

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.