LIVE STREAMING ~~AUDITING~~ --MODERATING METHOD AND APPARATUS, SERVER, AND STORAGE MEDIUM

[0001] The present disclosure is based on and claims priority to Chinese Patent Application No. 202011613190.6, filed on December 30, 2020, the disclosure of which is herein incorporated by reference in its entirety.

TECHNICAL FIELD

[0002] The present disclosure relates to the field of moderation of violation content in Internet resources, for example, relates to a method and apparatus for moderating living streaming, a server, and a storage medium.

BACKGROUND OF THE INVENTION

[0003] With the rapid development of Internet technology, there are more and more live streaming platforms on the Internet, such that Internet resources are spread more and more extensively. Meanwhile, as live streaming content of some live streaming rooms involves topics of terrorism, violence, pornography, political sensitivity, and the like, a large number of violation live streaming content is quickly spread over the Internet. Therefore, violation live streaming content needs to be moderated in real time for this type of live streaming rooms to prevent the spread of the violation live streaming content.

[0004] Violation content analysis is performed on live streaming visual features embodied by a plurality of live streaming video frames and live streaming audio features of live streaming voices in a live streaming process by a neural network model trained by single features, such that whether violation pictures exist in the plurality of live streaming video frames or whether violation audios related to terrorism, violence, pornography, political sensitivity, or other topics exist in the live streaming voices is determined, and suspicious violation live streaming rooms are found by moderation and are pushed to a manual moderation platform for manual moderation. In this case, in extracting the live streaming video frames and the live streaming voices in the live streaming process, a problem of inaccurate extraction results due to the instability of a live streaming network is present. Therefore, in reviewing whether violation content exists in a live streaming room based on the extracted live streaming video frames and

live streaming voices, the accuracy and recall rate of violation moderation of the live streaming room are not ensured, and false rejection or false acceptance is easily caused.

SUMMARY OF THE INVENTION

[0005] Embodiments of the present disclosure provide a method and apparatus for moderating live streaming, a server, and a storage medium, which can avoid false rejection or false acceptance of live streaming moderation and improve an accuracy and a recall rate of the live streaming moderation.

[0006] A method for moderating live streaming is provided. The method includes:

[0007] performing preliminary violation moderation on a current live streaming frame in a to-be-moderated live streaming room based on a high-accuracy moderation model and a high-recall moderation model that are cascaded with each other; and

[0008] acquiring a target violation score of the current live streaming frame by inputting a first violation score of the current live streaming frame under the high-accuracy moderation model, a second violation score of the current live streaming frame under the high-recall moderation model, and a multi-dimensional behavior feature of the current live streaming frame in the to-be-moderated live streaming room to a pre-constructed behavior moderation model in response to the current live streaming frame passing the preliminary violation moderation; wherein performing the preliminary violation moderation on the current live streaming frame in the to-be-moderated live streaming room based on the high-accuracy moderation model and the high-recall moderation model that are cascaded with each other includes: acquiring the first violation score of the current live streaming frame by inputting the current live streaming frame in the to-be-moderated live streaming room to the pre-constructed high-accuracy moderation model; and acquiring the second violation score of the current live streaming frame by continuously inputting the current live streaming frame to the pre-constructed high-recall moderation model in response to the first violation score failing to exceed a predetermined accuracy violation threshold; wherein the current live streaming frame is determined to pass the preliminary violation moderation in response to the first violation score of the current live streaming frame failing to exceed the predetermined accuracy violation threshold and the second violation score of the current live streaming frame failing to exceed a predetermined recall violation threshold.

[0009] An apparatus for moderating live streaming is further provided. The apparatus includes:

[0010] a preliminary moderating module, configured to perform preliminary violation moderation on a current live streaming frame in a to-be-moderated live streaming room based

on a high-accuracy moderation model and a high-recall moderation model that are cascaded with each other; and

[0011] a violation score determining module, configured to acquire a target violation score of the current live streaming frame by inputting a first violation score of the current live streaming frame under the high-accuracy moderation model, a second violation score of the current live streaming frame under the high-recall moderation model, and a multi-dimensional behavior feature of the current live streaming frame in the to-be-moderated live streaming room to a pre-constructed behavior moderation model in response to the current live streaming frame passing the preliminary violation moderation; wherein the preliminary moderating module (410), further configured to acquire the first violation score of the current live streaming frame by inputting the current live streaming frame in the to-be-moderated live streaming room to the pre-constructed high-accuracy moderation model; and acquire the second violation score of the current live streaming frame by continuously inputting the current live streaming frame to the pre-constructed high-recall moderation model in response to the first violation score failing to exceed a predetermined accuracy violation threshold; wherein the current live streaming frame is determined to pass the preliminary violation moderation in response to the first violation score of the current live streaming frame failing to exceed the predetermined accuracy violation threshold and the second violation score of the current live streaming frame failing to exceed a predetermined recall violation threshold.

[0012] A server for moderating live streaming is further provided. The server includes:

[0013] one or more processors; and

[0014] a memory, configured to store one or more programs.

[0015] The one or more processors, when loading and running the one or more programs, are caused to perform the method for moderating live streaming according to any embodiment of the present disclosure.

[0016] A non-transitory computer-readable storage medium is further provided, storing a computer program. The computer program, when loaded and run by a processor, causes the processor to perform the method for moderating live streaming according to any embodiment of the present disclosure.

BRIEF DESCRIPTION OF THE DRAWINGS

[0017] FIG. 1A is a flow chart of a method for moderating living streaming according to first embodiments of the present disclosure;

[0018] FIG. 1B is a schematic diagram of a live streaming moderation process according to first embodiments of the present disclosure;

[0019] FIG. 2A is a flow chart of a method for moderating living streaming according to second embodiment of the present disclosure;

[0020] FIG. 2B is a schematic diagram of a live streaming moderation process according to second embodiments of the present disclosure;

[0021] FIG. 3A is a flow chart of a method for moderating living streaming according to third embodiments of the present disclosure;

[0022] FIG. 3B is a schematic diagram of a live streaming moderation process according to third embodiments of the present disclosure;

[0023] FIG. 4 is a schematic structural diagram of an apparatus for moderating living streaming according to fourth embodiments of the present disclosure; and

[0024] FIG. 5 is a schematic structural diagram of a server according to fifth embodiments of the present disclosure.

DETAILED DESCRIPTION OF THE INVENTION

[0025] The present disclosure will be described hereinafter with reference to the accompanying drawings and embodiments. The embodiments described herein are merely for the purpose of explaining the present disclosure but not limiting the present disclosure. For convenience of description, only the portions relevant to the present disclosure are shown in the drawings, not all portions. In addition, the embodiments and features of the embodiments in the present disclosure may be combined with each other without conflict.

[0026] First embodiments

[0027] FIG. 1A is a flowchart of a method for moderating living streaming according to first embodiments of the present disclosure. The embodiments are applicable to violation moderation of live streaming screens in any live streaming room. The method for moderating living streaming in the embodiments is performed by an apparatus for moderating living streaming in the embodiments of the present disclosure. The apparatus is implemented in a software and/or hardware manner, and is integrated in a server performing the method. The server is a backend server having a video moderation capability.

[0028] Referring to FIG. 1A, the method includes the following steps.

[0029] In S110, preliminary violation moderation is performed on a current live streaming frame in a to-be-moderated live streaming room based on a high-accuracy moderation model and a high-recall moderation model that are cascaded with each other.

[0030] As a variety of audiences typically participate in a live streaming room, in the case that live streaming content of some live streaming rooms involves or implies topics, such as terrorism, violence, pornography, or political sensitivity, the violation content is quickly spread over the Internet. Therefore, for prevention of adverse effects of the violation content on the life of a user, whether the violation live streaming content exists needs to be reviewed in real time in a live streaming process.

[0031] In the embodiments, the whole moderation process of the live streaming moderation is divided into preliminary violation moderation and secondary violation moderation to ensure the comprehensiveness of violation moderation of the live streaming. In this case, in the preliminary violation moderation process, obvious violation live streaming that is easily distinguished can be quickly found by moderation. In the secondary violation moderation process, a stricter violation moderation than the preliminary violation moderation is used to perform violation moderation again on a live streaming frame that whether to have violation content cannot be accurately determined in the preliminary violation moderation, such that the problems of false rejection and false acceptance in the live streaming moderation are avoided, and the comprehensive violation moderation of the current live streaming frame in the to-be-moderated live streaming room is ensured.

[0032] In the embodiments, two network models, that is, a high-accuracy moderation model and a high-recall moderation model are pre-constructed in the preliminary violation moderation process. The high-accuracy moderation model has high-accuracy violation moderation capability. In the case that the violation moderation is performed on the current live streaming frame in the to-be-moderated live streaming room by the high-accuracy moderation model, the determined violation live streaming frame is very likely to be violation, such that the problem of a false rejection of the violation moderation is avoided. The high-recall moderation model has high-recall violation moderation capability. In the case that the violation moderation is performed on the current live streaming frame in the to-be-moderated live streaming room by the high-recall moderation model, the violation content in the current live streaming frame can be found even if the current live streaming frame has little violation content, such that the phenomenon of false acceptance of violation moderation can be reduced.

[0033] In the embodiments, a training sample set is pre-constructed in training the high-accuracy moderation model and the high-recall moderation model. The training sample set includes history live streaming frames in history live streaming videos in various kinds of live streaming rooms. Based on whether each history live streaming frame is really violation and a difficulty of accurately moderating the violation, training samples in the training sample set are divided into four types: easy positive samples, easy negative samples, hard positive samples, and hard negative samples. The easy positive samples and the easy negative samples respectively refer to positive samples and negative samples that are easily recognized for violation by the models, and the hard positive samples and the hard negative samples respectively refer to positive samples and negative samples that are hardly recognized for violation by the models. In this way, as the high-accuracy moderation model has a relatively high requirement on the accuracy of violation auditing, training samples for the high-accuracy auditing moderation model include the easy positive samples and the easy negative samples in the training sample set, and model parameters, network structures, and the like of the high-accuracy moderation model are continuously optimized by reviewing whether violation content exists in the easy positive samples and the easy negative samples. As the high-recall moderation model has a relatively high requirement on the recall property of violation moderation, training samples for the high-recall moderation model include the hard positive samples and the hard negative samples in the training sample set, and moderation training is performed based on the training samples that are hardly moderated for violation, such that the high-recall moderation model can also moderate and find violation live streaming frames that are hardly found by moderation, and the high-recall property of the high-recall moderation model for the violation moderation is ensured.

[0034] Meanwhile, as the high-accuracy moderation model is trained only for the accuracy of violation moderation and causes severe false acceptance of violation moderation, and the high-recall moderation model has high-recall violation moderation capability and causes the false acceptance of violation moderation as far as possible, the high-accuracy moderation model is cascaded to the high-recall moderation model to ensure high accuracy and high recall property of the preliminary violation moderation in the embodiments. As shown in FIG. 1B, the high-accuracy moderation model is firstly used to perform the preliminary violation moderation on each current live streaming frame in the to-be-moderated live streaming room, and the high-recall moderation model continuously performs the violation moderation on current live streaming frames that are not found to have a violation by the high-accuracy moderation model,

such that the two-layer cascaded moderation models are used to perform the violation moderation to complete the preliminary violation moderation process in the embodiments, and the high accuracy and high recall property of the preliminary violation moderation are ensured.

[0035] In a video live streaming process of the to-be-moderated live streaming room, a live streaming picture is captured from the to-be-moderated live streaming room every predetermined duration (such as every 2 s) and is determined as the current live streaming frame in the embodiments, and the current live streaming frame is input to the high-accuracy moderation model in the two-layer cascaded network. The high-accuracy moderation model determines whether violation content exists in the current live streaming frame. In the case that the current live streaming frame passes the violation moderation of the high-accuracy moderation model, the high-accuracy moderation model determines that no violation content exists in the current live streaming frame. As the high-accuracy moderation model causes false acceptance of violation moderation, the current live streaming frame needs to be further input to the high-recall moderation model in the two-layer cascaded network, and the high-recall moderation model continues to determine whether violation content exists in the current live streaming frame, such that the preliminary violation moderation performed on the current live streaming frame in the to-be-moderated live streaming room is completed.

[0036] In this case, the high-recall moderation model has two violation moderation results for the current live streaming frame. 1) In the case that the current live streaming frame fails in passing the violation moderation of the high-recall moderation model, the high-recall moderation model can determine that violation content exists in the current live streaming frame and directly confirms that the to-be-moderated live streaming room of the current live streaming frame is a violation live streaming room. 2) In the case that the current live streaming frame passes the violation moderation of the high-recall moderation model, the high-recall moderation model determines that no violation content exists in the current live streaming frame, and a pre-constructed behavior moderation model needs to be subsequently used to continuously perform the violation moderation on the current live streaming frame.

[0037] In S120, a target violation score of the current live streaming frame is obtained by inputting a first violation score of the current live streaming frame under the high-accuracy moderation model, a second violation score of the current live streaming frame under the high-recall moderation model, and a multi-dimensional behavior feature of the current live streaming frame in the to-be-moderated live streaming room to a pre-constructed behavior moderation

model in response to the current live streaming frame passing the preliminary violation moderation.

[0038] In some embodiments, in the case that the current live streaming frame in the to-be-moderated live streaming room passes the violation moderation of the high-accuracy moderation model and the violation moderation of the high-recall moderation model, the current live streaming frame passes the preliminary violation moderation, and the secondary violation moderation needs to be performed to ensure the accuracy and comprehensiveness of the violation moderation.

[0039] In performing the preliminary violation moderation on the current live streaming frame by the high-accuracy moderation model and the high-recall moderation model that are cascaded with each other, violation analysis is usually performed on information of a live streaming visual picture of the current live streaming frame. In this case, the high-accuracy moderation model outputs the first violation score of the current live streaming frame, and the high-recall moderation model outputs the second violation score of the current live streaming frame. Both the first violation score and the second violation score can represent visual picture features of the current live streaming frame.

[0040] However, live streaming content in a live streaming room is various, and a live streamer interacts with an audience in the live streaming room for the live streaming content, for example, rewarding the live streaming, commenting on a public screen, and showing history violations of the live streamer, such that interaction behaviors in the live streaming room can also be used as a reference for determining whether violation content exists in the current live streaming frame. Therefore, in the embodiments, upon the current live streaming frame passing the preliminary violation moderation, in performing the secondary violation moderation on the current live streaming frame, a multi-dimensional behavior feature of the current live streaming frame in the to-be-moderated live streaming room is determined by analyzing multiple live streaming interaction behaviors of the live streamer and the audience in the to-be-moderated live streaming room at live streaming time of the current live streaming frame in the to-be-moderated live streaming room in addition to referring to the visual picture features represented by the first violation score of the current live streaming frame under the high-accuracy moderation model and the second violation score of the current live streaming frame under the high-recall moderation model. And as shown in FIG. 1B, the visual picture features represented by the first violation score and the second violation score of the current live streaming frame and the multi-dimensional behavior feature determined by the multiple live streaming

interaction behaviors in the to-be-moderated live streaming room are combined and are jointly input to the pre-constructed behavior moderation model. The behavior moderation model can perform fusion analysis on the visual picture features represented by the first violation score and the second violation score and the multi-dimensional behavior feature determined by the multiple live streaming interaction behaviors in the to-be-moderated live streaming room to uniformly determine whether violation content exists in the current live streaming frame, such that the target violation score of the current live streaming frame is output. In this case, the secondary violation moderation is performed on the target violation score based on the multi-angle feature fusion analysis on the visual picture features and the multi-dimensional behavior feature, without an additional moderation device, such that the violation moderation costs are reduced, and the accuracy and recall property of violation moderation of live streaming rooms are ensured.

[0041] According to the technical solutions provided in the embodiments, the preliminary violation moderation is performed on the current live streaming frame in the to-be-moderated live streaming room by the cascaded high-accuracy moderation model and high-recall moderation model, such that a violation live streaming frame which is easily distinguished and has an obvious violation is moderated and found quickly. For the current live streaming frame that passes the preliminary violation moderation and is hard to distinguish or has an unobvious violation, the uniform feature analysis combining the visual picture features and the behavior features is performed again on the current live streaming frame by the behavior moderation model based on the first violation score of the current live streaming frame under the high-accuracy moderation model and the second violation score of the current live streaming frame under the high-recall moderation model and in conjunction with the multi-dimensional behavior feature of the current live streaming frame in the to-be-moderated live streaming room, such that the target violation score of the current live streaming frame is acquired, and the comprehensive violation moderation of the current live streaming frame in the to-be-moderated live streaming room is achieved. The problems of false rejection and false acceptance of live streaming moderation are avoided, and the accuracy and recall rate of the live streaming moderation are improved on the basis of reduction in the violation moderation costs.

[0042] Second embodiments

[0043] FIG. 2A is a flow chart of a method for moderating living streaming according to second embodiments of the present disclosure, and FIG. 2B is a schematic diagram of a live streaming moderation process according to second embodiments of the present disclosure. The embodiments are illustrated on the basis of the above embodiments. As shown in FIG. 2A, a process of the secondary violation moderation of the current live streaming frame in the to-be-moderated live streaming room and a construction process of the behavior moderation model are explained in the embodiments.

[0044] In some embodiments, as shown in FIG. 2A, the embodiments include the following steps.

[0045] In S210, the preliminary violation moderation is performed on the current live streaming frame in the to-be-moderated live streaming room based on the high-accuracy moderation model and the high-recall moderation model that are cascaded with each other.

[0046] In S220, the live streaming behavior feature of the current live streaming frame is acquired by combining the first violation score of the current live streaming frame under the high-accuracy moderation model, the second violation score of the current live streaming frame under the high-recall moderation model, and the multi-dimensional behavior feature of the current live streaming frame in the to-be-moderated live streaming room in response to the current live streaming frame passing the preliminary violation moderation,.

[0047] In some embodiments, in the case that the current live streaming frame passes the preliminary violation moderation, both the high-accuracy moderation model and the high-recall moderation model determine that no violation content exists in the current live streaming frame. Therefore, the secondary violation moderation is performed on the current live streaming frame based on the visual picture features of the current live streaming frame and the behavior features of the current live streaming frame to ensure the accuracy of live streaming moderation and avoid false acceptance of violation moderation. In this case, the first violation score of the current live streaming frame under the high-accuracy moderation model and the second violation score of the current live streaming frame under the high-recall moderation model are determined. The first violation score and the second violation score are determined as the visual picture features corresponding to the current live streaming frame. Meanwhile, the multi-dimensional behavior feature of the current live streaming frame in the to-be-moderated live streaming room is determined based on the multiple live streaming interaction behaviors in the to-be-moderated live streaming room at the live streaming time of the current live streaming frame. The live streaming behavior feature of the current live

streaming frame is acquired by combining the first violation score, the second violation score, and the multi-dimensional behavior feature and is determined as a multi-angle fusion feature under the secondary violation moderation. The live streaming behavior feature is used subsequently to determine whether violation content exists in the current live streaming frame, such that the comprehensiveness and accuracy of the live streaming moderation can be ensured.

[0048] For example, the live streaming behavior feature in the embodiments is X = (x1, x2, …, x11). x1 represents the first violation score of the current live streaming frame under the high-accuracy moderation model, x2 represents the second violation score of the current live streaming frame under the high-recall moderation model, x3 represents a real-time number of audiences in the to-be-moderated live streaming room in the multi-dimensional behavior feature, x4 represents a number of comments posted by users to the public screen in the to-be-moderated live streaming room in the multi-dimensional behavior feature, x5 represents a number of violation comments on the public screen in the to-be-moderated live streaming room in the multi-dimensional behavior feature, x6 represents a percentage of the violation comments on the public screen to all the comments on the public screen in the to-be-moderated live streaming room in the multi-dimensional behavior feature, x7 represents a number of live streaming rewards in the to-be-moderated live streaming room in the multi-dimensional behavior feature, x8 represents a number of history violations of the live streamer in the to-be-moderated live streaming room in the multi-dimensional behavior feature, x9 represents a percentage of the history violations of the live streamer in the to-be-moderated live streaming room in the multi-dimensional behavior feature, x10 represents a number of history voice violations of the live streamer in the to-be-moderated live streaming room in the multi-dimensional behavior feature, and x11 represents a percentage of the history voice violations of the live streamer in the to-be-moderated live streaming room in the multi-dimensional behavior feature.

[0049] The multi-dimensional behavior feature in the embodiments is dynamically configured to ensure the accuracy of the live streaming violation moderation, which is not limited.

[0050] In S230, the target violation score of the current live streaming frame is acquired by inputting the live streaming behavior feature to the behavior moderation model.

[0051] In some embodiments, upon acquisition of the live streaming behavior feature, the live streaming behavior feature is determined as a multi-angle violation fusion feature and input to the pre-constructed behavior moderation model. Fusion analysis is performed on a feature value of the live streaming behavior feature under each dimension by the behavior moderation

model to determine whether violation content information exists, the target violation score of the current live streaming frame is output, and then whether violation content exists in the current live streaming frame is determined based on the target violation score, such that the accuracy of the live streaming moderation is improved.

[0052] For example, in the embodiments the behavior moderation model is constructed by: extracting a corresponding history live streaming frame from each history live streaming video in a history live streaming video set, and determining a multi-dimensional history behavior feature of each history live streaming frame in a live streaming room where the each history live streaming frame is located and a violation label of the history live streaming frame; acquiring a history violation score of the history live streaming frame by inputting a first history violation score of the history live streaming frame under the high-accuracy moderation model, a second history violation score of the history live streaming frame under the high-recall moderation model, and a multi-dimensional history behavior feature of the history live streaming frame in the live streaming room to the pre-constructed behavior moderation model for each history live streaming frame; and continuously optimizing the behavior moderation model to minimize a difference between the history violation score and the violation label of each history live streaming frame.

[0053] The history live streaming video set for model construction is constructed by searching for history live streaming videos of a plurality of live streaming rooms. As the violation moderation is mainly performed on live streaming frames in the live streaming process in the violation moderation on the live streaming rooms, video frame extraction needs to be performed on each history live streaming video in the history live streaming video set. For example, a history live streaming picture is captured from each history live streaming video every predetermined duration and is determined as a history live streaming frame extracted from the history live streaming video. In this case, all the extracted history live streaming frames form model construction samples of the behavior moderation model.

[0054] In this case, for successful construction of the behavior moderation model, it is necessary to set a plurality of model construction features for reference in the model construction process and a model construction optimization goal of the behavior moderation model, such that actual model construction conditions under the plurality of model construction features are optimized based on the model construction optimization goal. For example, the model construction features in the embodiments are $X = (x1, x2, \ldots, x11)$. x1 represents the first history violation score of the history live streaming frame under the high-accuracy

moderation model, x2 represents the second history violation score of the history live streaming frame under the high-recall moderation model, x3 represents a real-time number of audiences in the live streaming room of the history live streaming frame in the set multi-dimensional history behavior feature, x4 represents a number of comments posted by users to the public screen in the live streaming room of the history live streaming frame in the multi-dimensional history behavior feature, x5 represents a number of violation comments on the public screen in the live streaming room of the history live streaming frame in the multi-dimensional history behavior feature, x6 represents a percentage of the violation comments on the public screen to all the comments on the public screen in the live streaming room of the history live streaming frame in the multi-dimensional history behavior feature, x7 represents a number of live streaming rewards in the live streaming room of the history live streaming frame in the multi-dimensional history behavior feature, x8 represents a number of history violations of the live streamer in the live streaming room of the history live streaming frame in the multi-dimensional history behavior feature, x9 represents a percentage of the history violations of the live streamer in the live streaming room of the history live streaming frame in the multi-dimensional history behavior feature, x10 represents a number of history voice violations of the live streamer in the live streaming room of the history live streaming frame in the multi-dimensional history behavior feature, and x11 represents a percentage of the history voice violations of the live streamer in the live streaming room of the history live streaming frame in the multi-dimensional history behavior feature.

[0055] As the multi-dimensional history behavior features, in the plurality of model construction features, of the history live streaming frames in the live streaming rooms are determined in the history live streaming process, upon extraction of the corresponding history live streaming frame from each history live streaming video, the multi-dimensional history behavior feature of each history live streaming frame in the live streaming room is determined by determining history live streaming interaction behaviors of the history live streaming frame in the live streaming room. Meanwhile, the violation label of each history live streaming frame is determined, as shown in FIG. 2B, such that continuous model construction optimization is subsequently performed based on the model construction optimization goal. For each history live streaming frame, the history live streaming frame is input to the high-accuracy moderation model and the high-recall moderation model in the embodiments, such that the first history violation score and the second history violation score of each history live streaming frame are acquired. On this basis, the first history violation score of each history live streaming frame,

the second history violation score of each history live streaming frame, and the multi-dimensional history behavior feature of each history live streaming frame in the live streaming room of the history live streaming frame are combined to form the model construction features corresponding to each history live streaming frame, and the history violation scores of each history live streaming frame are acquired by continuously inputting the model construction features corresponding to each history live streaming frame to the behavior moderation model.

[0056] In this case, the model construction optimization goal of the behavior moderation model in the embodiments is to minimize the difference between the history violation score and the violation label of each history live streaming frame. Therefore, the difference between the history violation score of each history live streaming frame and the violation label of each history live streaming frame is continuously analyzed in the model construction optimization process, and the difference is continuously reduced by optimizing mathematical parameters in the behavior moderation model, such that the goal of minimizing the difference between the history violation score and the violation label of the history live streaming frame, and a finally optimized behavior moderation model is acquired. In this case, the finally optimized behavior moderation model has high-accuracy violation moderation capability.

[0057] For example, the model construction optimization goal of the behavior moderation model in the embodiments is:

$$\min \sum_{x^i \in V} \left(\overset{\circ}{y}^i - y^i \right)^2 \; s.t. \; \overset{\circ}{y}^i = p\left(x^i \right)$$

. x^i represents a model construction feature of an ith history live streaming frame, y^i represents a violation label of the ith history live streaming frame, $y^i = 1$ indicates that violation content exists, $y^i = 0$ indicates that no violation content exists, $\overset{\circ}{y}^i$ represents a history violation score determined by the behavior moderation model for the ith history live streaming frame, $V = \left(x^1, x^2, \mathrm{L}\ , x^m \right)$ indicates a model construction sample including history live streaming frames, m represents a number of samples, and $p\left(x^i \right)$ represents a prediction probability of x^i acquired by the behavior moderation model.

[0058] According to the technical solutions in the embodiments, the behavior moderation model is constructed based on the model construction optimization goal of minimizing the difference between the history violation score and the violation label of each history live streaming frame, such that the accuracy of violation moderation based on the behavior

moderation model is ensured. The live streaming behavior feature of the current live streaming frame is subsequently acquired by combining the first violation score of the current live streaming frame in the to-be-moderated live streaming room under the high-accuracy moderation model, the second violation score of the current live streaming frame in the to-be-moderated live streaming room under the high-recall moderation model, and the multi-dimensional behavior feature of the current live streaming frame in the to-be-moderated live streaming room, and the multi-angle violation feature analysis is performed on the live streaming behavior feature based on the constructed behavior moderation model, such that the accuracy and recall rate of live streaming moderation are improved.

[0059] Third embodiments

[0060] FIG. 3A is a flow chart of a method for moderating living streaming according to third embodiments of the present disclosure; and FIG. 3B is a schematic diagram of a live streaming moderation process according to third embodiments of the present disclosure. The embodiments are illustrated on the basis of the above embodiments. As shown in FIG. 3A, the overall violation moderation process of the to-be-moderated live streaming room in the embodiments is explained.

[0061] In some embodiments, as shown in FIG. 3A, the embodiments include the following steps.

[0062] In S310, the first violation score of the current live streaming frame is acquired by inputting the current live streaming frame in the to-be-moderated live streaming room to the pre-constructed high-accuracy moderation model.

[0063] In some embodiments, in performing the violation moderation on the to-be-moderated live streaming room, whether violation content exists in a live streaming picture is reviewed in real time in the live streaming process, such that the violation moderation needs to be performed on the current live streaming frame in the live streaming process in real time. The current live streaming frame is input to the pre-constructed high-accuracy moderation model, the preliminary violation moderation is performed on the live streaming picture in the current live streaming frame by the high-accuracy moderation model, such that the first violation score of the current live streaming frame is output. Then, whether violation content exists in the current live streaming frame is determined by subsequently determining whether the first violation score exceeds a predetermined accuracy violation threshold.

[0064] In S320, whether the first violation score exceeds a predetermined accuracy violation threshold is determined; if yes, S380 is performed; and if no, S330 is performed.

[0065] S330, the second violation score of the current live streaming frame is acquired by continuously inputting the current live streaming frame to the pre-constructed high-recall moderation model.

[0066] In some embodiments, in the case that the first violation score fails to exceed the predetermined accuracy violation threshold, the high-accuracy moderation model determines that no violation content exists in the current live streaming frame. However, as the high-accuracy moderation model has false acceptance of violation moderation, the current live streaming frame needs to be continuously input to the pre-constructed high-recall moderation model. As shown in FIG. 3B, corresponding violation moderation is continuously performed on the live streaming picture in the current live streaming frame by the high-recall moderation model, such that the second violation score of the current live streaming frame is output. Then, whether violation content exists in the current live streaming frame is determined again by subsequently determining whether the second violation score exceeds a predetermined recall violation threshold.

[0067] In S340, whether the second violation score exceeds a predetermined recall violation threshold is determined; if yes, S380 is performed; and if no, S350 is performed.

[0068] In the case that the second violation score fails to exceed the predetermined recall violation threshold, the high-recall moderation model also determines that no violation content exists in the current live streaming frame. That is, the current live streaming frame passes the violation moderation of the high-accuracy moderation model and the violation moderation of the high-recall moderation model, namely, the current live streaming frame passes the preliminary violation moderation. The secondary violation moderation is continuously performed on the current live streaming frame by the behavior moderation model. The current live streaming frame is determined to pass the preliminary violation moderation in the case that the first violation score of the current live streaming frame fails to exceed the predetermined accuracy violation threshold and the second violation score of the current live streaming frame fails to exceed the predetermined recall violation threshold.

[0069] In S350, a target violation score of the current live streaming frame is acquired by inputting the first violation score of the current live streaming frame under the high-accuracy moderation model, the second violation score of the current live streaming frame under the high-recall moderation model, and the multi-dimensional behavior feature of the current live

streaming frame in the to-be-moderated live streaming room to the pre-constructed behavior moderation model.

[0070] In S360, whether the target violation score exceeds a predetermined behavior violation threshold is determined; if yes, S380 is performed; and if no, S370 is performed.

[0071] In some embodiments, upon outputting of the target violation score of the current live streaming frame by the behavior moderation model, whether violation content exists in the current live streaming frame is finally reviewed by determining whether the target violation score exceeds the predetermined behavior violation threshold, such that different live streaming processing processes are performed.

[0072] In S370, a new current live streaming frame is collected from the to-be-moderated live streaming room upon an interval of a predetermined live streaming duration, and violation moderation is continuously performed on the new current live streaming frame.

[0073] In some embodiments, in the case that the target violation score fails to exceed the predetermined behavior violation threshold, the behavior moderation model determines that no violation content exists in the current live streaming frame. In this case, the three models in the embodiments all determine that no violation content exists in the current live streaming frame, and it is determined that the current live streaming frame is a non-violation frame, which indicates that no violation content exists in the to-be-moderated live streaming room currently. In this case, for the instantaneity of violation moderation of the live streaming room, upon the interval (for example, every 2s) of the predetermined live streaming duration, a new live streaming picture is continuously captured from the to-be-moderated live streaming room at current live streaming time and is determined as a new collected current live streaming frame. Violation moderation is continuously performed on the new current live streaming frame according to S310 to S380 in the embodiments, such that real-time violation moderation is performed on the to-be-moderated live streaming room until the live streaming is ended or a violation live streaming frame is found.

[0074] In S380, live streaming information of the to-be-moderated live streaming room is pushed to a manual moderation platform.

[0075] In some embodiments, in the case that the first violation score exceeds the predetermined accuracy violation threshold, or the second violation score exceeds the predetermined recall violation threshold, or the target violation score exceeds the predetermined behavior violation threshold, the model used for current moderation determines that violation content exists in the current live streaming frame. As shown in FIG. 3B, an

additional manual moderation platform is further provided to avoid the false rejection in the violation moderation in the embodiments. In the case that violation content in the current live streaming frame is found by moderation, the live streaming information of the to-be-moderated live streaming room of the current live streaming frame is pushed to the manual moderation platform to further perform manual moderation on live streaming content of the to-be-moderated live streaming room, such that the accuracy of live streaming moderation is ensured.

[0076] In the embodiments, the live streaming information of the to-be-moderated live streaming room is a live streaming address of the to-be-moderated live streaming room, a live video picture, or the like, which is not limited herein.

[0077] In S390, live streaming interaction items in the to-be-moderated live streaming room are monitored in real time, live streaming information of the to-be-moderated live streaming room is pushed to the manual moderation platform in response to the live streaming interaction items exceeding a predetermined interaction threshold, and the violation moderation performed on the current live streaming frame in the to-be-moderated live streaming room is stopped.

[0078] In some embodiments, whether to directly push the live streaming information to the manual moderation platform for manual moderation is determined by monitoring the live streaming interaction items in the to-be-moderated live streaming room in real time on the basis of achieving live streaming violation moderation by constructing models to avoid an abnormality in the violation moderation and ensure a high recall rate of the live streaming moderation, and the manual moderation is determined as a moderation abnormality ensuring solution of the live streaming moderation. In this case, the live streaming interaction items include a number of live streaming views, a number of live streaming rewards, and the like in the to-be-moderated live streaming room, such that an acceptance degree of the to-be-moderated live streaming room is analyzed. In the case that the live streaming interaction items exceed the predetermined interaction threshold, the acceptance degree of the to-be-moderated live streaming room is extremely great. Violation content is quickly spread out once it appears. Therefore, the live streaming information of the to-be-moderated live streaming room needs to be directly pushed to the manual moderation platform for manual moderation to ensure high efficiency of the live streaming violation moderation. Meanwhile, the violation moderation performed on the current live streaming frame in the to-be-moderated live streaming room by the multiple constructed models is stopped.

[0079] According to the technical solutions in the embodiments, upon the violation content in the current live streaming frame being determined, the live streaming information of the to-be-

moderated live streaming room is directly pushed to the manual moderation platform for manual moderation, such that the false rejection of the violation moderation is avoided, and the accuracy of the live streaming violation moderation is improved. Meanwhile, an abnormality ensuring solution for the live streaming moderation is additionally set based on the live streaming interaction items in the to-be-moderated live streaming room, and the comprehensiveness of the live streaming violation moderation is ensured.

[0080] Fourth embodiments

[0081] FIG. 4 is a schematic structural diagram of an apparatus for moderating live streaming according to fourth embodiments of the present disclosure. As shown in FIG. 4, the apparatus for moderating live streaming includes:

[0082] a preliminary moderating module 410, configured to perform preliminary violation moderation on a current live streaming frame in a to-be-moderated live streaming room based on a high-accuracy moderation model and a high-recall moderation model that are cascaded with each other; and

[0083] a violation score determining module 420, configured to acquire a target violation score of the current live streaming frame by inputting a first violation score of the current live streaming frame under the high-accuracy moderation model, a second violation score of the current live streaming frame under the high-recall moderation model, and a multi-dimensional behavior feature of the current live streaming frame in the to-be-moderated live streaming room to a pre-constructed behavior moderation model in response to the current live streaming frame passing the preliminary violation moderation.

[0084] According to the technical solutions in the embodiments, the preliminary violation moderation is performed on the current live streaming frame in the to-be-moderated live streaming room by the cascaded high-accuracy moderation model and high-recall moderation model, such that a violation live streaming frame which is easily distinguished and has an obvious violation is moderated and found quickly. For the current live streaming frame that passes the preliminary violation moderation and is hard to distinguish or has an unobvious violation, the uniform feature analysis combining the visual picture features and the behavior features is performed again on the current live streaming frame by the behavior moderation model based on the first violation score of the current live streaming frame under the high-accuracy moderation model and the second violation score of the current live streaming frame under the high-recall moderation model and in conjunction with the multi-dimensional

behavior feature of the current live streaming frame in the to-be-moderated live streaming room, such that the target violation score of the current live streaming frame is acquired, and the comprehensive violation moderation of the current live streaming frame in the to-be-moderated live streaming room is achieved. The problems of false rejection and false acceptance of live streaming moderation are avoided, and the accuracy and recall rate of the live streaming moderation are improved on the basis of reduction in the violation moderation costs.

[0085] The apparatus for moderating live streaming in the embodiments is applicable to the method for moderating living streaming in any embodiment above, and has corresponding functions and effects.

[0086] Fifth embodiments

[0087] FIG. 5 is a schematic structural diagram of a server according to fifth embodiments of the present disclosure. As shown in FIG. 5, the server includes a processor 50, a memory 51, and a communication apparatus 52. The number of processors 50 in the server is one or more, and by way of example, one processor 50 is shown in FIG. 5. The processor 50, the memory 51, and the communication apparatus 52 in the server are connected by a bus or other means, and by way of example, they are connected by a bus in FIG. 5.

[0088] The server in the embodiments is applicable to the method for moderating living streaming in any embodiment above, and has corresponding functions and effects.

[0089] Sixth embodiments

[0090] The sixth embodiment of the present disclosure further provides a non-transitory computer-readable storage medium storing a computer program thereon. The computer program, when loaded and run by a processor, causes the processor to perform the method according to any one of the above embodiments. The method includes:

[0091] performing preliminary violation moderation on a current live streaming frame in a to-be-moderated live streaming room based on a high-accuracy moderation model and a high-recall moderation model that are cascaded with each other; and

[0092] acquiring a target violation score of the current live streaming frame by inputting a first violation score of the current live streaming frame under the high-accuracy moderation model, a second violation score of the current live streaming frame under the high-recall moderation model, and a multi-dimensional behavior feature of the current live streaming frame in the to-

be-moderated live streaming room to a pre-constructed behavior moderation model in response to the current live streaming frame passing the preliminary violation moderation.

[0093] The embodiments of the present disclosure provide a storage medium containing computer-executable instructions. The computer-executable instructions are not limited to performing the above method operations, but also can perform related operations in the method for moderating living streaming according to any embodiment of the present disclosure. The storage medium may be a non-transitory storage medium.

[0094] The present disclosure may be implemented by software, general-purpose hardware, or hardware. The technical solutions of the present disclosure are embodied in the form of a software product. The computer software product is stored in a computer-readable storage medium, such as, a floppy disk, a read-only memory (ROM), a random access memory (RAM), a flash (FLASH), a hard disk, or an optical disk of a computer, and includes a plurality of instructions for causing a computer device (which may be a personal computer, a server, a network device, or the like) to perform the method for moderating living streaming according to the embodiments of the present disclosure.

[0095] In the above embodiments of the apparatus for moderating live streaming, the plurality of units and modules are only divided according to a function logic, but are not limited to the above division, as long as corresponding functions are achieved. In addition, the names of the plurality of function units are merely for convenience of distinguishing from each other, and are not intended to limit the protection scope of the present disclosure.

CLAIMS

What is claimed is:

1. A method for moderating live streaming, comprising:

performing preliminary violation moderation on a current live streaming frame in a to-be-moderated live streaming room based on a high-accuracy moderation model and a high-recall moderation model that are cascaded with each other; and

acquiring a target violation score of the current live streaming frame by inputting a first violation score of the current live streaming frame under the high-accuracy moderation model, a second violation score of the current live streaming frame under the high-recall moderation model, and a multi-dimensional behavior feature of the current live streaming frame in the to-be-moderated live streaming room to a pre-constructed behavior moderation model in response to the current live streaming frame passing the preliminary violation moderation;

wherein performing the preliminary violation moderation on the current live streaming frame in the to-be-moderated live streaming room based on the high-accuracy moderation model and the high-recall moderation model that are cascaded with each other comprises:

acquiring the first violation score of the current live streaming frame by inputting the current live streaming frame in the to-be-moderated live streaming room to the pre-constructed high-accuracy moderation model; and

acquiring the second violation score of the current live streaming frame by continuously inputting the current live streaming frame to the pre-constructed high-recall moderation model in response to the first violation score failing to exceed a predetermined accuracy violation threshold;

wherein the current live streaming frame is determined to pass the preliminary violation moderation in response to the first violation score of the current live streaming frame failing to exceed the predetermined accuracy violation threshold and the second violation score of the current live streaming frame failing to exceed a predetermined recall violation threshold.

2. The method according to claim 1, wherein acquiring the target violation score of the current live streaming frame by inputting the first violation score of the current live streaming frame under the high-accuracy moderation model, the second violation score of the current live streaming frame under the high-recall moderation model, and the multi-dimensional behavior feature of the current live streaming frame in the to-be-moderated live streaming room to the pre-constructed behavior moderation model comprises:

acquiring a live streaming behavior feature of the current live streaming frame by combining the first violation score of the current live streaming frame under the high-accuracy moderation model, the second violation score of the current live streaming frame under the high-recall moderation model, and the multi-dimensional behavior feature of the current live streaming frame in the to-be-moderated live streaming room; and

acquiring the target violation score of the current live streaming frame by inputting the live streaming behavior feature to the behavior moderation model.

3. The method according to claim 1, wherein the behavior moderation model is constructed by:

extracting a corresponding history live streaming frame from each history live streaming video in a history live streaming video set, and determining a multi-dimensional history behavior feature of each history live streaming frame in a live streaming room where the each history live streaming frame is located and a violation label of the history live streaming frame;

acquiring a history violation score of the history live streaming frame by inputting a first history violation score of the history live streaming frame under the high-accuracy moderation model, a second history violation score of the history live streaming frame under the high-recall moderation model, and a multi-dimensional history behavior feature of the history live streaming frame in the live streaming room to the pre-constructed behavior moderation model for each history live streaming frame; and

continuously optimizing the behavior moderation model to minimize a difference between the history violation score and the violation label of each history live streaming frame.

4. The method according to claim 1, wherein upon acquiring the target violation score of the current live streaming frame by inputting the first violation score of the current live streaming frame under the high-accuracy moderation model, the second violation score of the current live streaming frame under the high-recall moderation model, and the multi-dimensional behavior feature of the current live streaming frame in the to-be-moderated live streaming room to the pre-constructed behavior moderation model, the method further comprises:

pushing live streaming information of the to-be-moderated live streaming room to a manual moderation platform in response to the to-be-moderated violation score exceeding a predetermined behavior violation threshold; and

collecting a new current live streaming frame from the to-be-moderated live streaming room upon an interval of a predetermined live streaming duration in response to the target violation score failing to exceed the predetermined behavior violation threshold, and continuously performing violation moderation on the new current live streaming frame.

5. The method according to claim 4, further comprising:

monitoring live streaming interaction items in the to-be-moderated live streaming room in real time, pushing live streaming information of the to-be-moderated live streaming room to the manual moderation platform in response to the live streaming interaction items exceeding a predetermined interaction threshold, and stopping the violation moderation performed on the current live streaming frame in the to-be-moderated live streaming room.

~~6. The method according to claim 1, wherein performing the preliminary violation moderation on the current live streaming frame in the to-be-moderated live streaming room based on the high-accuracy moderation model and the high-recall moderation model that are cascaded with each other comprises:~~

~~acquiring the first violation score of the current live streaming frame by inputting the current live streaming frame in the to-be-moderated live streaming room to the pre-constructed high-accuracy moderation model; and~~

~~acquiring the second violation score of the current live streaming frame by continuously inputting the current live streaming frame to the pre-constructed high-recall moderation model in response to the first violation score failing to exceed a predetermined accuracy violation threshold.~~

~~7. The method according to claim 6, wherein the current live streaming frame is determined to pass the preliminary violation moderation in response to the first violation score of the current live streaming frame failing to exceed the predetermined accuracy violation threshold and the second violation score of the current live streaming frame failing to exceed a predetermined recall violation threshold.~~

~~8~~6. The method according to claim ~~6~~1, further comprising:

pushing the live streaming information of the to-be-moderated live streaming room to the manual moderation platform in a case that the first violation score exceeds the predetermined

accuracy violation threshold or the second violation score exceeds the predetermined recall violation threshold.

~~9~~7. The method according to any one of claims 1 to ~~8~~6, wherein training samples of the high-accuracy moderation model comprise easy positive samples and easy negative samples in a training sample set, and training samples of the high-recall moderation model comprise hard positive samples and hard negative samples in the training sample set.

~~10~~8. An apparatus for moderating live streaming, comprising:

a preliminary moderating module, configured to perform preliminary violation moderation on a current live streaming frame in a to-be-moderated live streaming room based on a high-accuracy moderation model and a high-recall moderation model that are cascaded with each other; and

a violation score determining module, configured to acquire a target violation score of the current live streaming frame by inputting a first violation score of the current live streaming frame under the high-accuracy moderation model, a second violation score of the current live streaming frame under the high-recall moderation model, and a multi-dimensional behavior feature of the current live streaming frame in the to-be-moderated live streaming room to a pre-constructed behavior moderation model in response to the current live streaming frame passing the preliminary violation moderation;

wherein the preliminary moderating module, further configured to acquire the first violation score of the current live streaming frame by inputting the current live streaming frame in the to-be-moderated live streaming room to the pre-constructed high-accuracy moderation model; and acquire the second violation score of the current live streaming frame by continuously inputting the current live streaming frame to the pre-constructed high-recall moderation model in response to the first violation score failing to exceed a predetermined accuracy violation threshold;

wherein the current live streaming frame is determined to pass the preliminary violation moderation in response to the first violation score of the current live streaming frame failing to exceed the predetermined accuracy violation threshold and the second violation score of the current live streaming frame failing to exceed a predetermined recall violation threshold.

~~11~~9. A server for moderating live streaming, comprising:

one or more processors; and

a memory, configured to store one or more programs;

wherein the one or more processors, when loading and running the one or more programs, are caused to perform the method for moderating live streaming as defined in any one of claims 1 to ~~9~~7.

~~12~~10. A non-transitory computer-readable storage medium, storing a computer program, wherein the computer program, when loaded and run by a processor, causes the processor to perform the method for moderating live streaming as defined in any one of claims 1 to ~~9~~7.