

DATA PROCESSING METHOD AND APPARATUS, ELECTRONIC DEVICE AND STORAGE MEDIUM

CROSS-REFERENCE TO RELATED APPLICATION

[0001] This application claims priority to Chinese Patent Application No. 202010339381.1, filed with the Chinese Patent Office on April 26, 2020 and entitled “DATA PROCESSING METHOD AND APPARATUS, ELECTRONIC DEVICE AND STORAGE MEDIA”, the content of which is incorporated by reference in its entirety.

TECHNICAL FIELD

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

BACKGROUND

[0003] Evaluation of behavior state of a target object can be widely applied in various fields, and the obtained evaluation results can be used for analyzing the target object or behaviors of the target object. The more accurate the evaluation results are, the more real and meaningful the corresponding analysis will be.

[0004] Therefore, how to conveniently obtain more accurate behavior state evaluation results becomes an urgent problem to be solved at present.

SUMMARY

[0005] The present disclosure proposes a solution for data processing.

[0006] According to one aspect of the present disclosure, there is provided a data processing method, which includes the following operations.

[0007] Multimedia data of a target object is acquired; behavior state detection is performed on the target object in at least one detection dimension according to the multimedia data to obtain intermediate detection results of the target object in the at least one detection dimension; the intermediate detection results in the at least one detection dimension are processed to obtain a target detection result of the target object. The target detection result is used to indicate a behavior state of the target object.

[0008] According to one aspect of the present disclosure, there is provided a data processing apparatus, which includes an acquiring module, a detecting module and a processing module.

[0009] The acquiring module is configured to acquire multimedia data of a target object. The detecting module is configured to perform behavior state detection on the target object in at least one detection dimension according to the multimedia data to obtain intermediate detection results of the target object in the at least one detection dimension. The processing module is configured to process the intermediate detection results in the at least one detection dimension to obtain a target detection result of the target object. The target detection result is used to indicate a behavior state of the target object.

[0010] According to one aspect of the present disclosure, there is provided an electronic device, which includes one or more processors and a memory for storing instructions executable by the one or more processors. The one or more processors are configured to implement the data processing method as described above.

[0011] According to one aspect of the present disclosure, there is provided a computer readable storage medium, having stored thereon computer program instructions that, when being executed by a processor, cause to implement the data processing method as described above.

[0012] According to one aspect of the present disclosure, there is provided a computer program including computer readable codes which, when run in an electronic device, cause a processor in the electronic device to implement the data processing method as described above.

[0013] In the embodiments of the present disclosure, multimedia data of a target object is acquired, behavior state detection is performed on the target object in at least one detection dimension according to the multimedia data to obtain intermediate detection results of the target object in the at least one detection dimension, and the intermediate detection results in the at least one dimension are processed to obtain a target detection result, where the target detection result is used for indicating a behavior state of the target object. By the above process, the target detection result indicating the behavior state of the target object can be obtained based on performing the behavior state detection on the target object in the multiple detection dimensions. On one hand, automatic evaluation of the behavior state of the target object can be realized, and on the other hand, the comprehensiveness and accuracy of the final target detection result can be improved.

[0014] It is to be understood that the above general description and the following detailed description are only exemplary and explanatory, and are not intended to limit the disclosure. Other features and aspects of the present disclosure will become apparent according to the following detailed description of exemplary embodiments with reference to the accompanying drawings.

BRIEF DESCRIPTION OF THE DRAWINGS

[0015] The accompanying drawings, which are incorporated in and constitute a part of the description, illustrate embodiments consistent with the present disclosure and, together with the description, serve to illustrate the technical solution of the present disclosure.

[0016] FIG. 1 illustrates a flowchart of a data processing method according to an embodiment of the present disclosure.

[0017] FIG. 2 illustrates a block diagram of a data processing apparatus according to an embodiment of the present disclosure.

[0018] FIG. 3 illustrates a schematic diagram of a target detection result according to an application example of the present disclosure.

[0019] FIG. 4 illustrates a block diagram of an electronic device according to an embodiment of the present disclosure.

[0020] FIG. 5 illustrates another block diagram of an electronic device according to an embodiment of the present disclosure.

DETAILED DESCRIPTION

[0021] Various exemplary embodiments, features, and aspects of the present disclosure will be described in detail below with reference to the accompanying drawings. The same reference numerals in the accompanying drawings indicate components with the same or similar functions. Although various aspects of the embodiments are shown in the accompanying drawings, the accompanying drawings are not necessarily drawn to scale unless otherwise indicated.

[0022] The word “exemplary” as used herein means “serving as an example, embodiment, or illustration”. Any embodiment described herein as “exemplary” is not necessarily to be construed as superior to, or better than other embodiments.

[0023] The term “and/or”, as used herein, is merely an association that describes an associated object, meaning that there may be three relationships. For example, A and/or B, which may mean that independent existence of A, existence of both A and B, and independent existence of B. Additionally, the term “at least one” as used herein denotes

any combination of any one or at least two of a plurality. For example, including at least one of A, B or C, may denote the inclusion of any one or more elements selected from the group consisting of A, B, and C.

[0024] In addition, numerous specific details are set forth in the following detailed description in order to better illustrate the disclosure. It will be understood by those skilled in the art that the present disclosure may be performed without certain specific details. In some embodiments, methods, means, elements, and circuits well known to those skilled in the art have not been described in detail in order to highlight the principle of the disclosure.

[0025] FIG. 1 illustrates a flowchart of a data processing method according to an embodiment of the present disclosure, which may be applied to a data processing apparatus. The data processing apparatus may be a terminal device, a server, or other processing device, etc. The terminal device may be User Equipment (UE), a mobile device, a user terminal, a terminal, a cellular telephone, a cordless telephone, a personal digital assistant (PDA), a handheld device, a computing device, an in-vehicle device, a wearable device, or the like. 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 may be flexibly selected according to actual conditions.

[0026] In some possible implementations, the data processing method may also be implemented by a processor invoking computer readable instructions stored in a memory.

[0027] As shown in FIG. 1, in a possible implementation, the data processing method may include operations S11, S12 and S13.

[0028] In S11, multimedia data of a target object is acquired.

[0029] In S12, behavior state detection is performed on the target object in at least one detection dimension according to the multimedia data to obtain intermediate detection

results of the target object in the at least one detection dimension.

[0030] In S13, the intermediate detection results in the at least one detection dimension are processed to obtain a target detection result of the target object. The target detection result is used to indicate a behavior state of the target object.

[0031] The target object may be any object having a behavior state representation or evaluation requirement, and a specific implementation thereof may be flexibly determined according to an application scenario of an execution behavior performed by the target object. The specific behavior performed by the target object is not limited in the disclosed embodiment. In a possible implementation, the behavior may be a teaching behavior, a management behavior, a work behavior, or the like. Accordingly, the implementation of the target object may change along with different behaviors performed by the target object. In a possible implementation, in a case that the behavior is a teaching behavior, the target object may be a teacher. Further, the teaching behavior may also be a formal teaching behavior or a simulated teaching behavior, and accordingly, the target object may be a formal teaching teacher, a simulated teaching teacher, or a teacher who is not employed and at an interview stage, or the like. In a possible implementation, in a case that the behavior is a management behavior, the target object may be an object having management functions, such as a teaching manager or the like. In a possible implementation, in a case that the behavior is a work behavior, the target object may be an associated work object, such as an educator or the like. The subsequent embodiments of the present disclosure are described by taking the target object as a teacher and taking the execution behavior as a simulated teaching behavior (hereinafter referred to as the simulated teaching behavior) as an example. For other implementations of the target object and the behavior thereof, corresponding extensions may be made with reference to the subsequent embodiments of the present disclosure, and details are not described again.

[0032] The multimedia data of the target object may be data acquired by the target object when the corresponding behavior is executed, and the implementation thereof may be

flexibly determined according to actual conditions. In a possible implementation, the multimedia data of the target object may include video data and/or audio data. How to acquire the multimedia data of the target object specifically can be flexibly determined according to the actual conditions, which are detailed in the subsequent embodiments of the present disclosure and will not be expanded herein.

[0033] After the multimedia data of the target object is acquired, through S12, the behavior state detection is performed on the target object in at least one detection dimension to obtain intermediate detection results in the at least one detection dimension. In a possible implementation, in a case that the target object is a teacher and the execution behavior is a simulated teaching behavior, the state detection may be performed on the teaching behaviors of the target object in each detection dimension in the teaching behavior, such as gesture, emotion, eye contact, fluency, speech rate, pause or volume in the teaching process. The specific dimensions included and specific implementations of performing the behavior state detection in these dimensions can be detailed in the subsequent embodiments of the present disclosure, and will not be expanded herein.

[0034] After the intermediate detection results of the target object in the at least one detection dimension are obtained, through S13, the intermediate detection results in the at least one detection dimension can be processed to obtain the target detection result of the target object. The number of target detection results is not limited in the embodiments of the present disclosure, and can be flexibly set according to actual requirements. In a possible implementation, the target detection result may include an overall detection result for reflecting the overall situation of the behavior state of the target object. In a possible implementation, the target detection result may include an overall detection result and a plurality of detailed subdivision results for reflecting the overall situation and detailed situations of the behavior state of the target object. The specific implementation of the target detection result and a manner of obtaining the target detection result can be referred to the subsequent embodiments of the present disclosure, and will not be expanded herein.

[0035] In the embodiments of the present disclosure, multimedia data of a target object is acquired, behavior state detection is performed on the target object in at least one detection dimension according to the multimedia data to obtain intermediate detection results of the target object in the at least one detection dimension, and the intermediate detection results in the at least one dimension are processed to obtain a target detection result for indicating a behavior state of the target object. By the above process, the target detection result indicating the behavior state of the target object can be obtained based on performing the behavior state detection on the target object in the multiple dimensions. On one hand, automatic evaluation of the behavior state of the target object can be realized, and on the other hand, the comprehensiveness and accuracy of the final target detection result can be improved.

[0036] As described in the above embodiments of the present disclosure, the implementation of the multimedia data is not limited. In a possible implementation, the multimedia data may contain only audio data. In a possible implementation, the multimedia data may contain only video data, such as silent video, and the like. In a possible implementation, the multimedia data may contain both video data and audio data, such as audio video or the like. In an example, in a case that the multimedia data includes the video data, the resolution of the video data is not limited, and may be flexibly selected according to actual conditions, such as being 640P, 720P, and 1080P, etc. In an example, in a case that the multimedia data includes the audio data, the audio sampling frequency of the audio data is also not limited, and may be flexibly selected, such as being 8000Hz or 16000Hz, etc.

[0037] With different forms of the multimedia data, the generation manner of the multimedia data can also change flexibly. In a possible implementation, in a case that the target object is a teacher, and the execution behavior is a simulated teaching behavior, the audio data may be generated by recording the audio of a teacher-simulated-teaching process, and the video data may be generated by shooting the actions in the teacher in the teacher-simulated-teaching process. Therefore, in an example, the multimedia data may be generated by shooting the video of the teacher-simulated-teaching process.

[0038] In a possible implementation, the multimedia data may be acquired by the target object that performs teaching operations according to preset text data. The preset text data includes at least one instruction marker, and the instruction marker is used for segmenting and/or annotating at least part of contents in the preset text data.

[0039] The preset text data may be text content used by a teacher for teaching or simulated teaching, such as a verbatim manuscript of a simulated teaching, which contains related content to be taught by the teacher in the simulated teaching process. The instruction marker may be a marker located within the preset text data and used for segmenting or annotating part of contents in the preset text data. The position, specific contents, and functions of the instruction marker can be flexibly selected according to actual conditions, and are not limited to the following embodiments of the present disclosure.

[0040] In a possible implementation, in a case that the preset text data is the verbatim manuscript of the simulated teaching, the instruction marker may be a marker used for performing segmentation on the simulated teaching process to which the verbatim manuscript belongs, that is, the instruction marker may be a partial structural marker for the verbatim manuscript of the simulated teaching. The specific implementation of the instruction marker may be flexibly determined according to the segmenting situation of the simulated teaching process. In a possible implementation, the simulated teaching process may be segmented into the following stages: pre-class warm-up, knowledge teaching, in-class training, classroom test, etc., and the content of the verbatim manuscript of the simulated teaching may be segmented into these four stages by using instruction marker.

[0041] Further, how to segment the simulated teaching process into multiple stages through the instruction markers may also be flexibly selected. In a possible implementation, the corresponding stages may be annotated by using <start of start instruction> and <end of start instruction>, <start of end instruction> and <end of end instruction> etc. to realize structure segmenting of the verbatim manuscript of the simulated teaching. Specifically, the specific implementations of the annotations such as <start of start instruction>, <start

of end instruction> may be flexibly determined according to actual conditions. For example, some specific words or action descriptions may be used as the specific implementations of the annotations such as <start of start instruction> or <start of end instruction>.

[0042] Exemplarily, the content of the verbatim manuscript of the simulated teaching may be as follows, “<start of pre-class warm-up start instruction>, reading that “we perform the pre-class warm-up next”; <end of pre-class warm-up start instruction>; then there is a large section of course content. <start of pre-class warm-up end instruction>, reading that “ok, let's go to the next link”; <end of pre-class warm-up end instruction>; then there is a large section of course content.

[0043] <knowledge teaching link start instruction>, reading that “we perform the knowledge teaching next link”; <end of knowledge teaching link start instruction>.

[0044] <start of knowledge teaching link end instruction>, reading that “ok, let's go to the next link”; <end of knowledge teaching end instruction>”. As can be seen from the exemplary verbatim manuscript of the simulated teaching, in an example, the text content of the pre-class warm-up link may be annotated in the verbatim manuscript of the simulated teaching by using the markers such as <start of pre-class warm-up start instruction>, <end of pre-class warm-up start instruction>, <start of pre-class warm-up end instruction>, and <end of pre-class warm-up end instruction>, etc. Similarly, the text content of the knowledge teaching link may be further segmented from the verbatim manuscript of the simulated teaching by corresponding instruction markers. The specific words or action descriptions corresponding to these instruction markers are not limited in the embodiments of the present disclosure, and can be flexibly selected according to actual requirements.

[0045] The teacher performs simulated teaching according to the preset text data including instruction markers for segmenting structure to obtain corresponding multimedia data, so that the multimedia data can be marked with markers of different stages (such as

specific words or actions) or the like. These markers of different stages may be identified by the data processing apparatus automatically, and thus the data processing apparatus can segment the multimedia data automatically according to corresponding structures. In a possible implementation, the segmented multimedia data may be used to obtain the target detection results for each stage of the simulated teaching respectively, that is, the target detection results for each stage of the simulated teaching may be obtained respectively. In this way, the automation degree of the data processing process can be improved, and the pertinence and practicability of the target detection results obtained by data processing can be improved.

[0046] In a possible implementation, in a case that the preset text data is the verbatim manuscript of the simulated teaching, the instruction markers may be used to mark key contents of the verbatim manuscript or mark the locations where interactions are desired in the verbatim manuscript, that is, the instruction marker may be used to annotate the knowledge points and the interactions in the verbatim manuscript of the simulated teaching. The specific implementation of the instruction marker can be flexibly determined according to the locations where the key contents and the situation of the desired interactions. For example, the number of the instruction markers may be multiple when there are multiple key paragraphs or interaction positions.

[0047] Further, how to annotate the key contents and the interaction positions in the verbatim manuscript of the simulated teaching through the instruction markers also may be selected flexibly. In a possible implementation, the key knowledge points in the simulated teaching may be annotated by using <start of key point>, <end of key point> and the like. In a possible implementation, in a case that interactions are required during the simulated teaching, an identifier of <interaction needed> may be used as an instruction marker in the corresponding position in the verbatim manuscript of the simulated teaching. Specifically, the specific implementations of annotations such as <start of key point> and <interaction needed> may be determined flexibly according to actual conditions. For example, the specific words or action descriptions may be used as the specific implementation of the annotations such as <start of key point>, <interaction

needed> and the like.

[0048] Exemplarily, the content of the verbatim manuscript of the simulated teaching may also be as follows, “This is a part of teaching content, <start of key point>, reading that “let me take a look at the pictures, class; they are intersections; using your protractor, and let us see whether you have made any new discoveries”, <interaction needed>, reading that “the student in blue, please give your answer”.

[0049] <interaction needed>, reading that “you are right; class, if two straight lines intersect at right angles, we can say that they are perpendicular to each other”, <end of key point>”. As can be seen from the above exemplary verbatim manuscript of the simulated teaching, in an example, the key points can be segmented from the verbatim manuscript of the simulated teaching by using the markers such as <start of key point> and <end of key point>. Similarly, it is also possible to make use of <interaction needed> or the like to prompt the teacher of the simulated teaching to make corresponding actions at some stages during teaching. The specific words or action descriptions corresponding to these instruction markers are not limited in the embodiments of the present disclosure, and can be selected flexibly according to actual requirements.

[0050] The teacher performs simulated teaching according to the preset text data including instruction markers for annotating the key points and interactions to obtain corresponding multimedia data, so that the multimedia data can be marked at some important stages of the simulated teaching (for example, a stage of teaching the key points or a stage where interaction is needed). These markers may be identified by the data processing apparatus automatically, and thus the data processing apparatus may identify the teaching process of the key points or the interaction process in the multimedia data automatically. In a possible implementation, it is possible to focus on the intermediate detection results of the teaching process of the key points or the interaction process, so as to evaluate the simulated teaching state of the teacher more pertinently. In this way, the automation degree of the data processing process can be improved, and the pertinence and practicability of the target detection results obtained

by data processing can be improved.

[0051] As can be seen from the above embodiments of the present disclosure, the multimedia data, generated by the preset text data including the instruction markers, can be automatically recognized and processed, thereby improving the automation degree of the data processing method and improving the pertinence and practicability of the final target detection result.

[0052] As can be seen from the above embodiments of the present disclosure, the implementation and generation of the multimedia data may have multiple implementations. Accordingly, the manner of acquiring the multimedia data, i.e., S11, may be changed flexibly depending on the implementations of the multimedia data. In a possible implementation, the multimedia data may be pre-recorded multimedia data. In this case, the multimedia data may be acquired according to a storage location of the multimedia data, such as a Uniform Resource Locator (URL) link. In a possible implementation, the multimedia data may be data in a recording process, such as a live video. In this case, the multimedia data may be acquired according to a live link or live address of the multimedia data or the like.

[0053] Further, as described in the above embodiments of the present disclosure, the multimedia data may include video data and/or audio data. Therefore, the manner of acquiring the multimedia data may change flexibly depending on the specific contents of the multimedia data. In a possible implementation, when the multimedia data includes both video data and audio data and the audio data and the video data are integrated, the integrated audio-video data may be directly acquired, and then the video data and the audio data may be separated from the integrated audio-video data in a certain manner. The specific separation manner is not limited in the embodiments of the present disclosure, and may be selected flexibly according to actual conditions. In a possible implementation, the video data and the audio data may be acquired separately for subsequent detection, in a case that the multimedia data includes both the video data and the audio data and the audio data and the video data are independent of each other.

[0054] As described in the above embodiments of the present disclosure, the multimedia data may include multiple stages of the simulated teaching process, such as pre-class warm-up, knowledge teaching, in-class training, and classroom test, etc., and the stages may be identified based on words or actions corresponding to specific instruction markers. Therefore, in a possible implementation, after the multimedia data is acquired, the multimedia data may also be segmented according to the specific words or actions in the multimedia data so as to acquire the desired part of the multimedia data. For example, in an example, a part of multimedia data in the pre-class warm-up stage may be acquired from the multimedia data by identifying a pre-class warm-up start instruction and a pre-class warm-up end instruction in the multimedia data, and a subsequent target detection result may be obtained based on the part of multimedia data in the pre-class warm-up stage. In an example, each part of the multimedia data may also be acquired by identifying the start instructions and end instructions for the multiple stages in the multimedia data, and thus target detection results of each part of the multimedia data may be obtained by operations S12 and S13.

[0055] In a possible implementation, the multimedia data at different stages may also be acquired separately based on the recording time of each stage in the simulated teaching process. In an example, a teacher may record multimedia data through a client, and the implementation of the client is not limited in the embodiments of the present disclosure, which may be a mobile phone, a computer, or other user equipment. During recording, the client may display different stages of the simulated teaching on the client interface in the form of taps. Then, the teacher may enter a stage by clicking a tap, and record multimedia data of the simulated teaching at the stage. In this case, the multimedia data of the simulated teaching at the stage includes a recorded time stamp in addition to video data and audio data. Therefore, in the process of acquiring the multimedia data at S11, the data processing apparatus may determine a simulated teaching stage corresponding to the multimedia data by the time stamp included in the multimedia data, thereby obtaining each part of the multimedia data. In the present disclosure, for ease of description, subsequent embodiments of the present disclosure will illustrate the process

of data processing exemplarily by not segmenting the stages of the multimedia data. An implementation in which the multimedia data is segmented into a plurality of parts to obtain the target detection results of the parts respectively may be extended with reference to the subsequent embodiments of the present disclosure, and details are not described again.

[0056] In addition to the above content, in S11, the number of the acquired multimedia data of the target object is not limited, and may be multimedia data corresponding to one target object or multiple target objects. That is, in the data processing method of the embodiments of the present disclosure, the multimedia data of only one target object may be processed at a time, and the multimedia data of multiple target objects may be processed in batches at the same time. In a case that there are multiple pieces of multimedia data, in order to distinguish the multimedia data of different target objects, it may be added to the multimedia data other information for determining the target object to which the multimedia data belongs. Therefore, in a possible implementation, the multimedia data may include, in addition to the video data and the audio data, identity information such as a teacherID, a simulated teaching ID, and an ID of a group to which the teacher belongs (such as a vendeeID of a company or a school to which the teacher belongs). In addition, the multimedia data may include other related information, such as a multimedia data address (URL link), a structure of the multimedia data (such as simulated teaching stages corresponding to the multimedia data as mentioned in the above embodiments, and a start time stamp or an end time stamp of each stage), or related information of the multimedia detection (such as a video detection frame rate). Subsequent embodiments of the present disclosure are described by taking the acquired multimedia data as multimedia data of one target object as an example, and the process of acquiring the multimedia data of multiple target objects for data processing may be extended with reference to the subsequent embodiments, and details are not described again.

[0057] It should be noted that the implementations of the multimedia data and the manners of acquiring the multimedia data mentioned in each of the above-mentioned

embodiments of the present disclosure can be flexibly combined according to requirements, and are not limited in the disclosed embodiments.

[0058] After the multimedia data of the target object is acquired by any of the above embodiments of the present disclosure, behavior state detection may be performed on the target object in at least one detection dimension to obtain intermediate detection results of the target object in the at least one detection dimension by the operation S12. The implementation of S12 is not limited, and may be flexibly selected according to the actual conditions of the multimedia data, and is not limited to the following disclosed embodiments.

[0059] As described in the above embodiments of the present disclosure, in a possible implementation, the multimedia data may include the video data. In this case, S12 may include that the behavior state detection is performed on the target object based on the video data. Therefore, in a possible implementation, S12 may include operations S1211 and S1212.

[0060] In S1211, the target object is determined in the video data.

[0061] In S1212, behavior state detection is performed on the target object in at least one of the following detection dimensions: gesture, emotion or eye contact, to obtain the intermediate detection results of the target object in the at least one detection dimension.

[0062] The manner of determining the target object in S1211 is not limited, and can be flexibly determined according to the actual implementation of the target object. As described in the above embodiments of the present disclosure, in a possible implementation, the target object may be a teacher, and the execution behavior may be a simulated teaching behavior. In this case, the determination of the target object may be realized by determining the teacher who is performing teaching from the video data by means of face detection or face tracking. For example, in some possible implementations, the target object may be determined in the video data by invoking a mode of face detection and face tracking in the human action SDK.

[0063] After the target object is determined, the behavior state detection may be performed on the target object in at least one of the following detection dimensions: gesture, emotion or eye contact by the operation S1212. The specific detection dimensions and the detection order among the detection dimensions can be flexibly selected according to actual conditions. How to perform detection in each detection dimension is detailed with reference to the following embodiments of the disclosure and will not be expanded herein. In the present disclosure, each subsequent embodiment is described for example by simultaneously performing behavior state detection on video data in three detection dimensions of gesture, emotion and eye contact. Other implementations may be flexibly extended with reference to each subsequent embodiment of the present disclosure, and details are not described again.

[0064] As can be seen from the above embodiments of the present disclosure, in a possible implementation, when the multimedia data includes video data, the video data may be processed visually by a computer, so that the behavior state detection can be performed on a target object in the video data in a plurality of detection dimensions such as gesture, emotion, and eye contact. By the above process, the video data in the multimedia data can be sufficiently and effectively utilized to realize the detection of the target object in a plurality of different detection dimensions, thereby improving the diversity of the intermediate detection results, and further improving the comprehensiveness and reliability of the final target detection results.

[0065] The specific manner of obtaining corresponding intermediate detection results on each detection dimension can be flexibly determined. In a possible implementation, when the detection dimension includes a gesture detection dimension, an intermediate detection result of the target object in the gesture detection dimension may be obtained according to the following operations.

[0066] The number of times that the target object performs at least one target gesture in gesture detection periods is acquired according to the video data to obtain gesture detection results in the gesture detection periods. The target gesture includes at least one

of: holding a hand, raising a hand, or thumbs up.

[0067] The intermediate detection result of the target object in the gesture detection dimension is obtained according to the gesture detection result in at least one gesture detection period.

[0068] The video data may be divided into a plurality of video segments in order of time, and each video segment may be denoted as a gesture detection period. The division manner and the length of each gesture detection period after division are not limited. In a possible implementation, the video data may be divided according to the same time duration, and in this case, the time durations of the gesture detection periods corresponding to the different video segments are the same in length of time. In a possible implementation, the video data may also be randomly divided according to different time durations, and in this case, the time durations of the gesture detection periods corresponding to different segments of video are different in length of time. The embodiments of the present disclosure are described by using an example in which the time durations of the gesture detection periods are the same. In an example, the gesture detection period may be denoted as one minute, that is, the number of times that the target object performs the at least one target gesture in each minute are acquired respectively, to obtain a gesture detection result of the target object in each minute, and then the intermediate detection result of the target object in the gesture detection dimension is obtained in the complete video data according to the gesture detection result in each minute.

[0069] The target gesture may be set, and the gestures of the teacher during the simulated teaching, such as holding a hand (indicating that a student is asked to answer a question), raising a hand (indicating that the student is prompted to answer a question), or thumbs up (indicating that the behavior of the student is praised) may be determined as valid gestures. Which gestures can be used as the target gesture may be flexibly set according to actual conditions.

[0070] The gesture detection result in the gesture detection period is obtained by acquiring the number of times that the target object performs the at least one target gesture in the gesture detection period according to the video data, and then the intermediate detection result of the target object in the gesture detection dimension is obtained according to the gesture detection result in the at least one gesture detection period. Through the above process, the video data may be divided into a plurality of gesture detection periods, and a complete gesture detection process may be converted into multi-stage gesture detection processes, so that the difficulty of performing gesture detection each time is reduced and the efficiency of gesture detection is improved, and meantime, the intermediate detection results in the gesture detection dimension can effectively reflect the limbs movement degree of the teacher and the gesture interaction degree between the teacher and the students in the simulated teaching. Therefore, the accuracy and reliability of the data processing result can be improved while the efficiency of the data processing method is improved.

[0071] Specifically, how to acquire the number of times that the target object performs the at least one target gesture in the gesture detection period can be flexibly determined according to actual conditions. In a possible implementation, the operation that the number of times that the target object performs the at least one target gesture in the gesture detection period is acquired according to video data includes the following operations.

[0072] At least one gesture detection frame sequence of the video data in the gesture detection period is acquired. In a case that the number of frames containing a target gesture for the gesture detection frame sequence exceeds a first threshold, at least one frame in the gesture detection frame sequence is recorded as a gesture start frame. In a case that the number of frames not containing the target gesture for a gesture detection frame sequence after the gesture start frame exceeds a second threshold, at least one frame that is located in the gesture detection frame sequence after the gesture start frame is recorded as a gesture end frame. The number of times that the target object performs the at least one target gesture in the gesture detection period is acquired according to the

number of gesture start frames and the number of gesture end frames.

[0073] In a possible implementation, in the process of performing gesture detection on the video data, the gesture detection may be performed on the video data in each gesture detection period by a fixed detection frame rate. The value of the detection frame rate may be flexibly set according to actual conditions. In an example, the detection frame rate may be set to 10FPS, that is, gesture detection may be performed on 10 frames of video data per second.

[0074] If the detection frame rate is fixed, the number of frames to be detected in a gesture detection period is actually fixed, that is, the video data in the gesture detection period may correspond to a complete frame sequence, and the number of frames included in the frame sequence may be determined by the product of the time length of the gesture detection period and the detection frame rate. In a possible implementation, the gesture detection may be performed on a complete frame sequence corresponding to the gesture detection period directly, for example, the number of times that the target gesture is performed in the gesture detection period may be determined by the number of frames of the target gesture that are included in the complete frame sequence. In a possible implementation, also as described in the above embodiments of the present disclosure, at least one gesture detection frame sequence may be acquired from the complete frame sequence corresponding to the gesture detection period, and the number of target gestures in the gesture detection period may be determined according to the detection result of each gesture detection frame sequence.

[0075] The gesture detection frame sequence may be a plurality of frame sequences selected from the complete frame sequence corresponding to the gesture detection period. The specific selection manner may be selected flexibly and is not limited to the following embodiments of the present disclosure. In an example, a plurality of gesture detection frame sequences may be obtained by sliding of frame, which is specifically realized as follows. A length of each gesture detection frame sequence is set to X, and in a complete frame sequence corresponding to a gesture detection period, a first frame is used as a

start frame of a first gesture detection frame sequence and an X-th frame is used as an end frame of the first gesture detection frame sequence, to obtain the first gesture detection frame sequence. Then, in the complete frame sequence, the first gesture detection frame sequence is slid backward one frame to obtain a second gesture detection frame sequence, that is, the second frame in the complete frame sequence is used as a start frame of the second gesture detection frame sequence and an X+1-th frame is used as an end frame of the second gesture detection frame sequence. to obtain the second gesture detection frame sequence. By such analogy, multiple gesture detection frame sequences are obtained. The value of X may be selected flexibly according to the actual conditions, and is not limited in the embodiments of the disclosure. In an example, X may be consistent with the detection frame rate, that is, X may be set to 10 frames in a case that the detection frame rate is 10FPS.

[0076] After a plurality of gesture detection frame sequences is acquired, gesture start frames and gesture end frames may be obtained based on the plurality of gesture detection frame sequences. As described in the above embodiments of the disclosure, in a possible implementation, gesture detection is performed on each frame in each gesture detection frame sequence to determine the number of frames containing the target gesture. If the number of frames containing the target gesture exceeds the first threshold, it may be considered that the target gesture exists in the current gesture detection frame sequence, and at this time at least one frame may be selected from the current gesture detection frame sequence as the gesture start frame.

[0077] The manner of performing gesture detection on each frame may be flexibly selected according to the actual conditions. In a possible implementation, the gesture detection on the frame picture may be implemented by neural network having a gesture detection function. The implementation of the neural network having the gesture detection function may also be flexibly determined. As described in the embodiments of the disclosure as described above, the target gesture may include a plurality of gestures. In a possible implementation, the gesture detection may be performed on each frame of image by using a neural network that can simultaneously recognize a plurality of target

gestures. In a possible implementation, the gesture detection may also be performed for each type of the gestures by using a corresponding neural network. Specifically, in an example, the target gesture of raising a hand of the target object may be detected by invoking the human body detection and hand raising detection model in the insight SDK. In an example, other target gestures of the target object may be detected by invoking the gesture detection model in the human action SDK. In a possible implementation, if there is a plurality types of target gestures detected in the gesture detection frame sequence, it may be determined whether the number of frames of each target gesture exceeds the first threshold. If the number of frames of all the target gestures exceeds the first threshold, it may be indicated that there are a plurality of target gestures in the current gesture detection frame sequence. If the number of frames of some types of target gestures exceeds the first threshold, it may be indicated that there are target gestures of some types which exceeds the first threshold in the current gesture detection frame sequence.

[0078] The value of the first threshold may be flexibly set according to the actual conditions, and is not limited to the embodiments of the present disclosure. In an example, in a case that a gesture detection frame sequence includes 10 frames, the first threshold may be set to 6.

[0079] In the case that the number of frames containing the target gesture exceeds the first threshold, at least one frame may be selected from the current gesture frame sequence as a gesture start frame. Specifically, which frame is selected as the gesture start frame may be determined flexibly. In a possible implementation, the N-th frame including the target gesture in the gesture detection frame sequence may be used as the gesture start frame, and the time corresponding to the gesture start frame may be recorded as the start time of gesture interaction. The value of N can be flexibly selected. In an example, N can be consistent with the value of the first threshold. For example, in a case that the gesture detection frame sequence includes 10 frames and the first threshold is set to 6, if it is detected in the current gesture detection frame sequence that the number of frames including the target gesture is not less than 6, the sixth frame including the target gesture in the current gesture detection frame sequence may be recorded as the gesture start

frame, and the time of the gesture start frame in the video data is recorded as the start time of gesture.

[0080] After the gesture start frame is determined, the end time of gesture, i.e., the gesture end frame, may be further determined. The determination manner of the gesture end frame is similar to that of the gesture start frame. As described in the above embodiments of the disclosure, in a possible implementation, the gesture detection may be performed separately in a gesture detection frame sequence after the gesture start frame. If there is a certain gesture detection frame sequence in which the number of frames not containing the target gesture exceeds the second threshold, it may be considered that there is no target gesture in the gesture detection frame sequence, and at least one frame is selected from the gesture detection frame sequence as a gesture end frame. The value of the second threshold may be determined flexibly according to actual conditions, and may be the same as or different from the first threshold. In an example, the value of second threshold may correspond to the value of first threshold, each of which is 6. The process of selecting a gesture end frame from the gesture detection frame sequence may refer to the process of selecting a gesture start frame, and details are not described herein.

[0081] After a plurality of gesture start frames and gesture end frames are obtained, it is possible to determine the number of target gestures occurring in a gesture detection period based on the number of the gesture start frames and gesture end frames and the corresponding gesture start time and gesture end time.

[0082] Table 1 illustrates gesture detection rules according to an embodiment of the present disclosure.

Table 1 Gesture detection rules

Rule	Default value	Note
Detection frame rate	10FPS	Fixed detection frame rate
Interactive gesture detection period	10	With 10 frames as a period, a single frame is continuously scrolled, i.e., each frame is the end of

		the last period of 10 frames and the start of the next period of 10 frames.
Interactive gesture rule threshold	6	In the detection period, if the number of frames having a positive detection result (i.e., a gesture is detected) reaches threshold, it is determined that there is an interactive gesture, and the interactive event is marked to start. In the detection period, if the number of frames having a negative detection result (i.e., no gesture is detected) reaches threshold, it is determined that there is no interactive gesture, and the interactive event is marked to end.
Interactive gesture start time		If the rule threshold is 6, the time of the sixth frame having a positive detection result is the start time of gesture interaction event.
Interactive gesture end time		If the rule threshold is 6, the time of the sixth frame having a negative detection result is the end time of gesture interaction event.

[0083] The interactive gesture detection period in the Table 1 corresponds to the gesture detection frame sequence in the above embodiments of the disclosure, the interactive gesture rule threshold corresponds to the first threshold and the second threshold in the above embodiments of the disclosure, the start time of interactive gesture corresponds to the start time of gesture in the above embodiments of the disclosure, and the end time of interactive gesture corresponds to the end time of gesture in the above embodiments of the disclosure. As can be seen from Table 1, in an example, every 10 frames in a gesture detection period may be taken as a gesture detection frame sequence, and thus gesture detection is performed on each frame in each gesture detection frame sequence to determine gesture start frames and gesture end frames, and then the number of target gestures occurring in each gesture detection period is obtained.

[0084] By the above process, detection of the number of the target gestures can be realized based on a plurality of gesture detection frame sequences in a gesture detection period,

which effectively reduces the influence of inaccurate gesture detection results of individual frames on the gesture detection result, and improves the accuracy of gesture detection, thereby improving the accuracy and reliability of the entire data processing process.

[0085] Further, after the number of times that the target gesture is performed in the gesture detection period is acquired, the gesture detection result corresponding to the gesture detection period may be obtained according to the acquired number of times. In a possible implementation, the number of times that the target gesture is performed in the gesture detection period may be directly used as a gesture detection result. In a possible implementation, the number of times that the target gesture is performed in the gesture detection period may be mapped to a score according to a certain rule and the score is used as a gesture detection result. The mapping rule is not limited in the embodiments of the present disclosure.

[0086] Table 2 illustrates mapping rules for gesture detection results according to an embodiment of the present disclosure.

Table 2 Mapping rules for gesture detection results

Rule	Default value	Note
Interactive gesture scoring period	1 minute	The number of interactions is detected in each minute. If one gesture interaction crosses two of 1 minute, it is denoted as the 1 minute when the gesture interaction starts.
Interactive gesture scoring rule	10-point scale	The score of one interactive gesture is 1, the score of ten interactive gestures is 10, and the highest score is not more than 10. Raising a hand and AR gestures are used as interactive gestures to participate in the scoring.

[0087] The interactive gesture scoring period corresponds to the gesture detection period in the above embodiments of the disclosure. It can be seen from Table 2 that, in an example, one target gesture may be scored as 1 point in a gesture detection period, so as to

determine a score in the gesture detection period according to the number of target gestures. If a target gesture occurs more than 10 times in a gesture detection period, the gesture detection result of the gesture detection period is scored as 10 points.

[0088] By mapping the number of times that the target gesture is performed in the gesture detection period into a score according to a certain rule, the gesture detection result may be normalized, so that the standardization of the intermediate detection result determined based on the gesture detection result is improved, which facilitates combination of the intermediate detection result in the gesture dimension with the intermediate detection result in other dimensions to obtain a more visual target detection result.

[0089] After the gesture detection results in each gesture detection period are obtained, the intermediate detection result in the gesture detection dimension may be further obtained based on at least one gesture detection result. The manner of obtaining the intermediate detection result according to the gesture detection result can be flexibly determined, and is not limited to the following embodiments of the disclosure. In a possible implementation, the average value of the gesture detection results in each gesture detection period may be used as the intermediate detection result in the gesture detection dimension.

[0090] In a possible implementation, in a case that the detection dimension includes an emotion detection dimension, an intermediate detection result of the target object in the emotion detection dimension may be obtained according to the following operations.

[0091] An expression detection result and/or a smile detection result of the target object in an emotion detection period are acquired according to video data. The expression detection result includes an emotion result determined based on an expression of the target object, and the smile detection result includes a smile intensity of the target object.

[0092] The intermediate detection result of the target object in the emotion detection dimension is obtained according to the expression detection result and/or the smile detection result of the target object in at least one emotion detection period.

[0093] An implementation of the emotion detection period may refer to the implementation of the gesture detection period in the above embodiments of the disclosure, and details are not described herein. The length of the emotion detection period and the length of the gesture detection period may be the same or different, and may be flexibly selected according to actual conditions. In an example, the emotion detection period may be set to be the same as the gesture detection period, each of which is one minute.

[0094] The expression detection result may be an emotion result determined by performing expression detection on the target object, for example, the emotion of the target object is pleasure, calm, or sad. The implementation of the expression detection result may be flexibly set. The acquisition manner and implementation of the expression detection result may refer to the subsequent embodiments of the disclosure and will not expand them herein.

[0095] The smile detection result may be a related result determined by performing smile detection on the target object, which may reflect the smile intensity or the smile amplitude of the target object, etc. The implementation of the smile detection result may be flexibly set. The acquisition manner and implementation of the smile detection result may refer to the subsequent embodiments of the disclosure and will not expand them herein.

[0096] Further, how to obtain the intermediate detection result of the target object in the emotion detection dimension according to the expression detection result and the smile detection result may be determined according to the actual conditions of the expression detection result and the smile detection result, and details may also be described in the subsequent embodiments of the disclosure.

[0097] In some embodiments of the present disclosure, the smile detection and the expression detection may be two independent detections, and although both may be used to indicate the emotion state of the target object, they are implemented from two different perspectives. The intermediate detection result determined is based on both the

expression detection result and the smile detection result, which can indicate the emotion state of the target object in the emotion detection dimension more comprehensively and reliably, thereby improving the comprehensiveness and reliability of the final target detection result.

[0098] Specifically, how to obtain the expression detection result of the target object in a emotion period can be flexibly determined according to the actual conditions. In a possible implementation, the operation that the expression detection result of the target object in the emotion detection period is acquired according to the video data may include: performing expression detection on the target object in the emotion detection period and determining the number of times that the target object presents at least one target expression to obtain the expression detection result; where the target expression includes at least one of pleasure, calm, or others.

[0099] As can be seen from the above embodiments of the disclosure, in a possible implementation, the expression detection result may be acquired based on the number of times that the target object presents different target expressions in the emotion detection period. The target expression may be flexibly set according to actual conditions. In a possible implementation, the target expression may be set to be pleasure, calm, or other expressions. In a possible implementation, the other expressions may be further refined, for example, the target expression may be set to be pleasure, calm, sad, or angry.

[00100] The number of times for a target expression in the emotion detection period is determined by the number of detected frames containing the target expression. In a possible implementation, the detection frame rate of the emotion detection may be the same as the detection frame rate of the gesture detection, which is a certain fixed value, and thus the number of frames that can be detected in each emotion detection period is fixed. In a possible implementation, the number of times that the target object presents every target expression in the emotion detection period may be determined according to the number of frames where the target expression can be detected in the emotion detection period. In a possible implementation, the emotion detection period may also be

divided into a plurality of emotion detection sub-periods, and the target expression with the largest number of frames detected in each emotion detection sub-period is used as the expression of the emotion detection sub-period, so that the number of target expressions in the emotion detection period is determined based on the expression of each emotion detection sub-period.

[00101] Table 3 illustrates expression detection rules according to an embodiment of the present disclosure.

Table 3 Expression detection rules

Rule	Default value	Note
Detection frame rate	10FPS	Fixed detection frame rate
Expression rule		Three types of expressions are defined, i.e., pleasure, calm, and others. The expression detection result per second is the mode of results of all detection frames.

[00102] It can be seen from Table 3, when the emotion detection period is one minute, one second may be used as the length of the emotion detection sub-period to obtain 60 emotion detection sub-periods. Then, expression detection may be performed on each frame of the video data in each second to obtain a target expression corresponding to each frame, and a target expression with the largest number of frames in the second is used as the target expression corresponding to the second. Therefore, the occurrence times of different target expressions may be obtained in each detection period, and the sum of the occurrence times is 60.

[00103] The specific manner of performing the expression detection on each frame is not limited. In a possible implementation, the expression detection on each frame of image may be realized by the neural network having the expression detection function, that is, each frame of image is input to the neural network having the expression detection function, and the target expression corresponding to the target object may be output. A

specific implementation of the neural network having the expression detection function is not limited in the embodiments of the disclosure, and a suitable neural network can be flexibly selected according to actual conditions. In an example, expression detection of a target object may be achieved by invoking models such as face detection or face attributes in the human action SDK.

[00104] Further, after the number of times of each target expression in the emotion detection period is determined, the expression detection result in the emotion detection period can be acquired. Specifically, the mapping rules about how to convert the number of times of different target expressions into the expression detection result can be flexibly determined according to actual conditions, and are not limited to the following embodiments of the disclosure.

[00105] Table 4 illustrates corresponding rules of an emotion detection result according to an embodiment of the present disclosure.

Table 4 Corresponding rules of emotion detection result

Rule	Default value	Note
Emotion scoring period	1 minute	
Emotion score	10-point scale	Emotion score is the average of expression score and smile score.
Expression scoring rule	10-point scale	1 minute includes 60 seconds, which correspond to 60 expression detections and are assigned values respectively: 10 points for pleasure, 5 points for calm, 0 point for others. The average of all 60 scores is scored as the affinity score of the minute.
Smile scoring rule	10-point scale	The average of all 60 scores is scored as the smile score of the minute.

[00106] The expression score corresponds to the expression detection result in the above embodiments of the disclosure, and the smile score corresponds to the smile detection result in the above embodiments of the disclosure. As shown in Table 4, in an example,

different target expressions in the emotion detection period may be recorded as different scores, for example, pleasure may be scored as 10 points, calm as 5 points, others as 0 point, etc., and then the average score of the target expressions in the emotion detection period is used as an expression detection result in the emotion detection period.

[00107] Expression detection is performed on the target object in the emotion detection period to determine the number of times that the target object presents at least one target expression, so as to obtain an expression detection result. Based on a plurality of different target expressions of the target object in the emotion detection period, a more comprehensive and reliable expression detection result may be obtained, so that the emotion of the target object can be more accurately reflected and the accuracy of the emotion detection result can be improved.

[00108] Similarly, the implementation of how to acquire the smile detection result of the target object in the emotion period can be flexibly determined according to the actual conditions. In a possible implementation, the operation that the smile detection result of the target object in the emotion detection period is acquired according to the video data may include the following operations.

[00109] Smile detection is performed on the target object in the emotion detection period according to at least one frame of the video data to obtain a smile detection result corresponding to the at least one frame; and the smile detection result of the target object in the emotion detection period is determined according to the smile detection result corresponding to the at least one frame. As can be seen from the above embodiments of the disclosure, in a possible implementation, smile detection may be performed on each frame of the video data in the emotion detection period, and then an average is obtained based on the smile detection result of some or each of the frames, to obtain the smile detection result of the target object in the emotion detection period.

[00110] In a possible implementation, it is also possible to refer to the implementation of the expression detection. The emotion detection period is divided into a plurality of emotion

detection sub-periods, a smile detection result in each of the emotion detection sub-period is obtained, and then a smile detection result in the emotion detection period is obtained based on the smile detection result of each emotion detection sub-period. For the division manner of the emotion detection period, reference may be made to the embodiments of the present disclosure as described above, and details are not described herein. The determination manner of the smile detection result in each emotion detection sub-period can be flexibly determined according to actual conditions.

[00111] Table 5 shows smile detection rules according to an embodiment of the present disclosure.

Table 5 Smile detection rules

Rule	Default value	Note
Detection frame rate	10FPS	Fixed detection frame rate
Smile rules		The smile value of 0-1 detected each time is converted to 0-10 scores. The smile value of 10 frames detected per second are averaged, and the average result is taken as the smile value of the second.

[00112] As can be seen from the table, in an example, the emotion detection period may be further divided into a plurality of emotion detection sub-periods according to seconds. In each emotion detection sub-period, smile detection may be performed on each frame in the emotion detection sub-period, and then the average of the smile detection results of all frames is used as the smile detection result of the emotion detection sub-period.

[00113] The manner of performing smile detection on each frame of image is also not limited in the embodiments of the disclosure. In a possible implementation, a frame of image may be input to a neural network having a smile detection function, and a smile value corresponding to the frame of image may be output. The implementation of the neural network having the smile detection function is not limited in the embodiments of the disclosure, and any neural network which can reflect the smile amplitude or intensity of the target object in the image can be implemented as the neural network having the

smile detection function.

[00114] After the emotion detection period is divided into a plurality of emotion detection sub-periods, the manner of obtaining the smile detection result in the emotion detection period according to the smile detection results of the emotion detection sub-periods may be flexibly determined. As can be seen from Table 4 mentioned in the above embodiments of the disclosure, in an example, the average of the 60 smile detection results in one minute may be used as the smile detection result in the emotion detection period. That is, the smile detection result in the emotion detection period may be obtained by using the average of the smile detection results in the emotion detection sub-periods of the emotion detection period.

[00115] The smile detection result in the emotion detection period is acquired according to smile detection results of different frames in the emotion detection period. In this way, the influence of inaccuracy of smile detection result of the partial frames can be reduced, so that the obtained smile detection result in the emotion detection period has high reliability, and thus the reliability and accuracy of the final target detection result can be improved.

[00116] After the expression detection result and the smile detection result in the emotion detection period are obtained, the intermediate detection result of the target object in the emotion detection dimension may be further obtained based on the expression detection result and the smile detection result. In a possible implementation, emotion detection results of emotion detection periods may be obtained based on respective expression detection results and/or smile detection results in respective emotion detection periods, and the intermediate detection result of the target object in the emotion detection dimension may be obtained by performing average on the emotion detection results of different emotion detection periods.

[00117] The manner of acquiring the emotion detection results of the target object in respective emotion detection periods is not limited, as shown in Table 4. In a possible

implementation, an average of an expression detection result and a smile detection result in an emotion detection period may be used as the emotion detection result of the emotion detection period. In a possible implementation, weighted average may be performed on the expression detection result and the smile detection result to obtain the emotion detection result in the emotion detection period. The weights of the expression detection result and the smile detection result may be flexibly set according to actual conditions, and are not limited to the following embodiments of the disclosure as long as the sum of the two results is 1. In an example, the weight of the expression detection result may be set to 1 and the weight of the smile detection result may be set to 0, that is, the expression detection result may be directly used as the emotion detection result in the emotion detection period. In an example, the weight of the expression detection result may be set to 0 and the weight of the smile detection result may be set to 1, that is, the smile detection result may be directly used as the emotion detection result in the emotion detection period.

[00118] In a possible implementation, in a case that the detection dimension includes an eye contact detection dimension, an intermediate detection result of the target object in the eye contact detection dimension may be obtained according to the following operations.

[00119] Face angle detection is performed on the target object according to video data, and a time duration when a face angle of the target object is within a face angle threshold is determined as a face angle detection result. Eye closure detection is performed on the target object according to the video data, and a time duration when the target object performs an eye closure operation is determined as an eye closure detection result. A length of time for which the face angle of the target object is within the face angle threshold and no eye closure operation is performed is determined according to the face angle detection result and the eye closure detection result. The intermediate detection result of the target object in the eye contact detection dimension is obtained according to the length of time.

[00120] As can be seen from the above embodiment of the disclosure, in a possible

implementation, the detection of the target object in the eye contact detection dimension may be composed of two parts, including face angle detection and eye closure detection. In some embodiments of the disclosure, face orientation of the target object may be determined by the face angle detection. If the face orientation of the target object is within the face angle threshold, the viewing angle of the target object may be considered to be within the range of the eye contact. The specific value of the face angle threshold may be flexibly set according to the actual conditions. In a possible implementation, the face angle threshold may be a static value, that is, the value of the face angle threshold does not change at any time in the video data. In a possible implementation, the face angle threshold may also be set to a dynamic value, that is, the value of the face angle threshold changes flexibly according to the position change of the target object in the video data.

[00121] However, in some possible cases, although the viewing angle of the target object is within the range of the eye contact, it may merely be a habitual swing of the human face, rather than the eye contact with the object to communicate with. Therefore, in some embodiments of the present disclosure, it is also possible to further perform eye closure detection on the target object to determine whether the target object is in an eye closure state. If the viewing angle of the target object is within the range of eye contact, and the target object is in an eye-opening state (that is, not in the eye closure state), it may be considered that the target object is performing eye contact currently. Therefore, in a possible implementation, a length of time for which the face angle of the target object is within the face angle threshold and the target object performs no eye closure operation may be determined by the face angle detection and the eye closure detection, and the intermediate detection result of the target object in the eye contact detection dimension is obtained according to the proportion of the length of time in the video data.

[00122] The length of time for which the face angle of the target object is within the face angle threshold and no eye closure operation is performed is determined according to the face angle detection and the eye closure detection, and then the intermediate detection result of the target object in the eye contact detection dimension is obtained. Through the

above process, whether the target object is in a direction of the eye contact and whether the target object performs the eye closure operation in the direction can be considered during the detection of the eye contact, so as to comprehensively evaluate the degree of eye contact of the target object, thereby greatly improving the accuracy of the intermediate detection result in the eye contact detection dimension, and further improving the accuracy of the subsequent target detection result.

[00123] Furthermore, in a possible implementation, in order to increase the accuracy of the behavior state detection in the eye contact detection dimension, an eye contact detection period may be set with reference to the above-described embodiments of the disclosure, so that the length of time for which the face angle of the target object is within the face angle threshold and no eye closure operation is performed is determined in each eye contact detection period, to obtain an intermediate detection result in at least one eye contact detection period. An intermediate detection result of the target object in the eye contact detection dimension is obtained based on the intermediate detection result in the at least one eye contact detection period.

[00124] Implementation of the eye contact detection period may refer to the gesture detection period and the emotion detection period in the above-described embodiments of the disclosure, and details are not described herein. In a possible implementation, the length of the eye contact detection period may be set to one minute.

[00125] Specifically, in each eye contact detection period, the process of performing face angle detection may refer to the process of performing gesture detection. Therefore, in a possible implementation, the operation that face angle detection is performed in the eye contact detection period may include the following operations.

[00126] At least one face angle detection frame sequence of the video data in the eye contact detection period is acquired. If the number of frames where the face angle is within the face angle threshold in the face angle detection frame sequence exceeds a third threshold, at least one frame in the face angle detection frame sequence is recorded as a face

orientation start frame. If the number of frames where the face angle is outside the face angle threshold in a face angle detection frame sequence after the face orientation start frame exceeds the fourth threshold, at least one frame in the face angle detection frame sequence after the face orientation start frame is recorded as a face orientation end frame. The time duration for which the face angle of the target object is within the face angle threshold in the eye contact period is obtained according to the number and time of the face orientation start frames and the face orientation end frames.

[00127] The acquisition manner of the face angle detection frame sequence may refer to the that of the gesture detection frame sequence, the determination manner of the face orientation start frame may refer to that of the gesture start frame, and the determination manner of the face orientation end frame may refer to that of the gesture end frame, and details are not described herein again. The third threshold and the fourth threshold may be values flexibly set according to actual conditions, and may be the same as or different from the first threshold and the second threshold as long as they are flexibly set according to actual conditions. A manner of face angle detection may be flexibly determined according to actual conditions. In a possible implementation, each frame image may be input to a neural network having a face angle detection function to implement the face angle detection. An implementation of the neural network having a face angle detection function is not limited in the embodiments of the disclosure. In an example, the neural network performing the face angle detection may be obtained by invoking a model such as face detection or face tracking in the human action SDK.

[00128] Table 6 illustrates face angle detection rules according to an embodiment of the present disclosure.

Table 6 Face angle detection rules

Rule	Default value	Note
Viewing threshold settings		It is determined based on whether the yaw value and pitch value of the headpose are within a preset viewing range. The following 4 default settings for viewing threshold

		are supported, and 4 thresholds specified by the upload interface are also supported. Key_parameter_negative_yaw Key_parameter_positive_yaw Key_parameter_negative_pitch Key_parameter_positive_pitch
Detection frame rate	10FPS	Fixed detection frame rate
Viewing detection period	10	With 10 frames as a period, a single frame is continuously scrolled, i.e., each frame is the end of the last period of 10 frames and the start of the next period of 10 frames.
Viewing Rule threshold	8	In the detection period, if the number of frames where detection result is positive (within the set viewing range) reaches a threshold, it is determined that there is a viewing event, marking the start of the viewing event. In the detection period, if the number of frames whose detection result are negative (not within the set viewing range) reaches a threshold, it is determined that there is no viewing event, marking the end of the viewing event.
Start time of viewing event		If the rule threshold is 8, the time of the 8th frame where detection result is positive is the start time of the viewing event.
End time of viewing event		If the rule threshold is 8, the time of the 8th frame where detection result is negative is the end time of the viewing event.

[00129] The viewing threshold may correspond to the face angle threshold in the above embodiments of the disclosure, the viewing detection period may correspond to the face angle detection frame sequence in the above embodiments of the disclosure, the viewing rule threshold may correspond to the third threshold and the fourth threshold in the above embodiments of the disclosure, the start time of viewing event may correspond to the time of face orientation start frame in the above embodiments of the disclosure, and

end time of the viewing event may correspond to the time of face orientation end frame in the above embodiments of the disclosure. As shown in Table 6, in an example, the face angle threshold may include four parameters, which are positive yaw angle, negative yaw angle, positive pitch angle, and negative pitch angle. The specific values of the four parameters may be flexibly determined according to actual conditions. The yaw angle and the pitch angle may be flexibly determined according to a coordinate system set in the face angle detection, and the positive and the negative angles may be used to represent directions of the angles, etc. If it is detected that the face angle in a frame picture is within a range jointly determined by the four parameters, the face angle in the frame picture may be considered to be within the face angle threshold. The detection frame rate of the face angle detection may be set to 10FPS, the length of the face angle detection frame sequence may be set to 10, and both the third threshold and the fourth threshold may be set to 8. That is, in a face angle detection frame sequence, if the number of frames where the face angle is within the face angle threshold is not less than 8, the 8th frame where the face angle is within the face angle threshold may be used as the face orientation start frame, and the corresponding time is the start time when the face angle is within the face angle threshold. In the same way, the end time when the face angle is within the face angle threshold may be determined, and then the time range in which the face angle is within the face angle threshold in the eye contact period may be obtained.

[00130] Similarly, in each eye contact detection period, the process of performing eye closure detection may refer to the process of gesture detection and face angle detection. Therefore, in a possible implementation, the process of performing eye closure detection in the eye contact detection period may include the following operations.

[00131] At least one eye closure detection frame sequence of the video data in the eye contact detection period is acquired, When it is detected that the number of frames where both eyes are in the eye closure state in the eye closure detection frame sequence exceeds a fifth threshold, at least one frame in the eye closure detection frame sequence is recorded as an eye closure start frame. When it is detected that the number of frames

where both eyes are not in the eye closure state or only one eye is in the eye closure in the eye closure detection frame sequence after the eye closure start frame state exceeds the sixth threshold, at least one frame in the eye closure detection frame sequence after the eye closure start frame is recorded as an eye closure end frame. The time duration for which the target object is in the eye closure state in the eye contact period is obtained according to the number and time of the eye closure start frames and the eye closure end frames.

[00132] The acquisition manner of the eye closure detection frame sequence and the determination manner of the eye closure start frame and the eye closure end frame may refer to each of the above-described embodiments of the disclosure, and details are not described herein. The fifth threshold and the sixth threshold may be values flexibly set according to actual conditions, and may be the same as or different from the respective threshold mentioned above as long as they are flexibly set according to actual conditions. The manner of detecting whether the eye of the target object is closed may be flexibly determined according to the actual conditions. In a possible implementation, each frame picture may be input to a neural network having the eye closure detection function to implement eye closure detection. The implementation of the neural network having the eye closure detection function is not limited in the present disclosure embodiment. In an example, the neural network that can perform eye closure detection may be obtained by invoking a model such as face detection or face attribute in the human action SDK.

[00133] Table 7 shows eye closure detection rules according to an embodiment of the present disclosure.

Table 7 Eye closure detection rules

Rule	Default value	Note
Eye closure definition		It is eye closure if left and right eyes are both closed; It is not eye closure when one eye is not closed.
Detection frame rate	10FPS	Fixed detection frame rate
Eye closure	10	With 10 frames as a period, a single frame is

detection period		continuously scrolled, i.e., each frame is the end of the last period of 10 frames and the start of the next period of 10 frames.
Eye closure rule threshold	Forward threshold is 6 Negative threshold is 8	In the detection period, if the number of frames where detection results are positive (eye closure) reaches a threshold, that is, it is determined that there is eye closure, making the start of the eye closure event. In the detection period, if the number of frames where detection result are negative (non-eye closure) reaches a threshold, that is, it is determined that there is viewing, marking the end of the eye closure event.
start time of eye closure event		If the rule threshold is forward threshold, the time of the first frame where detection result is positive is the start time of the eye closure event.
End time of eye closure event		If the rule threshold is negative, the time of the first frame where detection result is negative is the end time of the eye closure event.

[00134] The eye closure detection period may correspond to the eye closure detection frame sequence in the above embodiments of the disclosure, the eye closure rule threshold may correspond to the fifth threshold and the sixth threshold in the above embodiments of the disclosure, the start time of eye closure event may correspond to the time of the eye closure start frame in the above embodiments of the disclosure, and the end time of eye closure event may correspond to the time of the eye closure end frame in the above embodiments of the disclosure. As shown in Table 7, in an example, a state of both eyes of the target object being closed is set to the eye closure state and the remaining states may be set to non-eye closure state. The detection frame rate of the eye closure detection may be set to 10FPS, the length of the eye closure detection frame sequence may be set to 10, the fifth threshold may be set to 6, and the sixth threshold may be set to 8. That is, in an eye closure detection frame sequence, if the number of frames being in the eye

closure state is not less than 6, the first frame in the eye closure state may be used as the eye closure start frame, and the corresponding time is the start time of the eye closure. Similarly, in the eye closure detection frame sequences after the start time of the eye closure, if the number of frames detected in the non-eye closure state is not less than 8, the first frame in the non-eye closure state may be used as the eye closure end frame, so that the end time of eye closure may be determined, and thus the time range in which the target object is in the eye closure state in the eye contact period may be obtained.

[00135] After the time range in which the face of the target object is within the face angle threshold and the time range in which the target object is within the eye closure state are respectively determined in the eye contact period, a time range in which the face of the target object is within the face angle threshold and not in the eye closure state in the eye contact period, i.e., the time range in which the target object performs eye contact in the eye contact period can be further obtained, and then the intermediate detection result of the eye contact period is determined. The specific manner of how to map the time range of the eye contact of the target object in the eye contact period to the intermediate detection result can be flexibly set according to the actual conditions, and is not limited to the following embodiments of the disclosure.

[00136] Table 8 shows eye contact detection result rules according to an embodiment of the present disclosure. The eye contact scoring period may correspond to the eye contact detection period in the above-described embodiment of the disclosure, and the eye contact score may correspond to the intermediate detection result in the eye contact detection dimension in the above-described embodiment of the disclosure.

Table 8 Eye contact detection result rules

Rule	Default value	Note
Eye contact scoring period	1 minute	
Eye contact scoring rules	10-point scale	The proportion of cumulative time length of eye contact per minute was converted to a score of 10-point scale. For example, in a 1-minute video, the cumulative time

		length of eye contact is 30 seconds, the corresponding proportion is 50%, and the corresponding score is 5.0.
--	--	---

[00137] As can be seen from Table 8, in an example, the time proportion of the eye contact in the eye contact detection period can be mapped to a score according to a certain rule, as the intermediate detection result in the eye contact detection dimension in the eye contact detection period.

[00138] Similarly, as described in the above-described embodiments, in a possible implementation, the multimedia data may include audio data, and in this case, the operation S12 may also include performing behavior state detection on behaviors of the target object according to the audio data. Therefore, in a possible implementation, S12 may include S1221 and S1222.

[00139] In S1221, the audio data is segmented according to sentences to obtain at least one audio sub-data.

[00140] In S1222, behavior state detection is performed on the at least one audio sub-data in at least one of the following detection dimensions: fluency, speech rate, pause, or volume, to obtain intermediate detection results of the target object in the at least one detection dimension

[00141] The implementation of segmenting the audio data according to sentences is not limited in the embodiments of the disclosure, and is not limited to the following embodiments of the disclosure. In a possible implementation, the audio data may be recognized by an audio data recognition neural network capable of recognizing the text in the audio data, thereby obtaining the recognition result of each sentence in the audio data, such as each sentence in the audio data, words included in each sentence, a start time stamp of each sentence, a time length of each sentence, a start time stamp of a word, and a time length of a word. The specific implementation of the audio data recognition neural network can be flexibly determined, and any neural network that can recognize the audio data may be used as an implementation of the audio data recognition neural

network.

[00142] One or more audio sub-data may be obtained after segmenting the audio data according to sentences. The implementation of the audio sub-data and the contents contained therein may be flexibly determined according to the actual conditions of segmentation of the audio data. In a possible implementation, in the case that the audio data is segmented according to sentences, each obtained audio sub-data may correspond to each complete sentence in the audio data.

[00143] After the audio sub-data are obtained, the behavior state detection may be performed on part or each of the audio sub-data according to the obtained audio sub-data. In the embodiment of the present disclosure, the detection of the audio sub-data may also be performed in different dimensions, for example, at least one of the following detection dimensions: fluency, speech rate, pause, or volume. The specific dimensions to be selected may be flexibly determined according to actual conditions, and is not limited in the embodiment of the present disclosure.

[00144] Specifically, the manner of detecting at least one of fluency, speech rate, pause, or volume of the audio sub-data is not limited. In a possible implementation, a plurality of neural networks having different functions, such as a fluency detection neural network, a speech rate detection neural network, a pause detection neural network, and a volume detection neural network, may be obtained by training. Audio sub-data are input to these neural networks, and detection results of the fluency, the speech rate, the pause and the volume or the like may be output respectively. The specific implementations of the above-mentioned neural network may be flexibly determined according to actual conditions, and are not limited in the embodiments of the present disclosure.

[00145] After the intermediate detection results of each audio sub-data in each detection dimension are obtained, the intermediate detection results of each audio sub-data in each detection dimension can be weighted and fused according to a proportion of time length, so that the weighted and fused result is used as the intermediate detection results of the

complete audio data in each detection dimension.

[00146] In a possible implementation, format of the audio data may also be transcoded before detecting the audio data, so that subsequent audio data detection processes are easier to implement. The manner of transcoding and the transcoded format may be flexibly determined according to actual detection requirements. In a possible implementation, the audio data may be transcoded into a pcm format (such as an uncompressed pcm file or a wav file) or a mono format of 16bit sampling bits. Similarly, the video data may be transcoded to a suitable video format before detecting the video data.

[00147] Audio data is segmented according to sentences to obtain at least one audio sub-data, and detection in one or more detection dimensions of fluency, speech rate, pause and volume is performed on the at least one audio sub-data. Through the above process, the detection process of the audio data can be converted into the detection process of each sub-data in the audio data, which reduces the detection difficulty and the amount of data to be processed each time, thereby improving the detection efficiency and detection accuracy of the audio data, and thus improving the efficiency and accuracy of data processing.

[00148] After the intermediate detection results of the target object in a plurality of detection dimensions are obtained based on any one of the above embodiments or a combination of the embodiments of the disclosure, the obtained intermediate detection results may be processed by the operation S13 to obtain the target detection result of the target object. The implementation of S13 may be flexibly determined according to actual conditions, and is not limited to the following embodiments of the disclosure.

[00149] In a possible implementation, S13 may include that: intermediate detection results in at least one detection dimension are combined according to a preset weight of the detection dimension to obtain the target detection result of the target object.

[00150] It can be seen from the above embodiments that, in a possible implementation, the multimedia data can be detected in one or more detection dimensions of gesture,

emotion, eye contact, fluency, speech rate, pause, and volume to obtain intermediate detection results in each of the detection dimensions respectively. Accordingly, the intermediate detection results in these detection dimensions may be fused or combined to obtain the target detection result.

[00151] The process of fusing or combining may be flexibly selected according to actual conditions. As described in the above embodiment, in a possible implementation, the intermediate detection results on each detection dimension may be weighted and averaged according to the preset weight of each detection dimension to obtain the target detection result of the target object. The values of the preset weights for respective detection dimensions may be flexibly set according to actual requirements. In a possible implementation, a detection dimension having a large influence on the state evaluation of the target object may be set with a higher preset weight. In a possible implementation, the preset weights of respective detection dimensions may be set to be consistent, and in this case, the average value of the intermediate detection results on each detection dimension may be directly used as the target detection result of the target object.

[00152] According to the above process, the final target detection result may be obtained based on the intermediate detection results of respective detection dimensions. Since the preset weight can be adjusted according to the actual requirement of the behavior state average, the obtained target detection result can better reflect the behavior state of the target object and has high reliability.

[00153] As already mentioned in the above embodiment, the implementation of the target detection result can be flexibly selected, for example, an overall detection result and a plurality of detailed subdivision results can be included for simultaneously reflecting the overall situation and detailed situation of the behavior state of the target object. Therefore, in a possible implementation, S13 may also include the operations S131-S133.

[00154] In S131, video sub-data corresponding to the audio sub-data is determined from the

video data included in the multimedia data according to a time duration that the audio sub-data occupies in the audio data.

[00155] In S132, the intermediate detection results of the audio sub-data in at least one detection dimension and the intermediate detection results of the video sub-data corresponding to the audio sub-data in the at least one detection dimension are combined according to a preset weight, to obtain target detection results of the at least one audio sub-data or the at least one video sub-data;

[00156] In S133, the target detection results of the at least one audio sub-data or the at least one video sub-data are combined to obtain the target detection result of the target object.

[00157] As described in the above embodiments of the disclosure, in a possible implementation, the multimedia data may include both the video data and the audio data. The video data and the audio data may be in a one-to-one correspondence, that is, the video data and the audio data may be separated from the audio-containing data separately. In the process of performing multi-dimensional behavior state detection on the audio data, a plurality of audio sub-data may be obtained by segmenting the audio data according to sentences, and intermediate detection results of the audio sub-data in a plurality of detection dimensions such as fluency, speech rate, pause, and volume are obtained respectively.

[00158] Therefore, in a possible implementation, in order to further obtain the target detection result of these audio sub-data, it is also possible to further obtain the intermediate detection results of the audio sub-data in the detection dimensions of gesture, emotion, and eye contact. The specific manner of obtaining may refer to S131. In a possible implementation, the video data may be segmented according to the segmenting manner of the audio sub-data in the audio data, thereby obtaining a plurality of video sub-data. Since the audio data corresponds to the video data, and the segmenting manner of the audio sub-data and the audio sub-data are the same, the obtained video sub-data and the audio sub-data is in one-to-one correspondence. Since

the behavior state detection can be performed on the video data in any of the above-described embodiments, intermediate detection results in a plurality of detection dimensions is obtained. Further, the intermediate detection results are mapped to respective video sub-data in a segmenting manner, and the intermediate detection results of the video sub-data in at least one detection dimension can be obtained.

[00159] After obtaining the intermediate detection results of each video sub-data in at least one detection dimension, the intermediate detection results in each dimension of the video sub-data and the intermediate detection results in each dimension of the audio sub-data may be combined according to the correspondence between the video sub-data and the audio sub-data, so as to obtain target detection results of respective audio sub-data. Since the audio sub-data correspond to the video sub-data, the target detection results of the respective audio sub-data may also be target detection results of respective video sub-data. The combination manner may refer to the above embodiments of the disclosure, and details are not described herein.

[00160] Further, after obtaining the target detection results of respective audio sub-data or video sub-data, the target detection results of different audio sub-data or different video sub-data again may be fused again in an inverse manner of the segmenting manner of the audio sub-data or video sub-data to obtain the overall target detection result of the target object.

[00161] By the above process, in addition to obtaining the overall target detection result of the target object, the target detection results of the target object in each sentence can be obtained, so that the behavior state of the target object is better reflected, and thus the reference value and the utilization range of the target detection result are improved.

[00162] FIG. 2 shows a block diagram of a data processing apparatus according to an embodiment of the present disclosure. As shown in FIG.2, the data processing apparatus 20 may include an acquiring module 21, a detecting module 22, and a processing module 23.

[00163] The acquiring module 21 is configured to acquire multimedia data of a target object.

The detecting module 22 is configured to perform behavior state detection on the target object in at least one detection dimension according to the multimedia data to obtain intermediate detection results of the target object in the at least one detection dimension.

The processing module 23 is configured to process the intermediate detection results in the at least one detection dimension to obtain a target detection result of the target object.

The target detection result is used to indicate a behavior state of the target object.

[00164] In a possible implementation, the multimedia data includes video data. The detecting module is configured to determine the target object in the video data; and perform behavior state detection on the target object in at least one of the following detection dimensions: gesture, emotion or eye contact, to obtain the intermediate detection results of the target object in the at least one detection dimension.

[00165] In a possible implementation, the at least one detection dimension includes a gesture detection dimension. The detecting module is further configured to acquire, according to the video data, the number of times that the target object performs at least one target gesture in a gesture detection period to obtain a gesture detection result in the gesture detection period, where the target gesture includes at least one of: holding a hand, raising a hand, or thumbs up; and obtain the intermediate detection result of the target object in the gesture detection dimension according to the gesture detection result in at least one gesture detection period .

[00166] In a possible implementation, the detecting module is further configured to acquire at least one gesture detection frame sequence of the video data in the gesture detection period; in a case that the number of frames containing the target gesture in the gesture detection frame sequence exceeds a first threshold, record at least one frame in the gesture detection frame sequence as a gesture start frame; in a case that the number of frames not containing the target gesture in a gesture detection frame sequence after the gesture start frame exceeds a second threshold, record at least one frame in the gesture detection frame sequence after the gesture start frame as a gesture end frame; and

acquire the number of times that the target object performs the at least one target gesture in the gesture detection period according to the number of gesture start frames and gesture end frames.

[00167] In a possible implementation, the at least one detection dimension includes an emotion detection dimension. The detecting module is further configured to acquire, according to the video data, an expression detection result and/or a smile detection result of the target object in an emotion detection period, where the expression detection result includes an emotion result determined based on an expression of the target object, and the smile detection result includes a smile intensity of the target object; and obtain the intermediate detection result of the target object in the emotion detection dimension according to the expression detection result and/or the smile detection result of the target object in at least one emotion detection period.

[00168] In a possible implementation, the detecting module is further configured to perform expression detection on the target object in the emotion detection period, and determine the number of times that the target object presents at least one target expression, to obtain the expression detection result, where the target expression includes at least one of pleasure, calm, or others.

[00169] In a possible implementation, the detecting module is further configured to perform smile detection on the target object in the emotion detection period according to at least one frame of the video data to obtain a smile detection result corresponding to the at least one frame; and determine the smile detection result of the target object in the emotion detection period according to the smile detection result corresponding to the at least one frame.

[00170] In a possible implementation, the at least one detection dimension includes an eye contact detection dimension. The detecting module is further configured to: perform face angle detection on the target object according to the video data, and determine a time duration when a face angle of the target object is within a face angle threshold as a face

angle detection result; perform eye closure detection on the target object according to the video data, and determine a time duration when the target object performs an eye closure operation as an eye closure detection result; determine, according to the face angle detection result and the eye closure detection result, a length of time for which the face angle of the target object is within the face angle threshold and no eye closure operation is performed; and obtain the intermediate detection result of the target object in the eye contact detection dimension according to the length of time.

[00171] In a possible implementation, the multimedia data includes audio data. The detecting module is configured to: perform segmentation on the audio data according to sentences to obtain at least one audio sub-data; and perform behavior state detection on the at least one audio sub-data in at least one of the following detection dimensions: fluency, speech rate, pause, or volume, to obtain intermediate detection results of the target object in the at least one detection dimension.

[00172] In a possible implementation, the processing module is configured to combine, according to a preset weight of each detection dimension, the intermediate detection results in the at least one detection dimension to obtain the target detection result of the target object.

[00173] In a possible implementation, the processing module is configured to: determine video sub-data corresponding to the audio sub-data from the video data included in the multimedia data according to a time duration that the audio sub-data occupies in the audio data; combine, according to a preset weight, the intermediate detection results of the audio sub-data in the at least one detection dimension with the intermediate detection results of the video sub-data corresponding to the audio sub-data in the at least one detection dimension, to obtain target detection results of the at least one audio sub-data or the at least one video sub-data; and combine the target detection results of the at least one audio sub-data or the at least one video sub-data to obtain the target detection result of the target object.

[00174] In a possible implementation, the multimedia data is obtained by the target object performing a teaching operation according to preset text data. The preset text data includes at least one instruction marker for segmenting and/or annotating at least part of contents in the preset text data.

[00175] Various embodiments of the present disclosure may be combined with each other without departing from the logic, the various embodiments have their respective emphasis, and the parts thereof that do not emphasize can be referred to the disclosure of other embodiments.

[00176] In some embodiments of the present disclosure, functions or modules included in the apparatus provided by the embodiments of the present disclosure may be configured to perform the method steps described in the above method embodiments, and specific implementations and technical effects thereof may be described with reference to the above method embodiments, and details are not described herein for brevity.

[00177] **Application scenario example**

[00178] For example, an application scenario is the teacher-simulated-teaching, that is, a teacher simulates teaching lessons, which may be that a plurality of teachers who face each other in an offline scene perform simulated teaching and give evaluations to each other. With the development of online teaching, the simulated teaching may also be converted to an online simulated teaching, that is, a teacher may record or broadcast a simulated teaching process through a terminal device (such as a mobile phone or a computer).

[00179] The simulated teaching can help teachers to preview the formal teaching, and the evaluation of the simulated teaching has high guiding value for teaching work of teachers. Therefore, a more reliable simulated teaching method can be effectively applied to the online simulated teaching process of the teachers, which plays a better auxiliary role in the formal teaching of the teachers.

[00180] The application example of the present disclosure provides a teacher simulated teaching system, which can effectively evaluate the behavior states of a teacher in the simulated teaching process through the data processing method as mentioned in the above embodiments of the disclosure.

[00181] The teacher simulated teaching system provided in the present disclosure may include a client (such as a mobile phone, a computer, a user device, and the like) and a server (such as a local server or a cloud server). The teacher may record or broadcast a simulated teaching process at the client, and upload the recorded or broadcast result as multimedia data to the server. The server may receive the multimedia data uploaded by the client, and process the multimedia data by the data processing method of the above embodiments of the disclosure, thereby obtaining a target detection result of a target object.

[00182] The process of the teacher performing a simulated teaching at the client may include the following operations.

[00183] The client can display four parts of the simulated teaching process through a presentation interface, which includes pre-class warm-up, knowledge teaching, in-class training, and classroom test. Each part corresponds to a tap in the presentation interface and the teacher can enter the part by clicking on the tap. During a teacher clicking on the tap, the server may collect time stamps when the teacher clicks on each tap, so that the multimedia data recorded by the teacher corresponds to one or more of the four parts.

[00184] The teacher may perform a simulated teaching according to an existing verbatim manuscript (i.e., the preset text data in the above embodiments of the disclosure). The verbatim manuscript may be a text file in txt format, and the verbatim manuscript may include an instruction marker for structuring annotation, knowledge point annotation and interaction annotation of the verbatim manuscript, so that the verbatim manuscript may be segmented into the above four parts, and in each part, the teacher may be prompted to make interactions in an appropriate position, including voice content, interaction

gestures, and the like.

[00185] The verbatim manuscript may be segmented by structuring annotation according to different parts of the simulated teaching. In the application example of the present disclosure, the start and end links of the four parts (pre-class warm-up, knowledge teaching, in-class training, classroom detection) are annotated with specific start and end instructions in the verbatim manuscript.

[00186] In an example, the corresponding instructions are annotated by using <start of start instruction>, <end of start instruction>, <start of end instruction> and <end of end instruction>, so that it is possible to determine and identify the part corresponding to the current content taught by the teacher according to the instruction marker of structuring annotation. In the application example of the present disclosure, reference may be made to the above embodiments of the disclosure for a specific implementation of the instruction marker of structuring annotation.

[00187] The knowledge point annotation and interaction annotation may be used to annotate knowledge points of the simulated teaching and interaction positions in the verbatim manuscript. In the application example of the present disclosure, the key point may be annotated by using <start of key point> and <end of key point>, thereby facilitating detection of the important paragraphs in the simulated teaching process. If interactions are needed in the simulated teaching process, the corresponding teaching content of the verbatim manuscript is annotated with a marker of <interaction needed>. In the application example of the present disclosure, reference may be made to the above embodiments of the disclosure for a specific implementation of the instruction marker of the knowledge point annotation and interaction annotation.

[00188] In the application example of the present disclosure, the teacher may record the simulated teaching process through the client while performing the simulated teaching by using the verbatim manuscript, thereby obtaining the multimedia data regarding the teacher and uploading the multimedia data to the server.

[00189] The process of performing data processing on the multimedia data uploaded from the client by the server may include the following operations: acquisition of multimedia data, processing of the multimedia data, and video detection.

[00190] Multimedia data is acquired. The server acquires to-be-processed multimedia data by initiating a request. The request initiated by the server may include a URL link of multimedia data (such as a MP4 file), vendeeID, teacherID, a simulated teaching ID, a multimedia data structure (that is, portions of the multimedia data formed by segmenting according to a simulated teaching process, start time stamp and end time stamp of each portion), a video detection frame rate, and the like. In the application example of the present disclosure, in a case that the multimedia data includes video data, the resolution of the video may include various forms, such as 640p, 720p, or 1080p, and in a case that the multimedia data includes audio data, the audio data may include various audio sampling rates, such as 8000Hz or 16000Hz, and the like. In addition, in a case that the teacher performs simulated teaching live broadcast through the client, the server can also acquire multimedia data (i.e., video and audio data) in real time.

[00191] Multimedia data is preprocessed (such as video transcoding or audio transcoding, etc.). The server may separate the video stream and the audio stream from the obtained multimedia data and transcode them into a format supported by video detection, speech recognition, or speech evaluation, respectively. For example, the separated audio stream may be converted to a pcm (uncompressed pcm file or wav file) or mono format of 16bit sample bit.

[00192] Video is detected. In the application example of the present disclosure, the face detection, face tracking, face attribute and gesture detection model in the human action SDK may be invoked, and the human body detection and hand raising detection mode in the insight SDK may be invoked to perform multi-dimensional detection on the video data. In the application example of the disclosure, multi-dimensional detection of video data may include gesture detection, emotion detection, eye contact detection, and the like.

[00193] The gesture detection may reflect the interaction degree during the teacher simulated teaching. The gesture detection may support detection of three kinds of gestures, including holding a hand (indicating that a student is asked to answer a question), raising a hand (indicating that the student is prompted to answer a question), or thumbs up (indicating that the behavior of the student is praised). The detection manner may be detected using the neural network for gesture detection, so that the number of times that each gesture occurs and the time stamp of each gesture detection can be output. For a specific implementation of gesture detection, reference may be made to each of the above embodiments of the disclosure. For a rule for obtaining an intermediate detection result in a gesture detection dimension, reference may be made to Table 1 and Table 2 in the above embodiments of the disclosure, and details are not described herein.

[00194] Emotion detection may reflect the affinity of the teacher in the simulated teaching, and may include two aspects, namely expression detection and smile detection. The expression detection may be performed by the neural network for expression detection. On the basis of a single-frame detection result, the expression detection result is output according to an emotion detection period (temporally set at a minute). An exemplary rule may be that the expression detected with the most times in the emotion detection period may be used as the expression detection result in the emotion detection period.

[00195] Similarly, on the basis of the single-frame detection result, the smile detection result is output according to the emotion detection period (temporally set at a minute). An exemplary rule may be that the arithmetic mean value of all the single-frame smile detection results in the emotion detection period may be used as the smile detection result of the emotion detection period.

[00196] For a specific implementation of the expression detection and the smile detection, reference may be made to the above embodiments of the disclosure. For a rule for obtaining an intermediate detection result in the emotion detection dimension, reference may be made to Table 3 to Table 5 in the above embodiments of the disclosure, and details are not described herein.

[00197] The eye contact detection may reflect the situation of eye contact with the students during the teacher simulated teaching, and may include face angle detection (i.e., headpose) and eye closure detection. The eye contact detection may be defined as an eye contact event, the face angle detection may be defined as a viewing event, and the eye closure detection may be defined as an eye closure event, and the eye contact event may be an intersection of the viewing event and non-eye closure event. In the application example of the present disclosure, the start time of the eye contact event may be set to a start time that is within a time range of the viewing event and is not within a time range of the eye closure event, and the end time of the eye contact event may be set to an end time of the viewing event or a start time of the eye closure event.

[00198] For specific implementations of face angle detection and eye closure detection, reference may be made to the above embodiments of the present disclosure, and a rule for obtaining the intermediate detection result in the eye contact detection dimension may be referred to Table 6 to Table 8 in the above embodiments of the present disclosure, and details are not described herein.

[00199] Audio is recognized.

[00200] In an application example of the present disclosure, the relevant recognition model for voice recognition may be invoked to input audio data, so as to obtain voice recognition results in real time, including sentences in the audio data, words in the sentences, and start time stamps and time lengths of each sentence and each word.

[00201] Audio is detected.

[00202] In the application example of the present disclosure, the audio of a sentence may be segmented based on the start time stamp and time length of each sentence in the voice recognition result, and the audio detection result of the sentence may be acquired and returned. The detection result includes fluency, speech rate, pause, volume, and the like.

[00203] By audio recognition and audio detection, it is possible to reflect the intermediate

detection results in the dimensions of fluency, speech rate, volume, etc., during a teacher simulated teaching. In an application example of the present disclosure, audio detection may support Chinese voice recognition for evaluation on the simulated teaching of non-English courses. The audio detection may also support voice recognition on mixed-reading Chinese and English for evaluation on simulated teaching of English courses.

[00204] The neural network model related to the voice recognition may be invoked in the voice recognition to return the recognition result in real time. The recognition result includes sentences and words in the sentences. The sentences returned may be recognized for the voice by the audio detection to obtain detection results in each of the above dimensions. Further, the audio detection for paragraphs may be added.

[00205] A target detection result is generated.

[00206] The target detection result may include an overall target detection result and a subdivided target detection result. The overall target detection result may include interaction, fluency, speech rate, and volume. The interaction may be further segmented into gesture interaction, emotion interaction, and eye contact interaction. FIG. 3 shows a schematic diagram of the target detection result according to an application example of the present disclosure. It can be seen from FIG. 3 that the overall target detection result may include an overall score calculated based on the intermediate detection results in each dimension and a score of the intermediate detection results in each dimension, etc. It should be noted that FIG. 3 is only an exemplary schematic diagram showing the target detection result. In practical application, the target detection result can be visually displayed in any form according to actual requirements.

[00207] The subdivided target detection result may be a detection result output based on each sentence obtained by voice recognition. In an example, the subdivided target detection result may include a sentence ID, a text of a sentence, a start time stamp of a sentence, a time length of a sentence, a fluency of a sentence, a speech rate of a sentence, a volume

of a sentence, a gesture of a sentence (supporting a plurality of gestures), an expression of a sentence, a smile value of a sentence, and the like.

[00208] The system mentioned in the application example of the present disclosure may be applied to, in addition to the analysis of teacher simulated teaching, other related fields such as formal teaching analysis of a teacher, or evaluation of the trial teaching of a teacher candidate.

[00209] It will be appreciated that each of the above-mentioned method embodiments in the present disclosure may be combined with each other without departing from the principle and logic to form combined embodiments, which will not be described in detail in the present disclosure.

[00210] It will be appreciated by those skilled in the art that in the above methods of the detailed description, the written order of the steps does not mean to limit the implementation process in a strict execution order, and the specific execution order of the steps should be determined in terms of their functions and possible intrinsic logics.

[00211] An embodiment of the present disclosure further provides a computer readable storage medium, having stored thereon computer program instructions that, when being executed by a processor, cause to implement the above method. The computer-readable storage medium may be a volatile computer-readable storage medium or a non-volatile computer-readable storage medium.

[00212] An embodiment of the present disclosure further provides an electronic device including a processor; and a memory for storing instructions executable by the processor. The processor is configured to implement the method as described above.

[00213] An embodiment of the present disclosure further provides a computer program including computer readable codes that, when run in an electronic device, cause a processor in the electronic device to implement the method as described above.

[00214] In practical applications, the memory may be a volatile memory, such as a RAM; or

non-volatile memory, such as a ROM, flash memory, Hard Disk Drive (HDD) or Solid-State Drive (SSD); or a combination of the above types of memories, and provide instructions and data to the processor.

[00215] The processor may be at least one of an ASIC, a DSP, a DSPD, a PLD, an FPGA, a CPU, a controller, a microcontroller, and a microprocessor. It will be appreciated that for different devices, the electronics for implementing the functions of above processor may be other devices, which are not specifically limited in the embodiments of the present disclosure.

[00216] The electronic device may be provided as a terminal, a server, or other devices.

[00217] Based on the same technical concept as the foregoing embodiments, an embodiment of the present disclosure further provides a computer program that, when executed by a processor, implements the method as described above.

[00218] FIG. 4 is a block diagram of an electronic device 800 according to an embodiment of the present disclosure. For example, the electronic device 800 may be a mobile phone, a computer, a digital broadcast terminal, a messaging device, a game console, a tablet device, a medical device, an exercise device, a personal digital assistant, or the like.

[00219] Referring to FIG. 4, the electronic device 800 may include one or more of 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.

[00220] The processing component 802 generally controls the overall operation of the electronic device 800, such as operations associated with displays, telephone calls, data communications, camera operations, and recording operations. The processing component 802 may include one or more processors 820 which execute instructions to perform all or a portion of the steps of the methods described above. In addition, the processing component 802 may include one or more modules which facilitate

interactions between the processing component 802 and other components. For example, the processing component 802 may include a multimedia module which facilitates interactions between the multimedia component 808 and the processing component 802.

[00221] The memory 804 is configured to store various types of data to support operations at the electronic device 800. Examples of such data include instructions for any application or method operating on the electronic device 800, contact data, phone book data, messages, pictures, video, etc. The memory 804 may be implemented by any type of volatile or non-volatile storage device or combination thereof, such as static random access memory (SRAM), electrically erasable programmable read-only memory (EEPROM), erasable programmable read-only memory (EPROM), programmable read-only memory (PROM), read-only memory (ROM), magnetic memory, flash memory, magnetic disk, or optical disk.

[00222] The power component 806 provides power to 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 generating, managing, and distributing power for the electronic device 800.

[00223] 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 input signals from a user. The TP includes one or more touch sensors to sense touch, slide, and gestures on the TP. The touch sensor may not only sense the boundary of a touch or sliding action, but also detect the time duration and pressure associated with the touch or sliding operation. In some embodiments, the multimedia component 808 includes a front-facing camera and/or a rear-facing camera. When the electronic device 800 is in an operation mode, such as a shooting mode or a video mode, the front-facing camera and/or the rear-facing camera may receive external multimedia data. Each of the front and rear cameras may be a fixed optical lens system or have a focal length and optical zoom capability.

[00224] The audio component 810 is configured to output and/or input audio signals. For example, the audio component 810 includes a microphone (MIC) configured to receive an external audio signal when the electronic device 800 is in an operating mode, such as a call mode, a recording mode, and a voice recognition mode. The received audio signal may be further stored in the memory 804 or transmitted via the communication component 816. In some embodiments, the audio component 810 further includes a speaker for outputting an audio signal.

[00225] The I/O interface 812 provides an interface between the processing component 802 and a peripheral interface module, which may be a keyboard, a click wheel, a button, or the like. These buttons may include, but are not limited to, a home page button, a volume button, an activation button, and a lock button.

[00226] The sensor component 814 includes one or more sensors for providing state assessment of various aspects of the electronic device 800. For example, the sensor component 814 may detect an on/off state of the electronic device 800, a relative positioning of the components, such as a display and keypad of the electronic device 800. The sensor component 814 may also detect a change in position of the electronic device 800 or one of the components of the electronic device 800, the presence or absence of user contact with the electronic device 800, an 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 the presence of a nearby object in the absence of any physical contact. The sensor component 814 may also include a light sensor, such as a CMOS or CCD picture sensor for use in imaging applications. 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.

[00227] The communication component 816 is configured to facilitate wired or wireless communication between electronic device 800 and other devices. The electronic device 800 may access a wireless network based on a communication standard, such as WiFi,

2G, 3G, 4G, 5G, or a combination thereof. In an exemplary embodiment, the communication component 816 receives broadcast signals or broadcast related personnel information from an external broadcast management system via 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 Radio Frequency Identification (RFID) technology, Infrared Data Association (IrDA) technology, Ultra Wideband (UWB) technology, Bluetooth (BT) technology, and other technologies.

[00228] In an exemplary embodiment, the electronic device 800 may be implemented by one or more application specific integrated circuits (ASICs), digital signal processors (DSPs), digital signal processing devices (DSPDs), programmable logic devices (PLDs), field programmable gate arrays (FPGAs), controllers, microcontrollers, microprocessors, or other electronic components for performing the methods as described above.

[00229] In an exemplary embodiment, a non-volatile computer-readable storage medium is also provided, such as a memory 804 including computer program instructions executable by a processor 820 of the electronic device 800 to perform the methods as described above.

[00230] FIG. 5 is a block diagram of an electronic device 1900 according to an embodiment of the present 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, which further includes one or more processors, and memory resources represented by a memory 1932 for storing instructions, such as applications, that may be executed by the processing component 1922. The application stored in the memory 1932 may include one or more modules each corresponding to a set of instructions. In addition, the processing component 1922 is configured to execute the instructions to perform the methods as described above.

[00231] The electronic device 1900 may also include: a power component 1926 configured

to perform power management of the electronic device 1900; a wired or wireless network interface 1950 configured to connect the electronic device 1900 to a network; and an input/output (I/O) interface 1958. The electronic device 1900 may operate based on an operating system stored in memory 1932, such as Windows Server™, Mac OS X™, Unix™, Linux™, FreeBSD™, or the like.

[00232] In an exemplary embodiment, a non-volatile computer-readable storage medium is also provided, such as a memory 1932 including computer program instructions executable by a processing component 1922 of an electronic device 1900 to perform the method as described above.

[00233] The present disclosure may provide a system, method, and/or computer program product. The computer program product may include a computer readable storage medium having stored thereon computer readable program instructions for causing a processor to implement various aspects of the present disclosure.

[00234] The computer readable storage medium may be a tangible device that may hold and store instructions used by the instruction execution device. The computer readable storage medium may be, for example, but not limited to, an electrical storage device, a magnetic storage device, an optical storage device, an electromagnetic storage device, a semiconductor storage device, or any suitable combination of the foregoing. More specific examples (non-exhaustive lists) of the computer readable storage medium include a portable computer disk, a hard disk, a random access memory (RAM), a read only memory (ROM), an erasable programmable read only memory (EPROM or flash memory), a static random access memory (SRAM), a portable compact disk read only memory (CD-ROM), a digital versatile disk (DVD), a memory stick, a floppy disk, a mechanical encoding device, e.g., a punch card or in-groove bump structure on which instructions are stored, and any suitable combination of the foregoing. As used herein, a computer readable storage medium is not to be construed as an instantaneous signal, such as a radio wave or other freely propagating electromagnetic wave, an electromagnetic wave propagating through a waveguide or other transmission medium

(e.g., an optical pulse through a fiber optic cable), or an electrical signal transmitted through a wire.

[00235] The computer readable program instructions described herein may be downloaded from a computer readable storage medium to various computing/processing devices, or downloaded via a network, such as the Internet, a local area network, a wide area network, and/or a wireless network, to an external computer or external storage device. The network may include copper transmission cables, fiber optic transmission, wireless transmission, routers, firewalls, switches, gateway computers, and/or edge servers. A network adapter card or network interface in each computing/processing device receives computer readable program instructions from the network and forwards the computer readable program instructions for storage in a computer readable storage medium in the respective computing/processing device.

[00236] The computer program instructions used to perform the operations of the present disclosure may be assembly instructions, instruction set architecture (ISA) instructions, machine instructions, machine-related instructions, microcode, firmware instructions, state setting data, or source or object code written in any combination of one or more programming languages, which includes object-oriented programming languages such as Smalltalk, C++, and the like, and conventional procedural programming languages such as "C" or a similar programming language. The computer readable program instructions may be executed entirely on the user computer, partly on the user computer, as a separate software package, partly on the user computer and partly on the remote computer, or entirely on the remote computer or server. In the case of a remote computer, the remote computer may be connected to the user computer through any kind of network, including a local area network (LAN) or a wide area network (WAN), or may be connected to an external computer (e.g., connected through the Internet using an Internet service provider). In some embodiments, various aspects of the present disclosure are implemented by customizing an electronic circuit, such as a programmable logic circuit, a field programmable gate array (FPGA), or a programmable logic array (PLA), and the electronic circuit can execute computer

readable program instructions.

[00237] Various aspects of the present disclosure are described herein with reference to flowcharts and/or block diagrams of methods, apparatus (systems), and computer program products in accordance with embodiments of the present disclosure. It should be understood that each block of the flowcharts and/or block diagrams, and combinations of blocks in the flowcharts and/or block diagrams, may be implemented by computer readable program instructions.

[00238] These computer readable program instructions may be provided to a processor of a general purpose computer, a special computer, or other programmable data processing apparatus to produce a machine that causes the instructions, when executed by the processor of the computer or other programmable data processing apparatus, to produce apparatus for implementing the functions/acts specified in one or more blocks of the flowchart and/or block diagram. The computer readable program instructions may also be stored in a computer readable storage medium that cause a computer, programmable data processing apparatus, and/or other device to operate in a particular manner, so that the computer readable medium having the instructions stored thereon includes an article of manufacture that includes instructions that implement various aspects of the functions/actions specified in one or more blocks of the flowchart and/or block diagram.

[00239] The computer readable program instructions may also be loaded onto a computer, other programmable data processing apparatus, or other device such that a series of operational steps are performed on the computer, other programmable data processing apparatus or other device to produce a computer-implemented process, so that the instructions executed on the computer, other programmable data processing apparatus, or other device implement the functions/actions specified in one or more blocks of the flowchart and/or block diagram.

[00240] The flowcharts and block diagrams in the drawings illustrate architectures, functions, and operations of possible implementations of systems, methods, and computer program

products in accordance with various embodiments of the present disclosure. In this regard, each block in a flowchart or block diagram may represent a module, program segment, or portion of an instruction that contains one or more executable instructions for implementing a specified logical function. In some alternative implementations, the functions noted in the blocks may also occur in an order different from that noted in the drawings. For example, two successive blocks may actually be executed substantially in parallel, and sometimes they may be executed in the reverse order, depending on the functionality involved. It is also noted that each block of the block diagrams and/or flowcharts, and combinations of blocks in the block diagrams and/or flowcharts, may be implemented with a dedicated hardware-based system that performs the specified functions or actions, or may be implemented with a combination of dedicated hardware and computer instructions.

[00241] The embodiments of the present disclosure have been described above, the foregoing description is illustrative, not exhaustive and is not limited to the various embodiments disclosed. Many modifications and variations will be apparent to those of ordinary skill in the art without departing from the scope and spirit of the illustrated embodiments. The choice of terms used herein is intended to best explain the principles, practical applications, or technical improvements in the marketplace of the various embodiments, or to enable others of ordinary skill in the art to understand the various embodiments disclosed herein.

CLAIMS

1. A data processing method, comprising:

acquiring multimedia data of a target object;

performing behavior state detection on the target object in at least one detection dimension according to the multimedia data to obtain intermediate detection results of the target object in the at least one detection dimension; and

processing the intermediate detection results in the at least one detection dimension to obtain a target detection result of the target object, wherein the target detection result is used to indicate a behavior state of the target object.

2. The method of claim 1, wherein the multimedia data comprises video data; and

wherein performing the behavior state detection on the target object in the at least one detection dimension according to the multimedia data to obtain the intermediate detection results of the target object in the at least one detection dimension comprises:

determining the target object in the video data; and performing behavior state detection on the target object in at least one of the following detection dimensions: gesture, emotion or eye contact, to obtain the intermediate detection results of the target object in the at least one detection dimension.

3. The method of claim 2, wherein the at least one detection dimension comprises a gesture detection dimension; and

wherein obtaining an intermediate detection result of the target object in the gesture detection dimension comprises:

acquiring, according to the video data, a number of times that the target object

performs at least one target gesture in a gesture detection period to obtain a gesture detection result in the gesture detection period, wherein the target gesture comprises at least one of: holding a hand, raising a hand, or thumbs up; and

obtaining the intermediate detection result of the target object in the gesture detection dimension according to the gesture detection result in at least one gesture detection period.

4. The method of claim 3, wherein acquiring, according to the video data, the number of times that the target object performs the at least one target gesture in the gesture detection period comprises:

acquiring at least one gesture detection frame sequence of the video data in the gesture detection period;

in a case that a number of frames containing the target gesture in the gesture detection frame sequence exceeds a first threshold, recording at least one frame in the gesture detection frame sequence as a gesture start frame;

in a case that a number of frames not containing the target gesture in a gesture detection frame sequence after the gesture start frame exceeds a second threshold, recording at least one frame in the gesture detection frame sequence after the gesture start frame as a gesture end frame; and

acquiring the number of times that the target object performs the at least one target gesture in the gesture detection period according to a number of gesture start frames and a number of gesture end frames.

5. The method of any one of claims 2 to 4, wherein the at least one detection dimension comprises an emotion detection dimension; and

wherein obtaining an intermediate detection result of the target object in the emotion detection dimension comprises:

acquiring, according to the video data, an expression detection result and/or a smile detection result of the target object in an emotion detection period, wherein the expression detection result comprises an emotion result determined based on an expression of the target object, and the smile detection result comprises a smile intensity of the target object; and

obtaining the intermediate detection result of the target object in the emotion detection dimension according to the expression detection result and/or the smile detection result of the target object in at least one emotion detection period.

6. The method of claim 5, wherein acquiring, according to the video data, the expression detection result of the target object in the emotion detection period comprises:

performing expression detection on the target object in the emotion detection period, and determining a number of times that the target object presents at least one target expression to obtain the expression detection result, wherein the target expression comprises at least one of pleasure, calm, or others.

7. The method of claim 5 or 6, wherein acquiring, according to the video data, the smile detection result of the target object in the emotion detection period comprises:

performing smile detection on the target object in the emotion detection period according to at least one frame of the video data to obtain a smile detection result corresponding to the at least one frame; and determining the smile detection result of the target object in the emotion detection period according to the smile detection result corresponding to the at least one frame.

8. The method of any one of claims 3 to 7, wherein the at least one detection dimension comprises an eye contact detection dimension; and

wherein obtaining an intermediate detection result of the target object in the eye contact detection dimension comprises:

performing face angle detection on the target object according to the video data, and determining a time duration when a face angle of the target object is within a face angle threshold as a face angle detection result;

performing eye closure detection on the target object according to the video data, and determining a time duration when the target object performs an eye closure operation as an eye closure detection result;

determining, according to the face angle detection result and the eye closure detection result, a length of time duration for which the face angle of the target object is within the face angle threshold and no eye closure operation is performed; and

obtaining the intermediate detection result of the target object in the eye contact detection dimension according to the length of time duration.

9. The method of any one of claims 2 to 8, wherein the multimedia data comprises audio data; and

wherein performing the behavior state detection on the target object in the at least one detection dimension according to the multimedia data to obtain the intermediate detection result of the target object in the at least one detection dimension comprises:

performing segmentation on the audio data according to sentences to obtain at least one audio sub-data; and performing behavior state detection on the at least one audio sub-data in at least one of the following detection dimensions: fluency, speech rate, pause, or volume, to obtain the intermediate detection result of the target object in the at least one detection dimension.

10. The method of any one of claims 1 to 9, wherein processing the intermediate detection results in the at least one detection dimension to obtain the target detection result of the target object comprises:

combining, according to a preset weight of each detection dimension, the intermediate detection results in the at least one detection dimension to obtain the target detection result of the target object.

11. The method of claim 9, wherein processing the intermediate detection results in the at least one detection dimension to obtain the target detection result of the target object comprises:

determining at least one video sub-data corresponding to the at least one audio sub-data from the video data comprised in the multimedia data according to a time duration that the audio sub-data occupies in the audio data;

combining, according to a preset weight, the intermediate detection results of the audio sub-data in the at least one detection dimension with the intermediate detection results of the video sub-data corresponding to the audio sub-data in the at least one detection dimension, to obtain target detection results of the at least one audio sub-data or the at least one video sub-data; and

combining the target detection results of the at least one audio sub-data or the at least one video sub-data to obtain the target detection result of the target object.

12. The method of any one of claims 1 to 11, wherein the multimedia data is obtained by the target object performing teaching operations according to preset text data, wherein the preset text data comprises at least one instruction marker for segmenting and/or annotating at least part of contents in the preset text data.

13. A data processing apparatus, comprising:

an acquiring module, configured to acquire multimedia data of a target object;

a detecting module, configured to perform behavior state detection on the target object

in at least one detection dimension according to the multimedia data to obtain intermediate detection results of the target object in the at least one detection dimension; and

a processing module, configured to process the intermediate detection results in the at least one detection dimension to obtain a target detection result of the target object, wherein the target detection result is used to indicate a behavior state of the target object.

14. An electronic device comprising:

one or more processors; and

a memory for storing instructions executable by the one or more processors;

wherein the one or more processors are configured to invoke the instructions stored in the memory to implement the method of any one of claims 1 to 12.

15. A computer readable storage medium, having stored thereon computer program instructions, that, when being executed by a processor, cause to implement the method of any one of claims 1 to 12.

16. A computer program comprising computer readable codes which, when run in an electronic device, cause a processor in the electronic device to implement the method of any one of claims 1 to 12.