is equal, and the like, which is not limited in this embodiment.

[00172] Furthermore, the target image data may be scaled to a preset size, and the target image data may further be ranked in time order, thereby facilitating identifying of the content moderation model.

[00173] In S604, an offense discrimination model is looked up.

[00174] In this embodiment, the image data containing offensive content may be pre-used as the training sample, and a network (e.g., CNN) is trained with a tag of the offense category. In the case that the training is completed, the offense discrimination model can be acquired. That is, the offense discrimination model can be configured to identify an image offense score of the content in the image data.

[00175] Optionally, different offense discrimination models may be trained for different offense categories. That is, an offense discrimination model can be configured to identify an image offense score in the image data of which the content belongs to the same offense category.

[00176] A generic offense discrimination model may also be trained for different offense categories, i.e., one offense discrimination model can be configured to identify the image offense score in the image data of which the content belongs to different offense category, which are not limited in this embodiment.

[00177] In S605, an image offense score of the content of the target image data is identified by inputting the target image data into the offense discrimination model.

[00178] In the case that the offense discrimination model is determined, the target image data of the target video file may be sequentially input into the offense discrimination model for processing, and the offense discrimination model outputs the image offense score of the target image data sequentially.

[00179] In S606, the target image data with the image offense score meeting a preset offense condition is selected.

[00180] In this embodiment, the offense condition may be preset, and the offense condition used to determine the target image data containing offensive content.

[00181] In the case that the image offense score of the target image data is determined, the target image data with the image offense score meeting the offense condition is determined.

[00182] In an example, the offense condition is that the image offense score is greater than an image score threshold, or the value of the image offense score is greatest.

[00183] In this embodiment, for the target image data, the image offense score of the target image data may be determined whether to be greater a preset image score threshold.

[00184] In the case that the image offense score of the target image data is greater than the preset image score threshold, the image offense score is determined to meet the preset offense condition.

[00185] In the case that the image offense score of the target image data is greater than the preset image score threshold, the image offense score is determined to meet the preset offense condition.

[00186] The above offense condition is merely an example, and when this embodiment is implemented, other offense condition may be set according to actual needs, for example, the top m of the value of the image offense score, and the like, which is not limited in the embodiments of the present disclosure. In addition, other offense condition, except for above offense condition, may be applied according to actual needs, which is not limited in the embodiments of the present disclosure.

[00187] In S607, the time point of the target image data meeting the preset offense condition in the target video file is determined.

[00188] In the case that a frame of target image data is determined to meet the offense condition, the time point of the target image data in the target video file is determined.

[00189] In S608, a time range containing the time point is determined.

[00190] On the timeline of the target video file, the time range containing the point is generated based on the time point.

[00191] In an example, assuming that the time point is T, then the time range of duration F[T-F/2, T+F/2] may be generated, wherein F is an adjustable parameter.

[00192] The fashion of generating the time range described above is merely an example, and when this embodiment is implemented, other fashion of generating the time range may be set according to actual needs, for example, [T−F/3, T+2F/3], [T−3F/4, T+F/4], and the like, which is not limited in the embodiments of the present disclosure. In addition, except for the fashion of generating the time range, other fashion of generating the time range may be applied according to actual needs, which is not limited in the embodiments of the present disclosure.

[00193] In S609, a salient region detection model is looked up.

[00194] In this embodiment, the salient region detection model is also preset, and the salient region detection model may be configured to identify a salient image region in the image data.

[00195] In an example, the salient region detection model may be applied in three classes of algorithms as follows.

[00196] A first class is a salience analysis algorithm based on low-level vision feature, for example, a salience algorithm (Itti algorithm), which is a selective attention algorithm that simulates a visual attention mechanism of an organism, and is adapted to process natural images.

[00197] A second class is a complete mathematical calculation method that is not based on any biological visual principle, for example, a full-resolution algorithm (Aho-Corasick, AC algorithm), a spectral residual algorithm (SR algorithm) based on a spatial frequency domain.

[00198] A third class combines the above two classes of the algorithms, for example, a graph theory based algorithm (GBVS algorithm), which simulates the visual principle similar to the Itti algorithm in the feature extraction process, and introduces a Markov chain in the salient image generation process and acquires a salience value with a complete mathematical calculation method.

[00199] In S610, the salient image region data in the image data is identified by inputting the image data within the time range into the salient region detection model.

[00200] The image data within the time range is extracted from the target video file, the image data is sequentially input into the salient region detection model for processing, and the salient region detection model outputs the salient image region data in the image data.

[00201] In S611, the content of the target video file is moderated by inputting the image region data and the target image data into a preset content moderation model.

[00202] In this embodiment, for the target image data containing offensive content, the time range is determined based on the time point, and it is highly probable that the content of the image data within the time range is offensive. Thus, extracting the salient image region data may quickly expand the training sample with reliability, thereby improving the performance of the content moderation model by training the content moderation model.

[00203] Fifth embodiment

[00204] FIG. 7 is a schematic structural diagram of an apparatus for training a content moderation model according to a fifth embodiment of the present disclosure. The apparatus includes:

[00205] a sample video file receiving module 701, configured to receive a sample video file, wherein the sample video file includes a plurality of frames of image data;

[00206] a sample image data extracting module 702, configured to extract part of image data of a sample video file as sample image data;

[00207] a time point positioning module 703, configured to position a time point of the sample image data in the sample video file in the case that the sample image data contains offensive content;

[00208] an image region data extracting module 704, configured to extract salient image region data from the image data around the time point; and

[00209] a model training module 705, configured to train the content moderation model based on the image region data and the sample image data.

[00210] The apparatus for training the content moderation model according to the embodiments of the present disclosure may perform the method for training the content moderation model according to any of the embodiments of the present disclosure, and has the functional modules and benefits corresponding to the method.

[00211] Sixth embodiment

[00212] FIG. 8 is a schematic structural diagram of an apparatus for moderating video content according a sixth embodiment of the present disclosure. The apparatus includes:

[00213] a target video file receiving module 801, configured to receive a target video file, wherein the target video file includes a plurality of frames of image data;

[00214] a target image data extracting module 802, configured to extract part of image data of a target video file as target image data;

[00215] a time point positioning module 803, configured to position a time point of the target image data in the target video file in the case that the target image data contains offensive content;

[00216] an image region data extracting module 804, configured to extract salient image region data from the image data around the time point; and

[00217] a video moderating module 805, configured to moderate the content of the target video file by inputting the image region data and the target image data into a preset content moderation model.

[00218] The apparatus for moderating video content according to the embodiments of the present invention may perform the method for moderating the video content according to any of the embodiments of the present disclosure, and has the functional modules and benefits corresponding to the method.

[00219] Seventh embodiment

[00220] FIG. 9 is a schematic structural diagram of a computer device according to a seventh embodiment of the present disclosure. As shown in FIG. 9, the computer device includes a processor 900, a memory 901, a communication module 902, an input device 903, and an output device 904. A quantity of the processors 900 in the computer device may be one or more, one processor 900 is taken as an example in FIG. 9. The processor 900, the memory 901, the communication module 902, the input device 903, and the output device 904 in the computer device may be connected by a bus or other means, and the bus is taken as an example in FIG. 9.

[00221] The computer device according to this embodiment may perform the method for training the content moderation model or the method for moderating the video content according to any of the embodiments of the present disclosure, and has the functional modules and benefits corresponding to the method.

[00222] Eighth embodiment

[00223] This embodiment provides a computer-readable storage medium storing a computer program. The computer program, when run by a processor of a computer device, causes the computer device to perform the method for training the content moderation model or the method for moderating the video content.

[00224] The method for training the content moderation model includes:

[00225] receiving a sample video file, wherein the sample video file includes a plurality of frames of image data;

[00226] extracting part of image data as sample image data

[00227] positioning a time point of the sample image data in the sample video file in the case that the sample image data contains offensive content;

[00228] extracting salient image region data from the image data around the time point; and

[00229] training the content moderation model based on the image region data and the sample image data.

[00230] The method for moderating the video content includes:

[00231] receiving a target video file, wherein the target video file includes a plurality of frames of image data;

[00232] extracting part of image data as target image data;

[00233] positioning a time point of the target image data in the target video file in the case that the target image data contains offensive content;

[00234] extracting salient image region data from the image data around the time point; and

[00235] moderating the content of the target video file by inputting the image region data and the target image data into a preset content moderation model.

[00236] For the computer readable storage medium according to the embodiment of the present disclosure, the computer program is not limited to the operation method described above, and may also perform related operations in the method for training the content moderation model or the method for moderating the video content according to any of the embodiments of the present disclosure.

[00237] Based on the above description of the embodiments, the present disclosure may be implemented through software and the necessary generic hardware, or may be implemented through hardware. The technical solutions of the present disclosure may be embodied in the form of a software product, and the software product may be stored in the computer-readable storage medium, for example, a floppy disk of a computer, a read-only memory (ROM), a random-access memory (RAM), a FLASH memory, a hard disk, an optical disk, and the like. The computer-readable storage medium includes various instructions to cause the computer device (may be a personal computer, a server, or a network device, and the like) to perform the methods described in various embodiments of the present disclosure.

[00238] In the above embodiment of the apparatus for training the content moderation model and the apparatus for moderating video content, the included units and modules are merely divided according to function logic, but are not limited to the division described above, so long as the corresponding functions can be implemented. In addition, the names of the

functional units are also merely for distinction, and are not intended to limit the scope of the present disclosure.

CLAIMS

What is claimed is:

1. A method for training a content moderation model, comprising:

extracting at least one frame of image data of a sample video file as sample image data;

positioning a time point of the sample image data on a timeline of the sample video file in the case of determining the sample image data contains offensive content by an offense discrimination model;

determining a time range of the sample video file, wherein a duration of the time range is a target duration, and the time range contains the time point;

looking up a salient region detection model configured to identify a salient image region in the image data;

identifying, by inputting a plurality of frames of image data within the time range of the sample video file into the salient region detection model, salient image region data in the plurality of frames of image data; and

training over a preset network with a training sample of the salient image region data and the sample image data, and a tag of an offense category marked on the sample video file and representing the offensive content, to acquire the content moderation model, wherein input of the preset network comprises the salient image region data and the sample image data, and the content moderation model is configured to moderate content of a target video file, to determine whether the content of the target video file is offensive.

2. The method according to claim 1, wherein extracting the at least one frame of image data of the sample video file as the sample image data comprises:

partitioning the sample video file into at least two sample video segments; and

extracting the at least one frame of the image data from each sample video segment as the sample image data.

3. The method according to claim 2, wherein extracting the at least one frame of image data of the sample video file as the sample image data comprises at least one of:

chronologically ranking the sample image data; and

scaling the sample image data to a preset size.

4. The method according to claim 1, wherein positioning the time point of the sample image data in the sample video file in the case that the sample image data contains the offensive content comprises:

looking up an offense discrimination model configured to identify an image offense score of the content in the image data;

identifying the image offense score of the content in the sample image data by inputting the sample image data into the offense discrimination model;

selecting the sample image data with the image offense score meeting a preset offense condition; and

determining the time point of the sample image data meeting the preset offense condition in the sample video file.

5. The method according to claim 4, wherein looking up the offense discrimination model comprises:

determining the offense category marked on the sample video file and representing the offensive content; and

looking up the offense discrimination model corresponding to the offense category, wherein the offense discrimination model is configured to identify the image offense score of the content that belongs to the offense category in the image data.

6. The method according to claim 4, wherein selecting the sample image data with the image offense score meeting the preset offense condition comprises:

determining, for the sample image data, whether the sample image data comprises the image offense score that is greater than a preset image score threshold;

determining that the image offense score meets the preset offense condition in the case that the sample image data comprises the image offense score that is greater than the preset image score threshold; and

determining that the image offense score of a maximum value meets the preset violation condition in the case that the sample image data does not comprise the image offense score that is greater than the preset image score threshold.

7. The method according to any one of claims 1 to 6, wherein the content moderation model is trained by:

determining the offense category marked on the sample video file and representing the offensive content;

acquiring a deep neural network and a pre-trained model;

initializing the deep neural network by the pre-trained model;

training, by backpropagation, the deep neural network as the content moderation model based on the image region data, the sample image data, and the offense category.

8. A method for moderating video content, comprising:

extracting at least one frame of image data of a target video file as target image data;

positioning a time point of the target image data on a timeline of the target video file in the case of determining the target image data contains offensive content by an offense discrimination model;

determining a time range of the target video file, wherein a duration of the time range is a target duration, and the time range contains the time point;

looking up a salient region detection model configured to identify a salient image region in the image data;

identifying, by inputting a plurality of frames of image data within the time range of the target video file into the salient region detection model, salient image region data in the plurality of frames of image data; and

moderating the content of the target video file by inputting the salient image region data and the target image data into a preset content moderation model.

9. The method according to claim 8, wherein moderating the content of the target video file by inputting the salient image region data and the target image data into the preset content moderation model comprises:

determining a file offense score of the content of the target video file by inputting the image region data and the sample image data into a preset content moderation model in the case that the content of the target video file belongs to a preset offense category;

determining a file score threshold;

determining the content of the target video file to be legal in the case that the file offense score is less than or equal to the file score threshold.

10. The method according to claim 9, wherein moderating the content of the target video file by inputting the salient image region data and the target image data into the preset content moderation model further comprises:

distributing the target video file to a designated client in the case that the file offense score is greater than the file score threshold;

determining the content of the target video file to be legal in the case that first moderation information is received from the client; and

determining the content of the target video file to be offensive in the case that second moderation information is received from the client.

11. The method according to claim 9 or 10, wherein determining the file score threshold comprises:

determining a total quantity of the target video files with a previous time period, wherein the file offense score of the target video file has been determined;

generating the file score threshold, such that a ratio of a moderation quantity to the total quantity matches a preset push ratio, wherein the moderation quantity is a quantity of the target video files of which the file offense scores are greater than the file score threshold.

12. An apparatus for training a content moderation model, comprising:

a sample image data extracting module, configured to extract at least one frame of image data of a sample video file as sample image data;

a time point positioning module, configured to position a time point of the sample image data on a timeline of the sample video file in the case of determining the sample image data contains offensive content by an offense discrimination model;

an image region data extracting module, configured to determine a time range of the sample video file, wherein a duration of the time range is a target duration, and the time range contains the time point; look up a salient region detection model configured to identify a salient image region in the image data; and identify, by inputting a plurality of frames of image data within the time range of the sample video file into the salient region detection model, salient image region data in the plurality of frames of image data; and

a model training module, configured to train over a preset network with a training sample of the salient image region data and the sample image data, and a tag of an offense category marked on the sample video file and representing the offensive content, to acquire the content moderation model, wherein input of the preset network comprises the salient image region data and the sample image data, and the content moderation model is configured to moderate content of a target video file, to determine whether the content of the target video file is offensive.

13. An apparatus for moderating video content, comprising:

a target image data extracting module, configured to extract at least one frame of image data of a target video file as target image data;

a time point positioning module, configured to position a time point of the target image data on a timeline of the target video file in the case of determining the target image data contains offensive content by an offense discrimination model;

an image region data extracting module, configured to determine a time range of the target video file, wherein a duration of the time range is a target duration, and the time range contains the time point; look up a salient region detection model configured to identify a salient image region in the image data; and identify, by inputting a plurality of frames of image data within the time range of the target video file into the salient region detection model, salient image region data in the plurality of frames of image data; and

a video moderating module, configured to moderate the content of the target video file upon inputting the salient image region data and the target image data into a preset content moderation model.

14. A computer device comprising:

one or more processors;

a memory configured to store one or more programs;

wherein the one or more processors, when running the one or more programs, is caused to perform the method for training the content moderation model as defined in any one of claims 1 to 7 or the method for moderating the video content as defined in any one of claims 8 to 11.

15. A computer readable storage medium, storing a computer program, wherein the computer program, when run by a processor of a computer device, causes the computer device to perform the method for training the content moderation model as defined in any one of claims 1 to 7 or the method for moderating the video content as defined in any one of claims 8 to 11.