

VIDEO PROCESSING METHOD AND APPARATUS, ELECTRONIC DEVICE, AND STORAGE MEDIUM

CROSS-REFERENCE TO RELATED APPLICATION

[0001] The present disclosure claims priority to Chinese Patent Application No. 202010442733.6, filed to the China National Intellectual Property Administration on May 22, 2020 and entitled "Video Processing Method and Apparatus, Electronic Device, and Storage Medium", the disclosure of which is hereby incorporated by reference in its entirety.

TECHNICAL FIELD

[0002] The disclosure relates to the field of computer vision, and particularly to a video processing method and apparatus, an electronic device, and a storage medium.

BACKGROUND

[0003] In the teaching process, teachers need to focus on teaching, so it is difficult for institutions or the teachers to master listening states of students, and parents may not know the performance of their children at school. Whether a student really attends a class and listens to a class carefully, and how is the participation of the student in class, these are not be quantitatively evaluated.

[0004] Therefore, how to master a learning state of each student in the teaching process at the same time of ensuring the teaching quality becomes a problem urgent to be solved at present.

SUMMARY

[0005] The disclosure discloses solutions to video processing.

[0006] According to an aspect of the disclosure, a video processing method is provided, which may include the following operations.

[0007] A video is acquired, at least part of video frames in the video including a target object. At least one type of learning behavior of the target object in a process of watching a teaching course is detected according to the video. Under the condition of detecting that the target object performs the at least one type of the learning behavior, learning state information is generated according to a duration for the target object to perform the at least one type of the learning behavior and/or the at least part of video frames including the at least one type of the learning behavior.

[0008] According to an aspect of the disclosure, a video processing apparatus is provided, which may include a video acquisition module, a detection module, and a generation module.

[0009] The video acquisition module may be configured to acquire a video, at least part of video frames in the video including a target object.

[0010] The detection module may be configured to detect at least one type of learning behavior of the target object in a process of watching a teaching course according to the video.

[0011] The generation module may be configured to, under the condition of detecting that the target object performs the at least one type of the learning behavior, generate learning state information according to a duration for the target object to perform the at least one type of the learning behavior and/or the at least part of video frames including the at least one type of the learning behavior.

[0012] According to an aspect of the disclosure, an electronic device is provided, which may include:

[0013] a processor; and a memory configured to store an instruction executable for the processor. The processor may be configured to execute the video processing method.

[0014] According to an aspect of the disclosure, a computer-readable storage medium

is provided, in which a computer program instruction may be stored, the computer program instruction being executed by a processor to implement the video processing method.

[0015] According to an aspect of the disclosure, a computer program is provided, which may include a computer-readable code, when the computer-readable code runs in an electronic device, a processor in the electronic device is configured to execute the video processing method.

[0016] In the embodiments of the disclosure, under the condition of detecting that the target object exists the at least one type of the learning behavior, visual learning state information may be generated by using the video frames including the learning behavior, and quantitative learning state information may be generated according to the duration of the learning behavior. In such a manner, learning state information with an evaluation value may be obtained flexibly, and a related person and institution such as a teacher or a parent may master a learning state of a student effectively and accurately.

[0017] It is to be understood that the above general description and the following detailed description are only exemplary and explanatory and not intended to limit the disclosure. According to the following detailed descriptions made to exemplary embodiments with reference to the drawings, other features and aspects of the disclosure may become clear.

BRIEF DESCRIPTION OF THE DRAWINGS

[0018] The accompanying drawings, which are incorporated in and constitute a part of this specification, illustrate embodiments consistent with the disclosure and, together with the specification, serve to describe the technical solutions of the disclosure.

[0019] FIG. 1 is a flowchart of a video processing method according to an embodiment of the disclosure.

[0020] FIG. 2 is a block diagram of a video processing apparatus according to an embodiment of the disclosure.

[0021] FIG. 3 is a schematic diagram of an application example according to the

disclosure.

[0022] FIG. 4 is a block diagram of an electronic device according to an embodiment of the disclosure.

[0023] FIG. 5 is a block diagram of an electronic device according to an embodiment of the disclosure.

DETAILED DESCRIPTION

[0024] Each exemplary embodiment, feature and aspect of the disclosure will be described below with reference to the drawings in detail. The same reference signs in the drawings represent components with the same or similar functions. Although each aspect of the embodiments is shown in the drawings, the drawings are not required to be drawn to scale, unless otherwise specified.

[0025] Herein, special term "exemplary" refers to "use as an example, embodiment or description". Herein, any "exemplarily" described embodiment may not be explained to be superior to or better than other embodiments.

[0026] In the disclosure, term "and/or" is only an association relationship describing associated objects and represents that three relationships may exist. For example, A and/or B may represent three conditions: i.e., independent existence of A, existence of both A and B and independent existence of B. In addition, term "at least one of..." in the disclosure represents any one of multiple or any combination of at least two of multiple. For example, including at least one of A, B or C may represent including any one or more elements selected from a set formed by A, B and C.

[0027] In addition, for describing the disclosure better, many specific details are presented in the following specific implementations. It is understood by those skilled in the art that the disclosure may still be implemented even without some specific details. In some examples, methods, means, components and circuits known very well to those skilled in the art are not described in detail, to highlight the subject of the disclosure.

[0028] FIG. 1 is a flowchart of a video processing method according to an embodiment of the disclosure. The method may be applied to a video processing apparatus. The video processing apparatus may be a terminal device, a server, another processing device, etc. The terminal device may be User Equipment (UE), a mobile device, a user terminal, a terminal, a cell phone, a cordless phone, a Personal Digital Assistant (PDA), a handheld device, a computing device, a vehicle device, a wearable device, etc. In an example, the data processing method may be applied to a cloud server or a local server. The cloud server may be a public cloud server or a private cloud server, and is selected flexibly according to a practical condition.

[0029] In some possible implementations, the video processing method may also be implemented by a processor by calling a computer-readable instruction stored in a memory.

[0030] As shown in FIG. 1, in a possible implementation, the video processing method may include the following operations.

[0031] In S11, a video is acquired, at least part of video frames in the video including a target object.

[0032] In S12, at least one type of learning behavior of the target object in a process of watching a teaching course is detected according to the video.

[0033] In S13, under the condition of detecting that the target object performs the at least one type of the learning behavior, learning state information is generated according to the at least part of video frames including the at least one type of the learning behavior and/or a duration for the target object to perform the at least one type of the learning behavior.

[0034] The target object may be any object whose learning state information is acquired, i.e., an object with a learning state evaluation requirement. A specific implementation form thereof may be determined flexibly according to the practical condition. In a possible implementation, the target object may be a student, such as an elementary school student, a middle school student, or a college student. In a possible implementation, the target object may be an adult who attends a refresher course, such as an adult who attends vocational education and training, or an old person who studies in a college for the aged.

[0035] In the embodiment of the disclosure, the video may be a video, recorded in the process of watching the teaching course, of the target object. An implementation form of the teaching course is not limited, and may be a prerecorded course video, or may be a live course, or a course taught on site by a teacher, etc. The at least part of video frames in the video may include the target object, namely an appearance condition of the target object in the recorded video may be determined flexibly according to the practical condition. In a possible implementation, the target object may appear throughout the video. In a possible implementation, the target object may not appear in video frames at some moments or in some time periods.

[0036] A scenario that the target object watches the teaching course may be determined flexibly according to the practical condition. In a possible implementation, the scenario may be an online scenario, namely the target object watches the teaching course through an online education manner such as an online class. In a possible implementation, the scenario may also be an offline scenario, namely the target object watches the teaching course taught on site by the teacher through a conventional face-to-face teaching manner, or the target object watches the teaching course played in a video form or another multimedia form in a specific teaching place such as a classroom.

[0037] A specific implementation form of the video may be determined flexibly according to an application scenario of the video processing method. In a possible implementation, the video may be a real-time video, such as a video recorded in real time by the target object in an online class learning process, or a real-time video collected through a camera arranged in the classroom in a process that the target object attends a class in the classroom. In a possible implementation, the video may also be a recorded video, such as a played-back video, recorded during online class learning of the target object, about learning of the target object; or a complete classroom learning video collected through the camera arranged in the classroom when the target object attends the class in the classroom.

[0038] For convenient description, a video processing process is described in each subsequent embodiment of the disclosure taking the video, recorded in real time, of the target object in the online class learning process as an example. A video processing process in another application scenario may be extended flexibly with reference to each subsequent

embodiment of the disclosure, and will not be elaborated herein.

[0039] After the video as described in each abovementioned embodiment of the disclosure is acquired through S11, the at least one type of the learning behavior of the target object in the process of watching the teaching course may be detected through S12. The type and number of the detected learning behavior may be determined flexibly according to the practical condition, and are not limited to each of the following embodiments of the disclosure. In a possible implementation, the learning behavior performed by the target object may include at least one type of the following behaviors: for example, making at least one target gesture, expressing a target emotion, paying attention to a display region of the teaching course, performing at least one interactive behavior with another object, not appearing in at least part of video frames in the video, closing the eyes, and eye contact in the display region of the teaching course.

[0040] The target gesture may be some preset gestures that may be made by the target object in the process of watching the teaching course. A specific implementation form thereof may be set flexibly according to the practical condition. Details refer to each subsequent embodiment of the disclosure, and are temporarily omitted herein.

[0041] The target emotion may be some emotions reflecting a true feeling of the target object about the teaching course in the process of watching the teaching course. A specific implementation form thereof may also be set flexibly according to the practical condition, and is temporarily not detailed herein.

[0042] Paying attention to the display region of the teaching course may reflect a degree of attention of the target object in the process of watching the teaching course. A specific regional range of the display region may be set flexibly according to the practical condition, and is not limited to each of the following embodiments of the disclosure. In a possible implementation, the display region may be a display region of the teaching course video in the online class. For example, when a student learns online with a terminal device through a computer, a mobile phone, or a tablet computer, the display region may be a screen that plays the teaching course in the terminal device. In a possible implementation, the display region may be a teaching region of the teacher in an offline class, such as a position of a platform or a blackboard in the classroom.

[0043] The at least one interactive behavior with the other object may be a learning related interaction of the target object with the other related object in the teaching course in the process of watching the teaching course. An implementation of the other object may be determined flexibly according to the practical condition. In a possible implementation, the other object may be a teaching object, such as the teacher. In a possible implementation, the other object may also be a learning object except the target object in the teaching process, such as a classmate of the target object. The interactive behavior with the other object may be changed flexibly according to different objects. In a possible implementation, under the condition that the other object is the teacher, the interaction with the other object may include receiving an award from the teacher, for example, receiving a red flower or praise by name from the teacher. In a possible implementation, under the condition that the other object is the teacher, the interaction with the other object may include answering a question asked by the teacher, speaking after called by name by the teacher, etc. In a possible implementation, under the condition that the other object is the classmate, the interaction with the other object may include group help, group discussion, group learning, etc.

[0044] Not appearing in at least part of video frames in the video may be a condition that the learning object leaves the teaching course at some moments or in some time periods, etc. For example, the target object may temporarily leave the present online learning device or leave a shooting range of the present online learning device for a private reason in the online learning process.

[0045] Closing the eyes may be an eye closing operation of the target object in the process of watching the teaching course. Eye contact in the display region of the teaching course may be watching the display region of the teaching course. Correspondingly, the condition that the target object does not watch the display region of the teaching course, etc., may further be determined according to the condition of eye contact of the target object in the display region of the teaching course in the video.

[0046] Behavior detection may be performed on the learning process of the target object comprehensively and flexibly through various types of learning behaviors mentioned in the abovementioned embodiment of the disclosure, so that the comprehensiveness and accuracy of the learning state information obtained according to detection are improved, and a

learning state of the target object may be mastered more flexibly and accurately.

[0047] Specifically, one or more types detected in S12 in various types of learning behaviors in the abovementioned embodiment of the disclosure may be set flexibly according to the practical condition. In a possible implementation, various types of learning behaviors mentioned in the abovementioned embodiment of the disclosure may be detected at the same time. Details about a specific detection manner and process refer to each of the following embodiments of the disclosure, and are temporarily omitted herein.

[0048] Under the condition of detecting that the target object performs the at least one type of the learning behavior, the learning state information may be generated according to the at least part of video frames including the at least one type of the learning behavior and/or the duration for the target object to perform the at least one type of the learning behavior. A specific implementation form of the learning state information may be determined flexibly according to the type of the learning behavior and a correspondingly performed operation. In a possible implementation, under the condition that the learning state information is generated according to the at least part of video frames including the at least one type of the learning behavior, the learning state information may include information consisting of the video frames. In a possible implementation, under the condition that the learning state information is generated according to the duration for the target object to perform the at least one type of the learning behavior, the learning state information may be data information in a digital form. In a possible implementation, the learning state information may also include information in both forms, i.e., video frame information and the data information. In a possible implementation, the learning state information may also include information of another state, etc. Specifically, how to generate the learning state information and the implementation form of the learning state information may refer to each subsequent embodiment of the disclosure, and are temporarily not detailed herein.

[0049] In the embodiment of the disclosure, under the condition of detecting that the target object performs the at least one type of the learning behavior, visual learning state information may be generated by using the video frames including the learning behavior, and quantitative learning state information may be generated according to the duration of the learning behavior. In such a manner, learning state information with an evaluation value may

be obtained flexibly, and a related person and institution such as a teacher or a parent may master a learning state of a student effectively and accurately.

[0050] As mentioned in the abovementioned embodiment of the disclosure, the video may be a video, recorded in the process of watching the teaching course, of the target object, while the scenario that the target object watches the teaching course may be determined flexibly according to the practical condition. Therefore, a manner for acquiring the video in S11 may also be correspondingly changed flexibly according to different scenarios. In a possible implementation, under the condition that the scenario that the target object watches the teaching course is the online scenario, namely under the condition that the target object watches the teaching course through the online class, the manner for acquiring the video may include that: if the video processing apparatus and the online learning device of the target object are the same device, video collection may be performed on the process that the target object watches the teaching course through the online learning device of the target object; and if the video processing apparatus and the online learning device of the target object are different devices, video collection may be performed on the process that the target object watches the teaching course through the online learning device of the target object, and the collected video is transmitted to the video processing apparatus in real time and/or not in real time. In a possible implementation, under the condition that the scenario that the target object watches the teaching course is the offline scenario (namely under the condition that the target object attends face-to-face teaching or watches the teaching course in a specific teaching scenario, etc.), the manner for acquiring the video may include that: the video of the target object is collected through an image collection device (such as an ordinary camera, and a shooting device arranged as required by safety) arranged on site. Furthermore, if the image collection device arranged on site may perform video processing, namely the image collection device may be used as the video processing apparatus, a video acquisition process in S11 has been completed. If the image collection device arranged on site may not perform video processing, the video collected by the image collection device arranged on site may be transmitted to the video processing apparatus in real time and/or not in real time.

[0051] As described in each abovementioned embodiment of the disclosure, a manner for detecting the learning behavior of the target object in S12 may be determined flexibly according to the practical condition. In a possible implementation, S12 may include the

following operations.

[0052] In S121, a target object detection is performed on the video to obtain the video frames including the target object.

[0053] In S122, the at least one type of the learning behavior is detected in the video frames including the target object.

[0054] Through the abovementioned embodiment of the disclosure, it can be seen that, in a possible implementation, target object detection may be performed on the video to determine the video frames including the target object in the video. After the video frames including the target object are determined, the at least one type of the learning behavior of the target object in the video frames including the target object may be detected.

[0055] A target object detection manner may be determined flexibly according to the practical condition, and is not limited to the following embodiments. In a possible implementation, the target object in the video may be detected by face detection, face tracking, or other manners. In a possible implementation, multiple objects may be detected after the video frames are detected by face detection, face tracking, or the other manners. In such case, detected face images may further be screened to select one or more objects as target objects. A specific screening manner may be set flexibly according to the practical condition, and is not limited in the embodiment of the disclosure.

[0056] In a possible implementation, after the video frames including the target object are obtained, the at least one type of the learning behavior may be detected in the video frames including the target object through S122. An implementation of S122 may be changed flexibly according to different learning behaviors. Details refer to each of the following embodiments of the disclosure, and are temporarily omitted herein. Under the condition that multiple types of learning behaviors of the target object are required to be detected, multiple manners may be combined to implement detection of the multiple types of learning behaviors.

[0057] In a possible implementation, after target object detection is performed on the video, learning behavior detection over the target object in the process of watching the teaching course may be completed. That is, target object detection may be performed on the video to determine the learning behavior of not appearing in at least part of video frames in

the video, mentioned in the abovementioned embodiment of the disclosure. Furthermore, the learning state information is obtained according to the video frames where the target object is not detected, or time period that the target object does not appear in the at least part of video frames in the video is counted, according to the video frames where the target object is not detected, as the learning state information.

[0058] In the embodiment of the disclosure, target object detection is performed on the video to obtain the video frames including the target object, and the at least one type of the learning behavior is detected in the video frames including the target object. Through the process, more targeted detection for the at least one type of the learning behavior of the target object may be implemented by target object detection over the video, so that the learning behavior may be detected more accurately, and furthermore, the accuracy and reliability of the subsequently obtained learning state information may be improved.

[0059] As described in each abovementioned embodiment of the disclosure, the implementation of S122 may be changed flexibly according to the difference of learning behaviors. In a possible implementation, the learning behavior may include making at least one target gesture.

[0060] In such case, the operation that the at least one type of the learning behavior is detected in the video frames including the target object may include the following operations.

[0061] The at least one target gesture is detected in the video frames including the target object.

[0062] Under the condition of detecting that the number of continuous video frames including the at least one target gesture exceeds a first threshold, at least one frame of the video frames including the target gesture is recorded as a gesture starting frame.

[0063] Under the condition that the number of continuous video frames, not including the target gesture in video frames after the gesture starting frame, exceeds a second threshold, at least one frame of the video frames not including the target gesture is recorded as a gesture ending frame.

[0064] A number and/or a time that the target object makes the at least one target

gesture in the video are/is determined according to the number of the gesture starting frames and the gesture ending frames.

[0065] Through the abovementioned embodiment of the disclosure, it can be seen that, under the condition that the learning behavior includes making the at least one target gesture, learning behavior detection over the video frames including the target object may include target gesture detection.

[0066] Specific gestures that the target gesture includes may be set flexibly according to the practical condition, and are not limited to the following embodiments of the disclosure. Exemplarily, the target gesture includes one or more of a hands-up gesture, a thumbs-up gesture, an OK gesture, and a victory gesture.

[0067] In a possible implementation, the target gesture may include a learning related gesture reflected by the target object according to a listening condition in the process of watching the teaching course, such as the hands-up gesture for answering questions, the thumbs-up gesture (turning up the thumb) for expressing appreciation to a taught content or the teacher, the OK gesture for expressing an understanding to the taught content or agreement with the taught content, and the victory gesture (such as the Yeah gesture) for interacting with the teacher.

[0068] Specifically, a manner for detecting the at least one target gesture in the video frames including the target object may be determined flexibly according to the practical condition, and is not limited to the following embodiments of the disclosure. In a possible implementation, the target gesture may be detected through an algorithm related to gesture recognition. For example, a hand key point of the target object or an image region corresponding to a hand detection box in the video frames may be recognized, gesture detection is performed based on the hand key point or the image region corresponding to the hand detection box, and whether the target object is performing the target gesture is determined based on a gesture detection result. In a possible implementation, the target gesture may be detected through a neural network with a gesture detection function. A specific structure and implementation of the neural network with the gesture detection function may be set flexibly according to the practical condition. Under the condition that the target gesture includes multiple gestures, in a possible implementation, the video frames

including the target object may be input to a neural network capable of simultaneously detecting multiple gestures to implement detection of the target gestures. In a possible implementation, the video frames including the target object may also be input to multiple neural networks with a single gesture detection function respectively to implement a detection of the multiple target gestures.

[0069] In a process of performing target gesture detection through any abovementioned embodiment of the disclosure, under the condition of detecting that the number of the continuous video frames including the at least one target gesture exceeds the first threshold, the at least one frame may be selected from these continuous video frames including the target gesture as the gesture starting frame. A value of the first threshold may be set flexibly according to the practical condition. Values of first thresholds corresponding to different target gestures may be the same or different. For example, the first threshold corresponding to the hands-up gesture may be set to 6, and the first threshold corresponding to the thumbs-up gesture may be set to 7. Under the condition of detecting that the number of continuous video frames including the hands-up gesture is not less than 6, at least one frame may be selected from the video frames including the hands-up gesture as a gesture starting frame of the hands-up gesture. Under the condition of detecting that the number of continuous video frames including the thumbs-up gesture is not less than 7, at least one frame may be selected from the video frames including the thumbs-up gesture as a gesture starting frame of the thumbs-up gesture. In a possible implementation, for conveniently detecting the target gesture, the first thresholds corresponding to different target gestures may be set to the same value. In an example, the value of the first threshold may be set to 6.

[0070] A manner for selecting the gesture starting frame may also be set flexibly according to the practical condition. In a possible implementation, the first frame in the detected continuous video frames including the target gesture may be determined as the gesture starting frame of the target gesture. In a possible implementation, for reducing a gesture detection error, a certain frame after the first frame in the detected continuous video frames including the target gesture may be determined as the gesture starting frame of the target gesture.

[0071] After the gesture starting frame is determined, the gesture ending frame may be

determined from the video frames after the gesture starting frame, namely ending time of the target gesture in the gesture starting frame is determined. A specific determination manner may be selected flexibly according to the practical condition, and is not limited to the following embodiments of the disclosure. In a possible implementation, under the condition of detecting that the number of the continuous video frames not including the target gesture in the gesture starting frame in the video frames after the gesture starting frame exceeds the second threshold, the at least one frame of the video frames not including the target gesture is recorded as the gesture ending frame. A value of the second threshold may also be set flexibly according to the practical condition. Values of second thresholds corresponding to different target gestures may be the same or different. A specific setting manner may refer to the first threshold, and will not be elaborated herein. In an example, the values of the second thresholds corresponding to different target gestures may be the same. For example, the second threshold may be set to 10, namely it may be determined that the target object stops making the target gesture under the condition of detecting that 10 continuous frames after the gesture starting frame do not include the target gesture in the gesture starting frame. In such case, the at least one frame may be selected from the continuous video frames not including the target gesture as the gesture ending frame. A selection manner may also refer to the gesture starting frame. In an example, the last frame in the continuous video frames not including the target gesture may be determined as the gesture ending frame. In an example, a certain frame before the last frame in the continuous video frames not including the target gesture may be determined as the gesture ending frame. In a possible implementation, if it is detected that there is one or more video frames not including the target gesture after the gesture starting frame, the one or more video frames not including the target object may also be determined as the gesture ending frame.

[0072] After the gesture starting frame and the gesture ending frame are determined, the number that the target object makes one or more target gestures may be determined according to the number of the gesture starting frames and the number of gesture ending frames in the video frames. Furthermore, a duration of making the one or more target gestures may be determined. The determination of contents related to the target gesture may be determined flexibly as required by the learning state information in S13. Details refer to each subsequent embodiment of the disclosure, and are temporarily omitted herein.

[0073] The at least one target gesture is detected in the video frames including the target object, and the gesture starting frame and the gesture ending frame are determined according to a detection condition to further determine the number and/or the time that the target object makes the at least one target gesture in the video. Through the process, gestures fed back by the target object according to the learning state in the video may be detected comprehensively and accurately, so that the comprehensiveness and accuracy of the subsequently obtained learning state information are improved, and furthermore, the learning state of the target object may be mastered accurately.

[0074] In a possible implementation, the learning behavior may include expressing the target emotion.

[0075] In such case, the operation that the at least one type of the learning behavior is detected in the video frames including the target object may include the following operations.

[0076] Expression detection and/or smile value detection are/is performed on the video frames including the target object.

[0077] Under the condition of detecting that the target object in a video frame shows at least one first target expression or a result of the smile value detection exceeds a target smile value, the detected video frame is determined as a first detection frame.

[0078] Under the condition of detecting that the number of continuous first detection frames exceeds a third threshold, it is determined that the target object has the target emotion.

[0079] The target emotion may be any emotion set as practically required. For example, the target emotion may be a happy emotion indicating that the target object is focusing on learning, or a bored emotion indicating that the target object is in a poor learning state. Descriptions will be made in each of the following embodiments of the disclosure taking the condition that the target emotion is the happy emotion as an example. The condition that the target emotion is another emotion may be correspondingly extended with reference to each subsequent embodiment of the disclosure.

[0080] Through the abovementioned embodiment of the disclosure, it can be seen that, under the condition that the learning behavior includes expressing the target emotion, learning

behavior detection over the target object may be implemented by a expression detection and/or a smile value detection. In a possible implementation, the learning behavior of expressing the target emotion may be detected only by expression detection or smile value detection. In a possible implementation, whether the target object expresses the target emotion may be determined by the expression detection and the smile value detection. Descriptions will be made in each subsequent embodiment of the disclosure taking determining whether the target object expresses the target emotion by expression detection and smile value detection as an example. Another implementation may be correspondingly extended with reference to each subsequent embodiment of the disclosure, and will not be elaborated herein.

[0081] Expression detection may include detecting an expression shown by the target object. For example, a specific expression shown by the target object may be detected. Specific expression division may be set flexibly according to the practical condition. In a possible implementation, the expression may be divided into happiness, calmness, and other, etc. Smile value detection may include detecting a smile intensity of the target object. For example, a smile amplitude of the target object may be detected. The result of the smile value detection may be fed back through a value. For example, the result of the smile value detection may be set in $[0, 100]$. If the value is greater, it indicates that the smile intensity or amplitude of the target object is higher. Specific expression detection and smile value detection manners may be determined flexibly according to the practical condition. Any manner that the expression or smile degree of the target object may be detected may be determined as the corresponding detection manner. The manners are not limited to each of the following embodiments of the disclosure. In a possible implementation, expression detection over the target object may be implemented through an expression recognition neural network. In a possible implementation, smile value detection over the target object may be implemented through a smile value detection neural network. Specific structures and implementations of the expression recognition neural network and the smile value detection neural network are not limited in the embodiment of the disclosure. Any neural network that may be trained to realize an expression recognition function and any neural network that may be trained to realize a smile value detection function may be applied to the embodiment of the disclosure. In a possible implementation, a face key point and mouth key point of the target object in the video may be detected to implement expression detection and smile value

detection respectively.

[0082] Specific detection results of expression detection and smile value detection may be set flexibly according to the practical condition to determine that the target object has the target emotion. In a possible implementation, it may be determined that the target object in the video frame expresses the target emotion under the condition of detecting that the target object in the video frame shows the at least one first target expression or the result of the smile value detection exceeds the target smile value. In such case, the video frame may be determined as the first detection frame. A specific expression type of the first target expression may be set flexibly according to the practical condition, and is not limited to the following embodiments of the disclosure. In a possible implementation, happiness may be determined as the first target expression, namely all detected video frames where the expression of the target object is happiness may be determined as first detection frames. In a possible implementation, happiness and calmness may be determined as first target expressions, namely all detected video frames where the expression of the target object is happiness or calmness may be determined as first detection frames. Similarly, a specific value of the target smile value may be set flexibly according to the practical condition, and is not specifically limited herein. Therefore, in a possible implementation, a video frame of which a result of the smile value detection exceeds the target smile value may be determined as a first detection frame.

[0083] In a possible implementation, it may be determined that the target object has the target emotion under the condition of detecting that a certain video frame is a first detection frame. In a possible implementation, for improving the detection accuracy and reducing the influence of a detection error on a learning behavior detection result, it may be determined that the target object has the target emotion under the condition of detecting that the number of the continuous first detection frames exceeds the third threshold. A sequence of video frames of which each frame in the continuous video frames is a first detection frame may be determined as continuous first detection frames. A value of the third threshold may be set flexibly according to the practical condition, and the value may be the same as or different from the first threshold or the second threshold. In an example, the value of the third threshold may be 6, namely it may be determined that the target object has the target emotion under the condition of detecting that 6 continuous frames are first detection frames.

[0084] Furthermore, after it is determined that the target object has the target emotion, one frame may be selected from the continuous first detection frames as a target emotion starting frame. Then, under the condition that the expression of the target object is detected not to be the first target expression in 10 continuous frames after the target emotion starting frame, or the result of the smile value detection of the target object in the 10 continuous frames do not exceed the third threshold, or the target object is not detected in one or more frames, a target emotion ending frame may further be determined. Then, the number and/or the time etc. that the target object has the target emotion are/is determined according to the target emotion starting frame or the target emotion ending frame. A specific process may refer to the corresponding process for the target gesture, and will not be elaborated herein.

[0085] Expression detection and/or smile value detection are/is performed on the video frames including the target object, and the first detection frame is determined according to the expression detection and the result of the smile value detection, so that it is determined that the target object has the target emotion under the condition of detecting that the number of the continuous first detection frames exceeds the third threshold. Through the process, the emotion of the target object in the learning process may be determined flexibly based on the expression and smile degree of the target object, so that an emotional state of the target object in the learning process may be perceived more comprehensively and accurately, and more accurate learning state information may be generated.

[0086] In a possible implementation, the learning behavior may include paying attention to the display region of the teaching course.

[0087] In such case, the operation that the at least one type of the learning behavior is detected in the video frames including the target object may include the following operations.

[0088] Expression detection and headpose detection are performed on the video frames including the target object.

[0089] Under the condition of detecting that the target object in a video frame shows at least one second target expression and a headpose is in a target headpose range, the detected video frame is determined as a second detection frame.

[0090] Under the condition of detecting that the number of continuous second

detection frames exceeds a fourth threshold, it is determined that the target object pays attention to the display region of the teaching course.

[0091] An implementation form of the display region of the teaching course may refer to each abovementioned embodiment of the disclosure, and will not be elaborated herein.

[0092] Through the abovementioned embodiment of the disclosure, it can be seen that, under the condition that the learning behavior includes paying attention to the display region of the teaching course, learning behavior detection over the target object may be implemented by expression detection and headpose detection. In a possible implementation, the learning behavior of paying attention to the display region of the teaching course may also be detected only by headpose detection. Descriptions will be made in each subsequent embodiment of the disclosure taking determining whether the target object pays attention to the display region of the teaching course by expression detection and headpose detection as an example. Another implementation may be correspondingly extended with reference to each subsequent embodiment of the disclosure, and will not be elaborated herein.

[0093] An implementation of expression detection may refer to each abovementioned embodiment of the disclosure, and will not be elaborated herein. A headpose detection may be a detection to angular orientation of the face. A specific headpose detection manner may be determined flexibly according to the practical condition. Any manner that the headpose of the target object may be detected may be determined as the headpose detection manner. The manner is not limited to each of the following embodiments of the disclosure. In a possible implementation, headpose detection over the target object may be implemented through a headpose detection neural network. A specific structure and implementation of the headpose detection neural network are not limited in the embodiment of the disclosure. Any neural network that may be trained to realize a headpose detection function may be applied to the embodiment of the disclosure. In a possible implementation, the face key point of the target object in the video may be detected to determine the headpose of the target object. A form of the headpose that may be detected by headpose detection may also be determined flexibly according to the practical condition. In a possible implementation, a yaw angle and a pitch angle of the face of the target object may be detected to determine the headpose of the target object.

[0094] Specific detection results of expression detection and headpose detection may be set flexibly according to the practical condition to determine that the target object pays attention to the display region of the teaching course. In a possible implementation, it may be determined that the target object in the video frame pays attention to the display region of the teaching course under the condition of detecting that the target object in the video frame shows the at least one second target expression and the detected headpose is in the target headpose range. In such case, the video frame may be determined as the second detection frame. A specific expression type of the second target expression may be set flexibly according to the practical condition, may be the same as the first target expression mentioned in the abovementioned embodiment of the disclosure, or may be different from the first target expression mentioned in the abovementioned embodiment of the disclosure, and is not limited to the following embodiment of the disclosure. In a possible implementation, calmness may be determined as the second target expression, namely all detected video frames where the expression of the target object is calmness and the headpose is in the target headpose range may be determined as second detection frames. In a possible implementation, an expression except other may be determined as the second target expression, namely all detected video frames where the headpose of the target object is in the target headpose range and the expression is not "other" may be determined as second detection frames. Similarly, a specific range value of the target headpose range may be set flexibly according to the practical condition, and is not specifically limited herein. In a possible implementation, the target headpose range may be static. In an example, a general position (for example, a platform region where the teacher is in the offline scenario) that the teacher may move to in teaching may be determined as the target headpose range. In an example, a fixed region (for example, a display screen that the target object pays attention to in the online scenario) that the target object may pay attention to in the process of watching the teaching course may be determined as the target headpose range. In a possible implementation, the target headpose range may also be dynamic. In an example, the target headpose range may be determined flexibly according to a present position that the teacher is in teaching, namely the value of the target headpose range may be dynamically changed with movement of the teacher.

[0095] In a possible implementation, it may be determined that the target object pays attention to the display region of the teaching course under the condition of detecting that a

certain video frame is a second detection frame. In a possible implementation, for improving the detection accuracy and reducing the influence of a detection error on the learning behavior detection result, it may be determined that the target object pays attention to the display region of the teaching course under the condition of detecting that the number of the continuous second detection frames exceeds the fourth threshold. A sequence of video frames of which each frame in the continuous video frames is a second detection frame may be determined as continuous second detection frames. A value of the fourth threshold may be set flexibly according to the practical condition, and the value may be the same as or different from the first threshold, the second threshold, or the third threshold. In an example, the value of the fourth threshold may be 6, namely it may be determined that the target object pays attention to the display region of the teaching course under the condition of detecting that 6 continuous frames are second detection frames.

[0096] Furthermore, after it is determined that the target object pays attention to the display region of the teaching course, one frame may be selected from the continuous second detection frames as an attention starting frame. Then, under the condition that the expression of the target object is detected not to be the second target expression in 10 continuous frames after the attention starting frame, or headposes of the target object in the 10 continuous frames are not in the target headpose range, or the target object is not detected in one or more frames, an attention ending frame may further be determined. Then, the number and/or the time etc., that the target object pays attention to the display region of the teaching course are/is determined according to the attention starting frame or the attention ending frame. A specific process may refer to the corresponding processes for the target gesture and the target emotion, and will not be elaborated herein.

[0097] Expression detection and headpose detection are performed on the video frames including the target object, and the second detection frame is determined according to expression detection and the result of headpose detection, so that it is determined that the target object pays attention to the display region of the teaching course under the condition of detecting that the number of the continuous second detection frames exceeds the fourth threshold. Through the process, whether the target object pays attention to the display region of the teaching course may be determined flexibly based on the expression and headpose of the target object, so that a concentration condition of the target object in the learning process

may be perceived more comprehensively and accurately, and more accurate learning state information may be generated.

[0098] In a possible implementation, the learning behavior may further include performing the at least one interactive behavior with the other object. An implementation of the interactive behavior may refer to each abovementioned embodiment of the disclosure, and will not be elaborated herein. In such case, a manner for performing interactive behavior detection on the video frames including the target object may be determined flexibly according to the practical condition. In a possible implementation, if the interactive behavior is an online interactive behavior, for example, receiving the red flower sent by the teacher through the online class, or speaking after called by name by the teacher in the online class, the manner for detecting the interactive behavior may be directly determining whether the target object performs the interactive behavior according to a signal transmitted by the other object. In a possible implementation, if the interactive behavior is an offline interactive behavior, for example, speaking after the target object is called by name by the teacher in the classroom, the manner for detecting whether the target object performs the interactive behavior may include that: a target action of the target object is recognized to determine whether the target object performs the interactive behavior. The target action may be set flexibly according to a practical condition of the interactive behavior. For example, the target action may include standing up to speak, or speaking facing the other object for a time more than a certain time value, etc.

[0099] In a possible implementation, the learning behavior may further include not appearing in at least part of video frames in the video. In such case, S12 may include the following operations.

[00100] Target object detection is performed on the video to obtain the video frames including the target object, and the video frames except the video frames including the target object in the video are determined as video frames where the target object is not detected.

[00101] Under the condition that the number of the video frames where the target object is not detected exceeds a preset video frame number, it is detected that the learning behavior includes not appearing in at least part of video frames in the video.

[00102] Details about the manner for performing target object detection on the video refer to each abovementioned embodiment of the disclosure, and will be omitted herein. In a possible implementation, besides the video frames including the target object, video frames in the video may further include video frames not including the target object. Therefore, these video frames not including the target object may be determined as video frames where the target object is not detected, and under the condition that the number of the video frames where the target object is not detected exceeds the preset video frame number, it is determined that the learning behavior of "not appearing in at least one part of video frames in the video" is detected. The preset video frame number may be set flexibly according to the practical condition. In a possible implementation, the preset video frame number may be set to 0, namely it is determined that the learning behavior of not appearing in at least part of video frames in the video is detected under the condition that the video includes the video frames where the target object is not detected. In a possible implementation, the preset video frame number may also be a number greater than 0. How to set may specifically be determined flexibly according to the practical condition.

[00103] In a possible implementation, the learning behavior may further include closing the eyes. In such case, the learning behavior detection manner may be eye closing detection. A specific eye closing detection process may be set flexibly according to the practical condition. In an example, the process may be implemented through a neural network with an eye closing detection function. In an example, whether the target object closes the eyes, etc., may also be determined by detecting key points of the eyes and in the eyeballs. For example, under the condition of detecting the key points in the eyeballs, it is determined that the target object opens the eyes. Under the condition of detecting the key points of the eyes and not detecting the key points in the eyeballs, it is determined that the target object closes the eyes. In a possible implementation, the learning behavior may further include eye contact in the display region of the teaching course. In such case, the learning behavior detection manner may refer to the process for paying attention to the display region of the teaching course in the abovementioned embodiment of the disclosure. A specific detection manner may be changed flexibly. For example, eye closing detection and headpose detection may be performed on the target object at the same time, a video frame where the headpose is in the target headpose range and the eyes are opened is determined as a third detection frame, and

under the condition that the number of third detection frames exceeds a certain set threshold, it is determined that the target object performs eye contact in the display region of the teaching course.

[00104] After the implementations of the abovementioned embodiments of the disclosure are combined freely to implement detection of the at least one type of the learning behavior of the target object, the learning state information may be generated through S13 under the condition of detecting that the target object performs the at least one type of the learning behavior. A specific implementation of S13 is not limited, may be changed flexibly according to a practical condition of the detected learning behavior, and is not limited to each of the following embodiments of the disclosure.

[00105] Through the practical content of S13 in the abovementioned embodiment of the disclosure, it can be seen that there may be multiple generation manners in a process of generating the learning state information in S13. For example, the learning state information may be generated according to the video frames including the at least one type of the learning behavior. Alternatively, the learning state information is generated according to the duration for the target object to perform the at least one type of the learning behavior. Alternatively, the two conditions are combined, namely a part of learning state information is generated according to the video frames including the at least one type of the learning behavior, and another type of learning state information is generated according to the duration for the target object to perform the at least one type of the learning behavior. Under the condition that not only may the learning state information generated according to the video frames including the learning behavior but also the learning state information may be generated according to the duration for the target object to perform the at least one type of the learning behavior, a specific mapping manner of generating learning state information according to a corresponding type of learning behavior may be set flexibly according to the practical condition. In a possible implementation, some positive learning behaviors may correspond to the process of generating, according to the video frames including the learning behavior, the learning state information. For example, under the condition that the target object makes the at least one target gesture, expresses a positive target emotion, pays attention to the display region of the teaching course, and performs the at least one interactive behavior with the other object, learning state information may be generated according to video frames including the

learning behaviors. In a possible implementation, for some negative learning behaviors, for example, under the condition that the target object does not appear in at least part of video frames in the video, closes the eyes, or does not perform eye contact in the display region of the teaching course, learning state information may be generated according to durations of the learning behaviors.

[00106] In a possible implementation, the operation that the learning state information is generated according to the at least part of video frames including the at least one type of the learning behavior may include the following operations.

[00107] In S1311, the video frames including the at least one type of the learning behavior in the video are acquired as a target video frame set.

[00108] In S1312, face quality detection is performed on at least one video frame in the target video frame set, and a video frame where face quality is greater than a face quality threshold is determined as a target video frame.

[00109] In S1313, the learning state information is generated according to the target video frame.

[00110] The video frame including the at least one type of the learning behavior may be a video frame where the target object is detected to perform the at least one type of behavior in a process of a learning behavior detection, such as the first detection frame, second detection frame, and third detection frame mentioned in the abovementioned embodiment of the disclosure, or a video frame which includes the target gesture and is between the gesture starting frame and the gesture ending frame.

[00111] After the video frames including the at least one type of the learning behavior are determined, an implementation for how to obtain the target video frame set may be determined flexibly. In a possible implementation, each video frame including each type of learning behavior is acquired according to the types of the learning behavior, thereby forming a target video frame set of each type of the learning behavior. In a possible implementation, part of frames including each type of learning behavior, etc., may be acquired according to the types of the learning behavior, and then the target video frame set of each type of learning behavior is obtained based on the part of frames of this type of learning behavior. A specific

manner for selecting the part of frames may be determined flexibly.

[00112] After the target video frame set corresponding to the learning behavior is obtained, the target video frame may be selected from the target video frame set through S1312. Through S1312, it can be seen that, in a possible implementation, face quality detection may be performed on the video frames in the target video frame set, and then the video frame where the face quality is greater than the face quality threshold is determined as the target video frame.

[00113] A face quality detection manner may be set flexibly according to the practical condition, and is not limited to the following embodiments of the disclosure. In a possible implementation, face recognition may be performed on the face in the video frame, thereby determining the completeness of the face in the video frame to determine the face quality. In a possible implementation, the face quality may also be determined based on a resolution of the face in the video frame. In a possible implementation, the face quality in the video frame may be comprehensively evaluated based on multiple parameters such as the completeness, the resolution and brightness of the face in the video frame. In a possible implementation, the video frame may be input to a face quality neural network to obtain the face quality in the video frame. The face quality neural network may be obtained by training through a large number of face images with scored and tagged face quality, and a specific implementation form thereof may be selected flexibly according to the practical condition, and is not limited in the embodiment of the disclosure.

[00114] A specific value of the face quality threshold may be determined flexibly according to the practical condition, and no limits are made thereto in the embodiment of the disclosure. In a possible implementation, a different face quality threshold may be set for each type of learning behavior. In a possible implementation, the same face threshold may also be set for each type of learning behavior. In a possible implementation, the face quality threshold may be set to a maximum value of the face quality in the target video frame set. In such case, the video frame with highest face quality under each type of learning behavior may be directly determined as the target video frame.

[00115] In some possible implementations, some video frames may simultaneously include multiple types of learning behaviors. In such case, a processing manner for the video

frames including the multiple types of learning behaviors may be changed flexibly according to the practical condition. In a possible implementation, these video frames may be divided to each type of learning behavior, and then a target video frame is selected from a video frame set corresponding to each type of learning behavior according to S1312. In a possible implementation, the video frames simultaneously including the multiple types of learning behaviors may also be directly selected as target video frames.

[00116] After the target video frame is determined through any abovementioned embodiment, the learning state information may be generated according to the target video frame through S1313. An implementation of S1313 may be selected flexibly according to the practical condition. Details refer to each of the following embodiments of the disclosure, and are temporarily omitted herein.

[00117] In the embodiment of the disclosure, the video frames including the at least one type of the learning behavior in the video frames are acquired as the target video frame set, thereby selecting the video frame with relatively high face quality as the target video frame according to the target video frame set of each type of learning behavior, and further generating the learning state information according to the target video frame. Through the process, the generated learning state information may be information obtained based on the video frames with relatively high face quality and including the learning behavior, and is more accurate, so that the learning state of the target object may be mastered more accurately.

[00118] As described in the abovementioned embodiments, the implementation of S1313 may be changed flexibly. In a possible implementation, S1313 may include the following operations.

[00119] At least one frame of the target video frame is determined as the learning state information; and/or,

[00120] a region where the target object is located in the at least one frame of the target video frame is recognized, and the learning state information is generated based on the region where the target object is located.

[00121] Through the abovementioned embodiment of the disclosure, it can be seen that, in a possible implementation, the at least one frame of the target video frame may be directly

determined as the learning state information. In an example, the obtained target video frame may further be selected. Such selection may be random, or may be conditional. Then, the selected target video frame is directly determined as the learning state information. In an example, each obtained target video frame may be directly determined as the learning state information.

[00122] In a possible implementation, the region where the target object is located in the target video frame may further be recognized, thereby generating the learning state information according to the region where the target object is located. A manner for recognizing the region where the target object is located is not limited in the embodiment of the disclosure. In a possible implementation, the manner may be implemented through the neural network with the target object detection function mentioned in the abovementioned embodiment of the disclosure. After the region where the target object is located in the target video frame is determined, corresponding processing may further be performed on the target video frame to obtain the learning state information. A processing manner may be determined flexibly. In an example, an image of the region where the target object is located in the target video frame may be determined as the learning state information. In an example, a background region except the region where the target object is located in the target video frame may be rendered, for example, adding another sticker, or adding mosaic to the background region, or replacing an image of the background region, to obtain the learning state information that does not display a present background of the target object, so that the privacy of the target object may be protected better. A rendering manner such as a sticker may also be used, so that the diversity and attractive appearance of the learning state information are improved.

[00123] The at least one frame of the target video frame is determined as the learning state information, and/or the learning state information is generated according to the region where the target object is located in the target video frame. In such a manner, the finally obtained learning state information may be more flexible, so that learning state information that highlights the target object more or learning state information that protects the privacy of the target object better may be obtained as required by the target object.

[00124] The abovementioned embodiments of the disclosure may be combined freely

to obtain the learning state information generated based on the video frames including the learning behavior. For example, Table 1 shows a generation rule of the learning state information according to an embodiment of the disclosure.

Table 1: Learning State Information Generation Rule

Learning behavior	Detection rule	Data processing	Image processing	Image uploading
Hands-up gesture	The hands-up is detected in N frames in M continuous frames of the same face track (frame sequence)	The timestamp of the first frame where the hands-up is detected is determined as hands-up starting time. If the hands-up is not detected in Z continuous frames, or the face track is lost, the timestamp of the last frame where the hands-up is detected is determined as hands-up ending time. A hands-up event is determined from the starting time to the ending time.	From the start till the end of the hands-up, the original image or matted image with the highest face quality score that is greater than the quality score threshold of the wonderful moment is extracted from all the frames where the hands-up is detected	MJPEG coding and uploading to the background. Y images are uploaded every X minutes.
Thumbs-up gesture	The thumb is detected in N frames in M continuous frames of the same face track (frame sequence)	Ditto	From the start till the end of the thumbs-up gesture, the original image or matted image with the highest face quality score that is greater than the quality score threshold of the wonderful moment is extracted from all the frames where the thumbs-up gesture is detected	MJPEG coding and uploading to the background. Y images are uploaded every X minutes.
OK gesture	The OK gesture is detected in N frames in M continuous frames of the	Ditto	From the start till the end of the OK gesture, the original image or matted image with the highest face quality score that is greater	MJPEG coding and uploading to the background.

	same face track (frame sequence)		than the quality score threshold of the wonderful moment is extracted from all the frames where the OK gesture is detected	Y images are uploaded every X minutes.
Victory gesture	The Yeah gesture is detected in N frames in M continuous frames of the same face track (frame sequence)	Ditto	From the start till the end of gesture Yeah, the original image or matted image with the highest face quality score that is greater than the quality score threshold of the wonderful moment is extracted from all the frames where the Yeah gesture is detected	MJPEG coding and uploading to the background. Y images are uploaded every X minutes.
Interactive behavior- the teacher sends the red flower	The Software Development Kit (SDK) calls the red flower sticker for rendering	Rendering starting time is determined as starting time, and rendering ending time is determined as ending time	None	None
Interactive behavior- the teacher calls by name for speaking	The SDK calls the speaking sticker for rendering	Ditto	None	None
Expressing the happy emotion	It is detected in N frames in M continuous frames of the same face track (frame sequence) that the expression is happiness or the smile	The timestamp of the first frame is determined as the starting time. If it is not detected in Z continuous frames that the expression is happiness or the smile value is greater than the certain threshold, or the face track is	From the start till the end, the original image or matted image with the highest face quality score that is greater than the quality score threshold of the wonderful moment is extracted for uploading from all the frames where it is detected that the	MJPEG coding and uploading to the background. Y images are uploaded every X minutes.

	value is greater than a certain threshold	lost, the timestamp of the last frame where it is detected that the expression is happiness or the smile value is greater than the certain threshold is determined as ending time. A happy event is determined from the starting time to the ending time.	expression is happiness or the smile value is greater than the certain threshold	
Paying attention to the display region of the teaching course (attentively)	It is detected in N frames in M continuous frames of the same face track (frame sequence) that the expression is calmness and headpose is in a watching threshold range	The timestamp of the first frame is determined as the starting time. If it is not detected in Z continuous frames that the expression is calmness and the headpose is in the watching threshold range, or the face track is lost, the timestamp of the last frame where it is detected that the expression is calmness and the headpose is in the watching threshold range is determined as the ending time. A concentration event is determined from the starting time to the ending time.	From the start till the end, the original image or matted image with the highest face quality score that is greater than the quality score threshold of the wonderful moment is extracted for uploading from all the frames where it is detected that the expression is calmness or the headpose is in the watching threshold range	MJPEG coding and uploading to the background. Y images are uploaded every X minutes.

[00125] M, N, X, Y, and Z are all positive integers, and specific values may be set as practically required. In addition, the parameters such as M in different rows in Table 1 may be the same or different. The parameters such as M are merely examples and not intended to limit the contents of the disclosure.

[00126] The wonderful moment is a moment corresponding to a positive learning behavior performed by the target object. It can be seen from Table 1 that, in an example, under the condition of detecting the learning behavior such as the target object performs the target gesture such as the hands-up gesture, the target object has the target emotion of happiness, or the target object pays attention to the display region of the teaching course, and the target object interacts with the teacher to speak after called by name, certain data processing may be performed on the video, and after data processing, image processing is further performed on the video frames to obtain the target video frame as the learning state information.

[00127] In a possible implementation, the operation that the learning state information is generated according to the duration for the target object to perform the at least one type of the learning behavior may include the following operations.

[00128] In S1321, under the condition of detecting that time when the target object performs the at least one type of the learning behavior is not less than a time threshold, the duration of the at least one type of the learning behavior is recorded.

[00129] In S1322, the duration corresponding to the at least one type of the learning behavior is determined as the learning state information.

[00130] The time threshold may be a certain value set flexibly according to the practical condition. Time thresholds corresponding to different types of learning behaviors may be the same or different. Under the condition of detecting that the target object performs a certain type of learning behavior in a certain time period, time when the target object performs the learning behavior may be counted for feedback to the teacher or the parent as the learning state information. Implementations of a specific statistical condition and a specific learning behavior that time is counted for may be set flexibly according to the practical condition.

[00131] In a possible implementation, under the condition of detecting that time when the target object does not appear in the video (for example, no person is in the video, a person is in the video frame but cannot be determined whether to be the target object, or a person is in the screen but is not the target object) exceeds a certain time, the target object closes the

eyes, or the target object does not watch the display region of the teaching course, time for these learning behaviors may be counted and determined as the learning state information.

[00132] In the embodiment of the disclosure, under the condition that the time when the target object performs the at least one type of the learning behavior is not less than the time threshold, the duration of the at least one type of the learning behavior is recorded and determined as the learning state information. Through the process, the learning state information may be quantized, so that the learning state of the target object may be mastered more intuitively and accurately.

[00133] In a possible implementation, the video processing method disclosed in the embodiment of the disclosure may further include the following operation.

[00134] Background regions in the at least part of video frames in the video are rendered, the background region being a region except the target object in the video frame.

[00135] A segmentation manner for the background region and a rendering manner for the background region may refer to recognition of the region where the target object is located in the target video frame and the rendering process after recognition in the abovementioned embodiment of the disclosure, and will not be elaborated herein. In a process of rendering the background region, in an example, rendering may be performed through a preset general template in the present video processing apparatus. In an example, rendering may also be implemented by calling another template, custom template, etc., in a database of an apparatus that is not a video processing apparatus. For example, another background template, etc., may be called from a cloud server that is not a video processing apparatus to render the background region in the video.

[00136] The background regions in the at least part of video frames in the video are rendered, so that the privacy of the target object in the video may be protected to reduce the probability that the privacy of the target object is leaked because of no appropriate video collection position on one hand, and on the other hand, the target object is more interested in watching the teaching course.

[00137] In a possible implementation, the video processing method disclosed in the embodiment of the disclosure may further include the following operations.

[00138] Learning state information of at least one target object is counted to obtain a statistical result of the at least one target object.

[00139] Learning state statistics data is generated according to the statistical result of the at least one target object.

[00140] In the embodiment of the disclosure, a video may include one or more target objects. In addition, the video processing method in the embodiment of the disclosure may be used to process a single video, or may be used to process multiple videos. Therefore, correspondingly, learning state information of one target object may be obtained, or learning state information of multiple target objects may be obtained. In such case, the learning state information of at least one target object may be counted to obtain a statistical result of the at least one target object. The statistical result may include the learning state information of the target object, and may also include other information related to teaching course watching of the target object. For example, in a possible implementation, before S12, namely before learning behavior detection is performed on the target object, sign-in data of the target object may also be acquired. The sign-in data of the target object may include identity information, sign-in time, etc., of the target object. A specific manner for acquiring the sign-in data may be determined flexibly according to a practical sign-in manner of the target object, and is not limited in the embodiment of the disclosure.

[00141] After the statistical result of the at least one target object is obtained, the learning state statistics data may be generated according to the at least one statistical result. Specifically, a generation manner and content of the learning state statistics data may be changed flexibly according to an implementation form of the statistical result. Details refer to each of the following embodiments of the disclosure, and are temporarily omitted herein.

[00142] In the embodiment of the disclosure, the learning state information of the at least one target object is counted to obtain the statistical result of the at least one target object, thereby generating the learning state statistics data according to the statistical result of the at least one target object. Through the process, learning states of multiple target objects may be comprehensively evaluated effectively, so that the teacher may master an overall learning condition of the whole class more conveniently, and other related persons may also conveniently know present learning positions, etc., of the target objects more

comprehensively.

[00143] In a possible implementation, the operation that the learning state statistics data is generated according to the statistical result of the at least one target object includes the following operations.

[00144] A statistical result of the target object of at least one classification is acquired according to a classification to which the at least one target object belongs, and learning state statistics data of the at least one classification is generated, classification to which the target object belongs including at least one of: a course that the target object attends, an institution that the target object is registered to, or a device used by the target object; and/or,

[00145] the statistical result of the at least one target object is visualized to generate the learning state statistics data of the at least one target object.

[00146] Classification to which the target object belongs may be a classification divided according to an identity of the target object. For example, the classification of the target object may include at least one of: the course that the target object attends, the institution that the target object is registered to, or the device used by the target object. The course that the target object attends may be the teaching course watched by the target object, mentioned in the abovementioned embodiments of the disclosure. The institution that the target object is registered to may be an educational institution that the target object belongs to, or a grade of the target object, or a class of the target object, etc. The device used by the target object may be a terminal device used by the target object to attend the online course in the online scenario, etc.

[00147] In the embodiment of the disclosure, the statistical result of the target object of the at least one classification may be acquired according to the classification to which the target object belongs, namely at least one statistical result under the classification to which the target object belongs may be summarized to obtain overall learning state statistics data under the classification. For example, statistical results of different target objects under the same device, statistical results of different target objects under the same course, and statistical results of different target objects under the same educational institution may be obtained by division according to the classifications such as the used device, the course, and the

educational institution. In an example, these statistical results may also be displayed in form of a report. In an example, the statistical result under each classification in the report may include overall learning state information of each target object, and may also include specific learning state information of each target object, such as a time length of paying attention to the display region of the teaching course and a smiling time length. In addition, other information related to teaching course watching may further be included, such as the sign-in time of the target object, the number of sign-ins of the target object, a matching condition of the target object and a face in a preset database, a sign-in device, and a course that the target object signs in to.

[00148] In addition, the statistical result of the at least one target object may also be visualized to obtain the learning state statistics data of the at least one target object. A visualization manner may be determined flexibly according to the practical condition. For example, the data may be organized into a form such as a chart or a video. The content in the learning state statistics data may be determined flexibly according to the practical condition. For example, the overall learning state information of the target object, the name of the teaching course watched by the target object, and the specific learning state information of the target object may be included. Data that is specifically included may be set flexibly according to the practical condition. In an example, contents such as the identity of the target object, the name of the teaching course watched by the target object, the duration when the target object pays attention to the display region of the teaching course, an attention degree of the target object, a data comparison result between the target object and another target object, the number of interactions of the target object, and the emotion of the target object may be organized into a visual report for sending to the target object or another person related to the target object, such as the parent of the target object.

[00149] In an example, besides a picture and a video, a form of a text content in the visualized learning state statistics data may be "the subject is XX, and student A keeps attentive for 30 minutes with the attentiveness of focused, higher than 10% of classmates, interacts for three times and smiles for five times, and is hereby praised and expected to keep it up" or "the subject is XX, student B is not so attentive, and interacts with the hands-up gesture and other gestures at a relatively low frequency, and the parent is advised to pay close attention and timely adjust the learning habits of the child".

[00150] In the embodiment of the disclosure, the classification to which the at least one target object belongs is acquired to generate the learning state statistics data of the at least one classification, and/or, the statistical result of the at least one target object is visualized to generate the learning state statistics data of the at least one target object. Through the process, the learning state of the target object may be mastered more intuitively and comprehensively in different data statistics manners.

[00151] FIG. 2 is a block diagram of a video processing apparatus according to an embodiment of the disclosure. As shown in FIG. 2, the video processing apparatus 20 may include a video acquisition module 21, a detection module 22, and a generation module 23.

[00152] The video acquisition module 21 is configured to acquire a video, at least part of video frames in the video including a target object.

[00153] The detection module 22 is configured to detect at least one type of learning behavior of the target object in a process of watching a teaching course according to the video.

[00154] The generation module 23 is configured to, under the condition of detecting that the target object performs the at least one type of the learning behavior, generate learning state information according to the at least part of video frames including the at least one type of the learning behavior and/or a duration for the target object to perform the at least one type of the learning behavior.

[00155] In a possible implementation, the learning behavior includes at least one type of the following behaviors: making at least one target gesture, expressing a target emotion, paying attention to a display region of the teaching course, performing at least one interactive behavior with another object, not appearing in at least part of video frames in the video, closing the eyes, and eye contact in the display region of the teaching course.

[00156] In a possible implementation, the detection module is configured to perform target object detection on the video to obtain the video frames including the target object, and detect the at least one type of the learning behavior in the video frames including the target object.

[00157] In a possible implementation, the learning behavior includes making the at

least one target gesture. The detection module is further configured to: detect the at least one target gesture in the video frames including the target object; under the condition of detecting that the number of continuous video frames including the at least one target gesture exceeds a first threshold, record at least one frame of the video frames including the target gesture as a gesture starting frame; under the condition that the number of continuous video frames not including the target gesture in the video frames after the gesture starting frame exceeds a second threshold, record at least one frame of the video frames not including the target gesture as a gesture ending frame; and the number and/or the time that the target object makes the at least one target gesture in the video according to the number of the gesture starting frames and the number of the gesture ending frames.

[00158] In a possible implementation, the learning behavior includes expressing the target emotion. The detection module is further configured to: perform expression detection and/or smile value detection on the video frames including the target object; under the condition of detecting that the target object in a video frame shows at least one first target expression or a result of the smile value detection exceeds a target smile value, determine the detected video frame as a first detection frame; and under the condition of detecting that the number of continuous first detection frames exceeds a third threshold, determine that the target object has the target emotion.

[00159] In a possible implementation, the learning behavior includes paying attention to the display region of the teaching course. The detection module is further configured to: perform expression detection and headpose detection on the video frames including the target object; under the condition of detecting that the target object in a video frame shows at least one second target expression or a headpose is in a target headpose range, determine the detected video frame as a second detection frame; and under the condition of detecting that the number of continuous second detection frames exceeds a fourth threshold, determine that the target object pays attention to the display region of the teaching course.

[00160] In a possible implementation, the generation module is configured to: acquire the video frames including the at least one type of the learning behavior in the video as a target video frame set; perform face quality detection on at least one video frame in the target video frame set and determine a video frame where face quality is greater than a face quality

threshold as a target video frame; and generate the learning state information according to the target video frame.

[00161] In a possible implementation, the generation module is further configured to: determine at least one frame of the target video frame as the learning state information; and/or, recognize a region where the target object is located in at least one frame of the target video frame and generate the learning state information based on the region where the target object is located.

[00162] In a possible implementation, the detection module is configured to: perform target object detection on the video to obtain the video frames including the target object, and determine video frames except the video frames including the target object in the video as video frames where the target object is not detected; and under the condition that the number of the video frames where the target object is not detected exceeds a preset video frame number, detect that the learning behavior includes not appearing in at least part of video frames in the video.

[00163] In a possible implementation, the generation module is configured to, under the condition of detecting that time when the target object performs the at least one type of the learning behavior is not less than a time threshold, record the duration of the at least one type of the learning behavior, and determine the duration corresponding to the at least one type of the learning behavior as the learning state information.

[00164] In a possible implementation, the apparatus is further configured to render background regions in the at least part of video frames in the video, the background region being a region except the target object in the video frame.

[00165] In a possible implementation, the apparatus is further configured to: count learning state information of at least one target object to obtain a statistical result of the at least one target object, and generate learning state statistics data according to the statistical result of the at least one target object.

[00166] In a possible implementation, the apparatus is further configured to: acquire a statistical result of the target object of at least one classification according to a classification to which the at least one target object belongs, and generate learning state statistics data of

the at least one classification, the classification to which the target object belongs including at least one of: a course that the target object attends, an institution that the target object is registered to, or a device used by the target object; and/or, visualize the statistical result of the at least one target object to generate the learning state statistics data of the at least one target object.

[00167] Different embodiments of the application may be combined without departing from logics, different embodiments are described with different emphases, and parts that are not emphasized may refer to records in the other embodiments.

[00168] In some embodiments of the disclosure, functions or modules of the apparatus provided in the embodiment of the disclosure may be configured to execute the method described in the method embodiment and specific implementation and technical effects thereof may refer to the descriptions about the method embodiment and, for simplicity, will not be elaborated herein.

[00169] Application scenario example

[00170] A learning manner for a student is usually that the teacher teaches and the student listens. The class is lack of interactions and is uninteresting, the student is not so interested in listening, and the student may not be positively motivated through the real-time performance of the student. In addition, the institution or the teacher may also not master the listening state of the student, and the parent may not know the performance of the child at school. Particularly, affected by the outbreak, the student attends online classes at most time. However, whether the student really attends a class and listens carefully and the participation of the student in class may not be quantitatively evaluated. Therefore, how to master a learning state of the student effectively becomes a problem urgent to be solved at present.

[00171] The application example of the disclosure discloses a set of learning system. The system may master a learning state of a student effectively through the video processing method disclosed in the abovementioned embodiments of the disclosure.

[00172] FIG. 3 is a schematic diagram of an application example according to the disclosure. As shown in the figure, in an example, the learning system may consist of three parts, i.e., a user end, an educational Software-as-a-Service (SaaS) background, and an

interactive class background. The student watches a teaching course through the user end. The user end may include two parts, i.e., a hardware device for learning (for example, a client installed with the Windows system or the IOS system and the SDK in the figure), and an Application (APP) configured for the student to sign in to an online class (i.e., the user APP in the figure). The educational SaaS background may be a platform constructed by a server of an educational institution that the student belongs to. The interactive class background may be a platform constructed by a server that summarizes data of different educational institutions and maintains the data. Either the educational SaaS background or the interactive class background may perform data interaction with the user end through an Application Programming Interface (API), thereby implementing learning state information generation and learning state statistics data generation mentioned in each abovementioned embodiments of the disclosure.

[00173] In the application example of the disclosure, a generation process of learning state information may include the following operations.

[00174] The user end collects a video of a process that the student watches the teaching course and processes the collected video to obtain learning state information of each student. The educational SaaS background and the interactive class background call the learning state information generated in different user ends through the API, and count the learning state information in any manner mentioned in the abovementioned embodiments of the disclosure to generate learning state statistics data.

[00175] In an example, the process that the user end processes the collected video to obtain the learning state information of each student may include the following operations.

[00176] A: a wonderful moment of the student in the class (i.e., the positive learning behavior mentioned in the abovementioned embodiments of the disclosure) is acquired.

[00177] In an example, a certain rule may be defined to make video highlights of the student. The performance of the student may be edited into a short video segment or some wonderful pictures and provided for the parent. As such, the parent may timely evaluate the performance of the student in class, and if the effect is good, may encourage the student to continue to attend the related course.

[00178] In an example, the wonderful moment of the student may be acquired after successful sign-in of the student, the video or picture of the wonderful moment may subsequently be uploaded to the background or the cloud, and in addition, the uploaded content of the wonderful moment may also be selected whether to be always visible to the student. In an example, a wonderful moment definition rule may include making at least one target gesture, the target gesture may include hands-up, thumbs-up, the OK gesture, the Yeah gesture, etc., and if it is detected in a period of time that the student makes the above gesture, picture extraction or video frame extraction may be performed on a video including the gesture. The rule may include expressing a target emotion of happiness, and if it is detected in a period of time that an expression of the student is happiness and a smile value reaches a certain target smile value (for example, a score of 99), picture extraction or video frame extraction may be performed on a video frame with a happiness tag or a video frame where the target smile value is reached. The rule may include paying attention to a display region of the teaching course, and if the student keeps facing the front in a period of time, namely the headpose is in a certain threshold range, picture extraction or video frame extraction may be performed on a video in this period of time.

[00179] B: learning condition detection is performed on a learning condition of the student (for the negative learning behavior mentioned in the abovementioned embodiments of the disclosure).

[00180] In an example, data of the condition that the student may be outside the screen or is not attentive may be pushed to the parent in real time by learning condition detection such that the parent may pay attention to the child at the first time and timely correct bad learning habits of the student to achieve an auxiliary supervision effect.

[00181] In an example, a process of performing learning condition detection on the student may be implemented after successful sign-in of the student. If the student does not appear in front of the camera for a long time, does not watch the screen, and closes the eyes, it is judged that the student is not so attentive. In such case, a duration of the above learning behavior of the student may be counted and determined as a learning condition detection result to obtain the corresponding learning state data. A specific learning condition detection configuration rule may refer to each abovementioned embodiment of the disclosure, and will

not be elaborated herein.

[00182] Through each abovementioned example of the disclosure, the learning state information including the wonderful moment and the learning condition detection may be obtained. Furthermore, a process that the educational SaaS background and the interactive class background call the learning state information generated in different user ends through the API to generate the learning state statistics data may include the following operations.

[00183] C: a report is generated (i.e., the learning state statistics data of the at least one classification is generated in the abovementioned embodiments of the disclosure).

[00184] In an example, the API of the background or the cloud may view sign-in information and learning state information of the student in different dimensions such as the device, the course, and the institution. Main data indexes may include: sign-in time, the number of sign-ins, matching with a face library (i.e., the matching condition of the target object and the face in the preset database in the abovementioned embodiments of the disclosure), sign-in device, a course that the student signs in to, a concentration time length, a smiling time length, etc.

[00185] D: the report is analyzed (i.e., the learning state statistics data of the at least one target object is generated by visualization in the abovementioned embodiments of the disclosure).

[00186] In an example, the educational SaaS background or the interactive class background may organize the performance of student in the online class into a complete learning condition analysis report in a unified manner. The report may indicate the learning condition of the student through a visual graphical interface. Furthermore, the background may further select a relatively good condition for pushing to the parent or the teacher for the teacher of the institution to analyze the condition of the student to help the student to improve the learning behavior step by step.

[00187] Besides the abovementioned process, the learning system may further perform background segmentation processing on the learning video of the student in a process that the student learns through the user end. In an example, the user end may provide a background segmentation function for the condition that there is no background position suitable for live

broadcast of the student or the student does not want to display a background picture for privacy protection. In an example, the SDK of the user end may support a plurality of different background templates. For example, a plurality of general templates may be preset. In an example, the student may also call a custom template from the interactive class background through the user end. In an example, the SDK may provide a background template previewing interface for the APP of the user end such that the student may preview, through the APP, the custom template that may be called. During the course, the student may also render the live broadcast background through a background segmentation sticker in the APP of the user end. In an example, the student may manually trigger the sticker to be turned off if unsatisfied with the sticker. The APP of the user end may report data about sticker use of the student to the corresponding background (the educational SaaS background or the interactive class background), and the corresponding background may analyze information of background stickers used by the student and the number of the used stickers as additional learning state information, etc.

[00188] The learning system disclosed in the application example of the disclosure may be applied to online classes, and may further be extended to other related fields such as online meetings.

[00189] It can be understood that each method embodiment mentioned in the disclosure may be combined to form combined embodiments without departing from principles and logics. For saving the space, elaborations are omitted in the disclosure.

[00190] It can be understood by those skilled in the art that, in the method of the specific implementations, the writing sequence of each step does not mean a strict execution sequence and is not intended to form any limit to the implementation process and a specific execution sequence of each step should be determined by functions and probable internal logic thereof.

[00191] The embodiments of the disclosure also disclose a computer-readable storage medium, in which a computer program instruction is stored, the computer program instruction being executed by a processor to implement the abovementioned method. The computer-readable storage medium may be a volatile computer-readable storage medium or a non-volatile computer-readable storage medium.

[00192] The embodiments of the disclosure disclose an electronic device, which includes a processor and a memory configured to store an instruction executable for the processor, the processor being configured for the abovementioned method.

[00193] The embodiments of the disclosure also disclose a computer program, which includes a computer-readable code, the computer-readable code running in an electronic device to enable a processor of the electronic device to execute the abovementioned method.

[00194] In practical applications, the memory may be a volatile memory such as a Random Access Memory (RAM), or a non-volatile memory such as a Read-Only Memory (ROM), a flash memory, a Hard Disk Drive (HDD) or a Solid-State Drive (SSD), or a combination of the memories, and provides instructions and data for the processor.

[00195] The processor may be at least one of an Application Specific Integrated Circuit (ASIC), a Digital Signal Processor (DSP), a Digital Signal Processing Device (DSPD), a Programmable Logic Device (PLD), a Field Programmable Gate Array (FPGA), a Central Processing Unit (CPU), a controller, a microcontroller, or a microprocessor. It can be understood that, for different devices, other electronic components may be configured to realize functions of the processor, and no specific limits are made in the embodiment of the disclosure.

[00196] The electronic device may be provided as a terminal, a server or a device in another form.

[00197] Based on the same technical concept of the abovementioned embodiments, the embodiments of the disclosure also provide a computer program, which is executed by a processor to implement the abovementioned method.

[00198] FIG. 4 is a block diagram of an electronic device 800 according to an embodiment of the disclosure. For example, the electronic device 800 may be a terminal such as a mobile phone, a computer, a digital broadcast terminal, a messaging device, a gaming console, a tablet, a medical device, exercise equipment and a PDA.

[00199] Referring to Fig. 4, the electronic device 800 may include one or more of the following components: a processing component 802, a memory 804, a power component 806,

a multimedia component 808, an audio component 810, an Input/Output (I/O) interface 812, a sensor component 814, and a communication component 816.

[00200] The processing component 802 typically controls overall operations of the electronic device 800, such as the operations associated with display, telephone calls, data communications, camera operations, and recording operations. The processing component 802 may include one or more processors 820 to execute instructions to perform all or part of the steps in the abovementioned method. Moreover, the processing component 802 may include one or more modules which facilitate interaction between the processing component 802 and the other components. For instance, the processing component 802 may include a multimedia module to facilitate interaction between the multimedia component 808 and the processing component 802.

[00201] The memory 804 is configured to store various types of data to support the operation of the electronic device 800. Examples of such data include instructions for any APPs or methods operated on the electronic device 800, contact data, phonebook data, messages, pictures, video, etc. The memory 804 may be implemented by a volatile or non-volatile storage device of any type or a combination thereof, for example, a Static Random Access Memory (SRAM), an Electrically Erasable Programmable Read-Only Memory (EEPROM), an Erasable Programmable Read-Only Memory (EPROM), a Programmable Read-Only Memory (PROM), a ROM, a magnetic memory, a flash memory, a magnetic disk or an optical disk.

[00202] The power component 806 provides power for various components of the electronic device 800. The power component 806 may include a power management system, one or more power supplies, and other components associated with generation, management and distribution of power for the electronic device 800.

[00203] The multimedia component 808 includes a screen providing an output interface between the electronic device 800 and a user. In some embodiments, the screen may include a Liquid Crystal Display (LCD) and a Touch Panel (TP). If the screen includes the TP, the screen may be implemented as a touch screen to receive an input signal from the user. The TP includes one or more touch sensors to sense touches, swipes and gestures on the TP. The touch sensors may not only sense a boundary of a touch or swipe action but also detect a

duration and pressure associated with the touch or swipe action. The touch sensors may not only sense a boundary of a touch or swipe action but also detect a duration and pressure associated with the touch or swipe action. The front camera and/or the rear camera may receive external multimedia data when the electronic device 800 is in an operation mode, such as a photographing mode or a video mode. Each of the front camera and the rear camera may be a fixed optical lens system or have focusing and optical zooming capabilities.

[00204] The audio component 810 is configured to output and/or input an audio signal. For example, the audio component 810 includes a Microphone (MIC), and the MIC is configured to receive an external audio signal when the electronic device 800 is in the operation mode, such as a call mode, a recording mode and a voice recognition mode. The received audio signal may further be stored in the memory 804 or sent through the communication component 816. In some embodiments, the audio component 810 further includes a speaker configured to output the audio signal.

[00205] The I/O interface 812 provides an interface between the processing component 802 and a peripheral interface module, and the peripheral interface module may be a keyboard, a click wheel, a button, etc. The button may include, but not limited to: a home button, a volume button, a starting button and a locking button.

[00206] The sensor component 814 includes one or more sensors configured to provide status assessment in various aspects for the electronic device 800. For instance, the sensor component 814 may detect an on/off status of the electronic device 800 and relative positioning of components, such as a display and small keyboard of the electronic device 800, and the sensor component 814 may further detect a change in a position of the electronic device 800 or a component of the electronic device 800, presence or absence of contact between the user and the electronic device 800, orientation or acceleration/deceleration of the electronic device 800 and a change in temperature of the electronic device 800. The sensor component 814 may include a proximity sensor configured to detect presence of an object nearby without any physical contact. The sensor component 814 may also include a light sensor, such as a Complementary Metal Oxide Semiconductor (CMOS) or Charge Coupled Device (CCD) image sensor, configured for use in an imaging APP. In some embodiments, the sensor component 814 may also include an acceleration sensor, a gyroscope sensor, a

magnetic sensor, a pressure sensor, or a temperature sensor.

[00207] The communication component 816 is configured to facilitate wired or wireless communication between the electronic device 800 and another device. The electronic device 800 may access a communication-standard-based wireless network, such as a Wireless Fidelity (Wi-Fi) network, a 2nd-Generation (2G), 3rd-Generation (3G), 4th-Generation (4G) or 5th-Generation (5G) network, or a combination thereof. In an exemplary embodiment, the communication component 816 receives a broadcast signal or broadcast associated personal information from an external broadcast management system through a broadcast channel. In an exemplary embodiment, the communication component 816 further includes a Near Field Communication (NFC) module to facilitate short-range communication. For example, the NFC module may be implemented based on a Radio Frequency Identification (RFID) technology, an Infrared Data Association (IrDA) technology, an Ultra-Wide Band (UWB) technology, a BlueTooth (BT) technology and another technology.

[00208] In the exemplary embodiment, the electronic device 800 may be implemented by one or more ASICs, DSPs, DSPDs, PLDs, FPGAs, controllers, micro-controllers, microprocessors or other electronic components, and is configured to execute the abovementioned method.

[00209] In the exemplary embodiment, a non-volatile computer-readable storage medium is also provided, for example, a memory 804 including a computer program instruction. The computer program instruction may be executed by a processor 820 of an electronic device 800 to implement the abovementioned method.

[00210] FIG. 5 is a block diagram of an electronic device 1900 according to an embodiment of the disclosure. For example, the electronic device 1900 may be provided as a server. Referring to Fig. 5, the electronic device 1900 includes a processing component 1922, further including one or more processors, and a memory resource represented by a memory 1932, configured to store an instruction executable for the processing component 1922, for example, an APP. The APP stored in the memory 1932 may include one or more than one module of which each corresponds to a set of instructions. In addition, the processing component 1922 is configured to execute the instruction to execute the abovementioned method.

[00211] The electronic device 1900 may further include a power component 1926 configured to execute power management of the electronic device 1900, a wired or wireless network interface 1950 configured to concatenate the electronic device 1900 to a network and an I/O interface 1958. The electronic device 1900 may be operated based on an operating system stored in the memory 1932, for example, Windows Server™, Mac OS X™, Unix™, Linux™, FreeBSD™ or the like.

[00212] In the exemplary embodiment, a non-volatile computer-readable storage medium is also provided, for example, a memory 1932 including a computer program instruction. The computer program instruction may be executed by a processing component 1922 of an electronic device 1900 to implement the abovementioned method.

[00213] The disclosure may be a system, a method and/or a computer program product. The computer program product may include a computer-readable storage medium, in which a computer-readable program instruction configured to enable a processor to implement each aspect of the disclosure is stored.

[00214] The computer-readable storage medium may be a physical device capable of retaining and storing an instruction used by an instruction execution device. For example, the computer-readable storage medium may be, but not limited to, an electric storage device, a magnetic storage device, an optical storage device, an electromagnetic storage device, a semiconductor storage device, or any appropriate combination thereof. More specific examples (non-exhaustive list) of the computer-readable storage medium include a portable computer disk, a hard disk, a RAM, a ROM, an EPROM (or a flash memory), an SRAM, a Compact Disc Read-Only Memory (CD-ROM), a Digital Video Disk (DVD), a memory stick, a floppy disk, a mechanical encoding device, a punched card or in-slot raised structure with an instruction stored therein, and any appropriate combination thereof. Herein, the computer-readable storage medium is not explained as a transient signal, for example, a radio wave or another freely propagated electromagnetic wave, an electromagnetic wave propagated through a wave guide or another transmission medium (for example, a light pulse propagated through an optical fiber cable) or an electric signal transmitted through an electric wire.

[00215] The computer-readable program instruction described here may be downloaded from the computer-readable storage medium to each computing/processing

device or downloaded to an external computer or an external storage device through a network such as the Internet, a Local Area Network (LAN), a Wide Area Network (WAN) and/or a wireless network. The network may include a copper transmission cable, optical fiber transmission, wireless transmission, a router, a firewall, a switch, a gateway computer and/or an edge server. A network adapter card or network interface in each computing/processing device receives the computer-readable program instruction from the network and forwards the computer-readable program instruction for storage in the computer-readable storage medium in each computing/processing device.

[00216] The computer program instruction configured to execute the operations of the disclosure may be an assembly instruction, an Instruction Set Architecture (ISA) instruction, a machine instruction, a machine related instruction, a microcode, a firmware instruction, state setting data or a source code or target code edited by one or any combination of more programming languages, the programming language including an object-oriented programming language such as Smalltalk and C++ and a conventional procedural programming language such as "C" language or a similar programming language. The computer-readable program instruction may be completely executed in a computer of a user or partially executed in the computer of the user, executed as an independent software package, executed partially in the computer of the user and partially in a remote computer, or executed completely in the remote server or a server. Under the condition that the remote computer is involved, the remote computer may be concatenated to the computer of the user through any type of network including an LAN or a WAN, or may be concatenated to an external computer (for example, concatenated by an Internet service provider through the Internet). In some embodiments, an electronic circuit such as a programmable logic circuit, an FPGA, or a Programmable Logic Array (PLA) may be customized by use of state personal information of a computer-readable program instruction, and the electronic circuit may execute the computer-readable program instruction, thereby implementing each aspect of the disclosure.

[00217] Herein, each aspect of the disclosure is described with reference to flowcharts and/or block diagrams of the method, device (system) and computer program product according to the embodiments of the disclosure. It is to be understood that each block in the flowcharts and/or the block diagrams and a combination of each block in the flowcharts

and/or the block diagrams may be implemented by computer-readable program instructions.

[00218] These computer-readable program instructions may be provided for a universal computer, a dedicated computer or a processor of another programmable data processing device, thereby generating a machine to further generate a device that realizes a function/action specified in one or more blocks in the flowcharts and/or the block diagrams when the instructions are executed through the computer or the processor of the other programmable data processing device. These computer-readable program instructions may also be stored in a computer-readable storage medium, and through these instructions, the computer, the programmable data processing device and/or another device may work in a specific manner, so that the computer-readable medium including the instructions includes a product including instructions for implementing each aspect of the function/action specified in one or more blocks in the flowcharts and/or the block diagrams.

[00219] These computer-readable program instructions may further be loaded to the computer, the other programmable data processing device or the other device, so that a series of operating steps are executed in the computer, the other programmable data processing device or the other device to generate a process implemented by the computer to further realize the function/action specified in one or more blocks in the flowcharts and/or the block diagrams by the instructions executed in the computer, the other programmable data processing device or the other device.

[00220] The flowcharts and block diagrams in the drawings illustrate probably implemented system architectures, functions and operations of the system, method, and computer program product according to multiple embodiments of the disclosure. On this aspect, each block in the flowcharts or the block diagrams may represent part of a module, a program segment or an instruction, and part of the module, the program segment or the instruction includes one or more executable instructions configured to realize a specified logical function. In some alternative implementations, the functions marked in the blocks may also be realized in a sequence different from those marked in the drawings. For example, two continuous blocks may actually be executed substantially concurrently and may also be executed in a reverse sequence sometimes, which is determined by the involved functions. It is further to be noted that each block in the block diagrams and/or the flowcharts and a

combination of the blocks in the block diagrams and/or the flowcharts may be implemented by a dedicated hardware-based system configured to execute a specified function or operation or may be implemented by a combination of a special hardware and a computer instruction.

[00221] Each embodiment of the disclosure has been described above. The above descriptions are exemplary, non-exhaustive and also not limited to each disclosed embodiment. Many modifications and variations are apparent to those of ordinary skill in the art without departing from the scope and spirit of each described embodiment of the disclosure. The terms used herein are selected to explain the principle and practical application of each embodiment or improvements in the technologies in the market best or enable others of ordinary skill in the art to understand each embodiment disclosed herein.

CLAIMS

1. A video processing method, comprising:

acquiring a video, at least part of video frames in the video comprising a target object;
detecting at least one type of learning behavior of the target object in a process of watching a teaching course according to the video; and

under a condition of detecting that the target object performs the at least one type of the learning behavior, generating learning state information according to a duration for the target object to perform the at least one type of the learning behavior and/or at least part of video frames comprising the at least one type of the learning behavior.

2. The method of claim 1, wherein the learning behavior comprises at least one type of: making at least one target gesture, expressing a target emotion, paying attention to a display region of the teaching course, performing at least one interactive behavior with another object, not appearing in the at least part of the video frames in the video, closing the eyes, or eye contact in the display region of the teaching course.

3. The method of claim 1 or 2, wherein detecting the at least one type of the learning behavior of the target object according to the video comprises:

performing target object detection on the video to obtain the video frame comprising the target object; and

detecting the at least one type of the learning behavior in the video frames comprising the target object.

4. The method of claim 3, wherein the learning behavior comprises making the at least one target gesture, and

detecting the at least one type of the learning behavior in the video frames comprising the target object comprises:

detecting the at least one target gesture in the video frames comprising the target object;

under a condition of detecting that a number of continuous video frames comprising the at least one target gesture exceeds a first threshold, recording at least one frame of the

video frames comprising the target gesture as a gesture starting frame;

under a condition that a number of continuous video frames, not comprising the target gesture in video frames after the gesture starting frame, exceeds a second threshold, recording at least one frame of the video frames not comprising the target gesture as a gesture ending frame; and

determining a number and/or a time that the target object makes the at least one target gesture in the video according to a number of gesture starting frames and a number of gesture ending frames.

5. The method of claim 3 or 4, wherein the learning behavior comprises expressing the target emotion, and

detecting the at least one type of the learning behavior in the video frames comprising the target object comprises:

performing a expression detection and/or a smile value detection on the video frames comprising the target object;

under a condition of detecting that the target object in a video frame shows at least one first target expression or detecting that a result of the smile value detection exceeds a target smile value, determining the detected video frame as a first detection frame; and

under a condition of detecting that a number of continuous first detection frames exceeds a third threshold, determining that the target object has the target emotion.

6. The method of any one of claims 3-5, wherein the learning behavior comprises paying attention to the display region of the teaching course, and

detecting the at least one type of the learning behavior in the video frames comprising the target object comprises:

performing a expression detection and a headpose detection on the video frames comprising the target object;

under a condition of detecting that the target object in a video frame shows at least one second target expression or a headpose is within a target headpose range, determining the detected video frame as a second detection frame; and

under a condition of detecting that a number of continuous second detection frames exceeds a fourth threshold, determining that the target object pays attention to the display region of the teaching course.

7. The method of any one of claims 1-6, wherein generating the learning state information according to the at least part of the video frames comprising the at least one type of the learning behavior comprises:

acquiring the video frames comprising the at least one type of the learning behavior in the video as a target video frame set;

performing a face quality detection on at least one video frame of the target video frame set, and determining a video frame where a face quality is greater than a face quality threshold as a target video frame; and

generating the learning state information according to the target video frame.

8. The method of claim 7, wherein generating the learning state information according to the target video frame comprises:

determining at least one frame of the target video frame as the learning state information; and/or

recognizing a region where the target object is located in at least one frame of the target video frame, and generating the learning state information based on the region where the target object is located.

9. The method of claim 1 or 2, wherein detecting the at least one type of the learning behavior of the target object according to the video comprises:

performing a target object detection on the video to obtain the video frames comprising the target object, and determining video frames except the video frames comprising the target object in the video as video frames where the target object is not detected; and

under a condition that a number of the video frames where the target object is not detected exceeds a preset video frame number, detecting that the learning behavior comprises not appearing in the at least part of the video frames in the video.

10. The method of any one of claims 1-9, wherein generating the learning state information according to the duration for the target object to perform the at least one type of the learning behavior comprises:

under a condition of detecting that a time when the target object performs the at least

one type of the learning behavior is not less than a time threshold, recording the duration of the at least one type of the learning behavior; and

determining the duration corresponding to the at least one type of the learning behavior as the learning state information.

11. The method of any one of claims 1-10, further comprising:

rendering background region in the at least part of the video frames in the video, the background region being a region except the target object in the video frames.

12. The method of any one of claims 1-11, further comprising:

counting learning state information of at least one target object to obtain a statistical result of the at least one target object; and

generating learning state statistics data according to the statistical result of the at least one target object.

13. The method of claim 12, wherein generating the learning state statistics data according to the statistical result of the at least one target object comprises:

acquiring a statistical result of the target object of at least one classification according to a classification to which the at least one target object belongs, and generating learning state statistics data of the at least one classification, the classification to which the target object belongs comprising at least one of: a course that the target object attends, an institution that the target object is registered to, or a device used by the target object; and/or,

visualizing the statistical result of the at least one target object to generate the learning state statistics data of the at least one target object.

14. A video processing apparatus, comprising:

a video acquisition module, configured to acquire a video, at least part of video frames in the video comprising a target object;

a detection module, configured to detect at least one type of learning behavior of the target object in a process of watching a teaching course according to the video; and

a generation module, configured to, under a condition of detecting that the target object performs the at least one type of the learning behavior, generate learning state information according to a duration for the target object to perform the at least one type of the

learning behavior and/or the at least part of video frames comprising the at least one type of the learning behavior.

15. An electronic device, comprising:

a processor; and

a memory, configured to store an instruction executable for the processor.

wherein the processor is configured to call the instruction stored in the memory to execute the method of any one of claims 1-12.

16. A computer-readable storage medium having stored therein a computer program instruction that when executed by a processor, implements the method of any one of claims 1-13.

17. A computer program, comprising a computer-readable code, when the computer-readable code runs in an electronic device, a processor in the electronic device is configured to execute the method of any one of claims 1-13.