

VOICE REPLY METHOD, DEVICE, SYSTEM, STORAGE MEDIUM AND PRODUCT

[0001] The present disclosure claims priority to Chinese Patent Application No. 202411397888.7, filed with China National Intellectual Property Administration on September 30, 2024 and entitled "VOICE REPLY METHOD AND DEVICE, SYSTEM, STORAGE MEDIUM AND PRODUCT", which is incorporated herein by reference in its entirety.

TECHNICAL FIELD

[0002] Embodiments of the present disclosure relate to the field of computer technologies and, in particular, to a voice reply method, a device, a system, a storage medium and a product.

BACKGROUND

[0003] Intelligent conversation implements interaction between a user and an intelligent device in a question-and-answer mode, and the intelligent conversation is widely used in scenarios such as a voice assistant and an intelligent customer service.

[0004] At present, intelligent conversation is mainly implemented by a terminal device. For example, a user inputs voice information or text information through an interface of the terminal device, and the terminal device outputs corresponding reply information. This manner has the problem of low conversation efficiency.

SUMMARY

[0005] Embodiments of the present disclosure provide a voice reply method and apparatus, a device, a computer-readable storage medium and a product, which are used to solve the technical problem of low efficiency in the prior art when a terminal device is used to implement an intelligent conversation.

[0006] In a first aspect, an embodiment of the present disclosure provides a voice reply method, which is applied to a headphone device carrying a camera, where the method includes: collecting first inquiry voice information; obtaining a target image collected by the camera; obtaining first reply voice information of the first inquiry voice information, where the first reply voice information is information related to the target image, and the target image and the first inquiry voice information are processed by a pre-trained dialogue model to obtain the first reply voice information; and playing the first reply voice information.

[0007] In a second aspect, an embodiment of the present disclosure provides a headphone device, including: the headphone device carries a camera, and the headphone device includes: a collecting unit, configured to collect first inquiry voice information; an obtaining unit, configured to obtain a target image collected by the camera; a processing unit, configured to obtain first reply voice information of the first inquiry voice information, where the first reply voice information is information related to the target image, and the target image and the first inquiry voice information are processed by a pre-trained dialogue model to obtain the first reply voice information; and a playing unit, configured to play the first reply voice information.

[0008] In a third aspect, an embodiment of the present disclosure provides a voice reply system, including the headphone device of the second aspect and a terminal device, where: the headphone device is configured to: collect first inquiry voice information; obtain a target image collected by a camera, and send the target image and the first inquiry voice information to the terminal device; the terminal device is configured to process the target image and the first inquiry voice information by a pre-trained dialogue model to obtain first reply voice information of the first inquiry voice information, and send the first reply voice information to the headphone device; and the headphone device is configured to receive the first reply voice information sent by the terminal device, and play the first reply voice information.

[0009] In a fourth aspect, an embodiment of the present disclosure provides an electronic device, including: a processor and a memory; the memory stores a computer-executable instruction; and the processor executes the computer-executable instruction stored in the memory to cause at least one processor to execute the voice reply method according to the first aspect above and various possibilities of the first aspect.

[0010] In a fifth aspect, an embodiment of the present disclosure provides a computer-readable storage medium, where a computer-executable instruction is stored in the computer-readable storage medium, and when the processor executes the computer-executable instruction, the voice reply method according to the first aspect above and various possibilities of the first aspect is implemented.

[0011] In a sixth aspect, an embodiment of the present disclosure provides a computer program product, including a computer program, where when the computer program is executed by a processor, the voice reply method according to the first aspect above and various possibilities of the first aspect is implemented.

[0012] The voice reply method and apparatus, the device, the storage medium and the product provided by the embodiments include: collecting first inquiry voice information; obtaining a target image collected by a camera; obtaining first reply voice information of the first inquiry voice information, where the first reply voice information is information related to the target image, and the target image and the first inquiry voice information are processed by a pre-trained dialogue model to obtain the first reply voice information; and playing the first reply voice information. First, in the present disclosure, a headphone is used to implement an intelligent conversation, thus improving the efficiency of the intelligent conversation. In addition, in the present disclosure, the first reply voice information may be quickly determined by using the dialogue model. Furthermore, by obtaining the target image to reply to the first inquiry voice information, a reply to a question proposed by a user for a real-time environment may be implemented.

BRIEF DESCRIPTION OF DRAWINGS

[0013] In order to more clearly explain the technical solutions in the embodiments of the present disclosure or in the prior art, the following will briefly introduce the drawings that need to be used in the description of the embodiments or the prior art. Obviously, the drawings in the following description are some embodiments of the present disclosure. For those of ordinary skill in the art, other drawings may be obtained based on these drawings without creative efforts.

- [0014] Fig. 1 is an example diagram of an application scenario of a voice reply method provided by an embodiment of the present disclosure;
- [0015] Fig. 2 is a flowchart of a voice reply method provided by an embodiment of the present disclosure;
- [0016] Fig. 3 is a flowchart of another voice reply method provided by an embodiment of the present disclosure;
- [0017] Fig. 4 is a schematic structural diagram of a headphone device provided by an embodiment of the present disclosure; and
- [0018] Fig. 5 is a schematic structural diagram of an electronic device provided by an embodiment of the present disclosure.

DETAILED DESCRIPTION OF EMBODIMENTS

- [0019] In order to make the objectives, technical solutions and advantages of the embodiments of the present disclosure clearer, the technical solutions in the embodiments of the present disclosure will be described clearly and completely below in combination with the drawings in the embodiments of the present disclosure. Obviously, the described embodiments are part of the embodiments of the present disclosure, but not all of the embodiments. Based on the embodiments in the present disclosure, all other embodiments obtained by those of ordinary skill in the art without paying creative labor belong to the protection scope of the present disclosure.
- [0020] It may be understood that before using the technical solutions disclosed in each embodiment of the present disclosure, the target user should be informed of the type, range of use, use scenarios, etc. of the personal information involved in the present disclosure in an appropriate manner and obtain the authorization of the target user in accordance with relevant laws and regulations.
- [0021] It may be understood that the above process of notifying and obtaining the authorization of the target user is only illustrative, and does not constitute a limitation on the implementation of the present disclosure, and other methods that meet relevant laws and regulations may also be applied to the implementation of the present disclosure.
- [0022] In the related art, if a user needs to inquire about the relevant information of an object in the surrounding environment, a mobile terminal with a camera needs to be used to take a picture of the object to obtain an image, and then a relevant recognition application in the mobile terminal is used to scan the image, and the relevant information of the object is displayed on the interface of the mobile terminal. If the user needs to know further information about the object, it needs to be searched for in the application based on the relevant information. For example, when the user is in the current environment and sees a bunch of flowers, the user does not know the name of the flower, so the user uses the mobile terminal to take a picture of the flower to obtain an image containing the flower, and then the user uses the recognition application on the mobile terminal to scan the image to obtain the name of the flower. Then, if the user needs to know the flowering period, planting place, etc. of the flower, the user needs to input a question about the flower in

the application to get a corresponding answer. It may be seen that the process of acquiring the relevant information of the object is complicated and inefficient.

[0023] In order to solve the above technical problem, the present disclosure provides a voice reply method, in which a camera is deployed on a headphone device to control the camera to obtain a target image in real time, and the headphone device collects first inquiry voice information, obtains first reply voice information for the target image and the first inquiry voice information, and plays the first reply voice information through the headphone device, so that the user may quickly and easily obtain desired information, thereby improving the conversation efficiency.

[0024] It should be noted that the voice reply method and apparatus, the device, the system, the storage medium and the product provided by the present disclosure may be applied to any headphone device configured with a camera.

[0025] The voice reply method provided by the embodiments of the present disclosure will be described in detail below in combination with the drawings.

[0026] Fig. 1 is an example diagram of an application scenario of a voice reply method provided by an embodiment of the present disclosure. As shown in Fig. 1, a user wears a headphone device, and the headphone device includes: a first headphone and a second headphone, one ear of the user wears the first headphone, and the other ear wears the second headphone. The headphone device carries a camera, and in the embodiment of the present disclosure, the first headphone and/or the second headphone may carry the camera. In addition, the first headphone and/or the second headphone carry a microphone, and the microphone is used to collect inquiry voice information of the user. It may be understood that in the present disclosure, the camera carried by the headphone device collects a target image, the microphone collects the inquiry voice information of the user, reply voice information may be generated according to the target image and the inquiry voice information of the user, and the reply voice information may be played through the headphone device, making the whole conversation process simple and efficient, and improving the user experience.

[0027] Fig. 1 is only an exemplary application scenario, which may be used in any headphone device configured with a camera, such as an in-ear headphone, a semi-in-ear headphone, an over-ear headphone, a wired headphone, a TWS (True Wireless Stereo, true wireless headphone), etc. The present disclosure does not limit the type of the headphone device.

[0028] Fig. 2 is a flowchart of a voice reply method provided by an embodiment of the present disclosure, and the voice reply method may include the following steps.

[0029] S201, collecting first inquiry voice information.

[0030] The first inquiry voice information is collected by a microphone of the headphone device. It may be understood that one headphone (the first headphone or the second headphone) of the headphone device carries the microphone, and the microphone may collect the first inquiry voice information of the user in real time.

- [0031] It may be understood that the collected first inquiry voice information needs to be processed by a preset application using a dialogue model. In an embodiment, the preset application has been woken up, and the first inquiry voice information is any inquiry voice information input by the user, such as "What's in front?", "What's the name of this flower", "What variety does this flower belong to" or "What color does this flower have", and so on.
- [0032] In another embodiment, the preset application is not woken up, and the first inquiry voice information includes: a preset wake-up word of the preset application and inquiry voice information, such as "XX, what's in front", "XX, XX, what's in front", "XX, look what's in front", where "XX" is the preset wake-up word of the preset application.
- [0033] S202, obtaining a target image collected by a camera.
- [0034] The camera collects at least one target image or a series of continuous target images.
- [0035] In an embodiment, before the obtaining the target image collected by the camera, the method further includes: waking up the camera in response to a wake-up operation of the camera. Obtaining the target image collected by the camera includes: controlling the woken-up camera to collect the target image.
- [0036] It may be understood that in the case that the camera has been woken up, if the first inquiry voice information is collected, the camera is controlled to collect the target image after the first inquiry voice information is collected.
- [0037] The camera may be woken up by a wake-up word or by touching the headphone device. The wake-up word of the camera may be the same as the wake-up word of the preset application. For example, the wake-up words are such as "XX", or "XX, look", "XX, scan", "XX, look again", "XX, take another look", "XX, look closely".
- [0038] In another embodiment, the obtaining the target image collected by the camera includes: determining a shooting intention of the first inquiry voice information by a recognition model; and controlling the camera to collect the target image according to the shooting intention.
- [0039] It may be understood that after the first inquiry voice information is collected, it is first recognized the wake-up word of the camera that is included in the first inquiry voice information, then the camera is woken up, and the shooting intention of the first inquiry voice information is analyzed by the recognition model. For example, if the first inquiry voice information is "What's the name of this flower", the shooting intention is to shoot a flower, and then the camera may be controlled to focus on the flower to shoot to obtain the target image. By controlling the camera to collect the target image according to the shooting intention, an accurate target image may be obtained, which may then be used to accurately reply to the user subsequently.
- [0040] The recognition model is a network model deployed on the headphone device.
- [0041] In the embodiment of the present disclosure, the headphone device includes: a first headphone and a second headphone, the first headphone carries a first camera, the second headphone carries a second

camera, and the target image includes: a target image collected by the first camera and a target image collected by the second camera.

[0042] It may be understood that in an embodiment, at least one camera may be set only on one headphone, for example, only the first headphone is provided with the first camera, and only the first camera needs to collect the target image. In another embodiment, both headphones may be provided with a camera, then the cameras of both headphones collect the target images.

[0043] The target image collected by the first camera and the target image collected by the second camera are different angles of the same object, which is used for the subsequent determination of the first reply voice information, and may improve the reply accuracy of the first reply voice information.

[0044] S203, obtaining first reply voice information of the first inquiry voice information.

[0045] The first reply voice information is information related to the target image, and the target image and the first inquiry voice information are processed by a pre-trained dialogue model to obtain the first reply voice information.

[0046] The dialogue model is a LLM (Large Language Model), and the dialogue model is a pre-trained multimodal language model, which may process the target image and the first inquiry voice information in natural language to obtain the first reply voice information. The first reply voice information is information related to the target image. For example, if the first inquiry voice information is "What's in front?", and the target image correspondingly includes a "flower", the first reply voice information may be "There is a pot of flower in front". If the first inquiry voice information is "What is this flower?", and the target image includes a "rose", the first reply voice information may be "This is a rose".

[0047] In an embodiment, a dialogue model is deployed in the headphone device, and the obtaining the first reply voice information of the first inquiry voice information includes: inputting the target image and the first inquiry voice information into the pre-trained dialogue model for processing to obtain the first reply voice information of the first inquiry voice information.

[0048] It may be understood that the dialogue model is deployed on the headphone device side, and after collecting the first inquiry voice information and the target image, the headphone device inputs the first inquiry voice information and the target image into the dialogue model for processing to obtain the first reply voice information of the first inquiry voice information.

[0049] In the embodiment of the present disclosure, the first inquiry voice information in a voice format may be input into the dialogue model, or the first inquiry voice information may be converted into first inquiry text information by STT (Speech To Text) technology, and the first inquiry text information is input into the dialogue model. The dialogue model may also output first reply text information, and the headphone device uses TTS (Text To Speech) technology to convert the first reply text information from a text format to a voice format to obtain the first reply voice information.

[0050] In the embodiment of the present disclosure, it may also be the case that a dialogue application is installed on the headphone device side, and a dialogue model is deployed in a cloud server, and the

headphone device may upload the target image and the first inquiry voice information to the dialogue model of the cloud service through the dialogue application for processing to obtain the first reply voice information.

[0051] In another embodiment, the obtaining the first reply voice information of the first inquiry voice information includes: sending the target image and the first inquiry voice information to a terminal device, the terminal device is configured to process the target image and the first inquiry voice information by a pre-trained dialogue model to obtain the first reply voice information of the first inquiry voice information; and receiving the first reply voice information sent by the terminal device.

[0052] It may be understood that the dialogue model is deployed on the terminal device side, and the headphone device sends the collected target image and the collected first inquiry voice information to the terminal device, and the terminal device processes the target image and the first inquiry voice information through the dialogue model to obtain the first reply voice information.

[0053] In the embodiment of the present disclosure, it may also be the case that a dialogue application is installed on the terminal device side, and a dialogue model is deployed in a cloud server, and the terminal device may upload the target image and the first inquiry voice information to the dialogue model of the cloud service through the dialogue application for processing to obtain the first reply voice information, and then send the first reply voice information to the headphone device for playing.

[0054] In summary, in the present disclosure, the first reply voice information may be obtained by processing the target image and the first inquiry voice information through the dialogue model in various ways.

[0055] S204, playing the first reply voice information.

[0056] In the embodiment of the present disclosure, the first reply voice information may be played in one headphone (the first headphone or the second headphone), or the first reply voice information may be played in both the first headphone and the second headphone.

[0057] In summary, in the present disclosure, by deploying a camera on a headphone device, controlling the camera to obtain a target image in real time, collecting first inquiry voice information by the headphone device, obtaining first reply voice information for the target image and the first inquiry voice information, and playing the first reply voice information through the headphone device, a user may quickly and easily obtain desired information, thereby improving conversation efficiency.

[0058] Fig. 3 is a flowchart of another voice reply method provided by an embodiment of the present disclosure, and the voice reply method may include the following steps.

[0059] S301, collecting first inquiry voice information.

[0060] For the specific implementation process of this step, please refer to S201, which will not be repeated here.

[0061] S302, obtaining a target image collected by a camera.

- [0062] For the specific implementation process of this step, please refer to S202, which will not be repeated here.
- [0063] S303, sending the target image and the first inquiry voice information to a terminal device.
- [0064] In the embodiment of the present disclosure, the headphone device is in communication connection with the terminal device, and the headphone device sends the target image and the first inquiry voice information to the terminal device through Bluetooth or WIFI (mobile hotspot). When the headphone device sends the target image and the first inquiry voice information to the terminal device through WIFI, the headphone device may connect to the hotspot of the terminal device.
- [0065] The user may set the camera of a single-sided headphone (the first headphone or the second headphone) to shoot the target image, or may set the cameras of both-sided headphones (the first headphone and the second headphone) to shoot the target image.
- [0066] In an embodiment, the sending the target image to the terminal device includes: sending the target image collected by the first camera to the terminal device by the first headphone; and sending the target image collected by the second camera to the terminal device by the second headphone. When the cameras of both-sided headphones are set to shoot the target image, one headphone (for example, the first headphone) may be set as the main headphone, and the other headphone (for example, the second headphone) may be set as the slave headphone.
- [0067] It may be understood that the first headphone is in communication connection with the terminal device, and the second headphone is in communication connection with the terminal device. Then the first headphone and the second headphone may separately send target images to the terminal device.
- [0068] If the first headphone side carries a microphone, the first headphone sends the target image collected by the first camera and the first inquiry voice information collected by the first headphone to the terminal device, and the second headphone sends the target image collected by the second camera to the terminal device. If the second headphone side carries a microphone, the first headphone sends the target image collected by the first camera to the terminal device, and the second headphone sends the target image collected by the second camera and the second inquiry voice information collected by the second headphone to the terminal device.
- [0069] In the embodiment of the present disclosure, the first headphone and the second headphone separately communicate with the terminal device, and the connection between the headphones does not need to be considered, where the headphone carrying the microphone is the main headphone.
- [0070] In another embodiment, the sending the target image to the terminal device includes: receiving, by the first headphone, the target image collected by the second camera and sent by the second headphone; and sending, by the first headphone, the target image collected by the first camera and the target image collected by the second camera to the terminal device.
- [0071] It may be understood that the first headphone is in communication connection with the terminal device, the second headphone is not in communication connection with the terminal device, and the first

headphone and the second headphone are in communication connection with each other. In this embodiment, the first headphone is the main headphone and the second headphone is the slave headphone. The first headphone and the second headphone are communicatively connected by Bluetooth. In this case, the second headphone sends the target image collected by the second camera to the first headphone, and the first headphone sends the target image collected by the first camera and the target image collected by the second camera to the terminal device.

[0072] Further, in the case that the first headphone is in communication connection with the terminal device, the second headphone is not in communication connection with the terminal device, and the first headphone and the second headphone are communicatively connected to each other, if the first headphone carries a microphone, the first headphone collects the first inquiry voice information, and the first headphone sends the first inquiry voice information to the terminal device. If the second headphone carries a microphone, the second headphone collects the first inquiry voice information, the second headphone sends the first inquiry voice information to the first headphone, and the first headphone sends the first inquiry voice information to the terminal device.

[0073] In the embodiment of the present disclosure, after receiving the target image sent by the second headphone, the first headphone then sends the target image to the terminal device, and only one headphone needs to establish a communication connection with the terminal device. The link for transmitting data from one headphone to the terminal device is short.

[0074] Further, in the present disclosure, the target image shot by the camera of the main headphone may be sent to the terminal device first, and then the target image shot by the camera of the slave headphone may be sent to the terminal device, or they may be sent at the same time, which is not limited here.

[0075] In addition, the first inquiry voice information is encoded and then sent to the terminal device, for example, the first inquiry voice information is encoded by an OPUS encoding (a voice encoding mode). After receiving the encoded first inquiry voice information, the terminal device decodes the first inquiry voice information by an OPUS decoding (a voice decoding mode), and then inputs it into a question answering model for processing.

[0076] S304, receiving first reply voice information sent by the terminal device.

[0077] In an embodiment, the first reply voice information sent by the terminal device is received by the first headphone, and the first reply voice information is forwarded to the second headphone by the first headphone.

[0078] In the case that only the first headphone is communicatively connected to the terminal device, and the first headphone and the second headphone are communicatively connected to each other, the terminal device sends the first reply voice information to the first headphone, and the first headphone then forwards the first reply voice information to the second headphone.

- [0079] In another embodiment, the first reply voice information sent by the terminal device is received by the first headphone, and the first reply voice information sent by the terminal device is received by the second headphone.
- [0080] In the case that the first headphone and the second headphone respectively establish a communication connection with the terminal device, the terminal device sends the first reply voice information to the first headphone and the second headphone respectively for playing.
- [0081] S305, playing the first reply voice information.
- [0082] In the embodiment of the present disclosure, the terminal device may encode the first reply voice information and then send it to the headphone device. After receiving the encoded first reply voice information, the headphone device decodes it and then plays the first reply voice information.
- [0083] The headphone device plays the first reply voice information through a speaker.
- [0084] S306, collecting second inquiry voice information.
- [0085] It may be understood that the user may continue to inquire about the first reply voice information played by the headphone device, thus implementing a multi-round conversation.
- [0086] S307, obtaining second reply voice information of the second inquiry voice information.
- [0087] The second reply voice information is information related to the target image, and the target image and the second inquiry voice information are processed by the dialogue model to obtain the second reply voice information.
- [0088] In the embodiment of the present disclosure, the user may implement a polling conversation for the target image, for example, if the first inquiry voice information is "What's in front?", and the target image includes a corresponding "flower", the first reply voice information may be "There is a pot of flower in front", the second inquiry voice information is "the name of this flower", and then the second inquiry voice information and the target image are input into the dialogue model for processing to obtain the second reply voice information.
- [0089] For the determination manner of the second reply voice information, please refer to the first reply voice information, which will not be repeated here.
- [0090] S308, playing the second reply voice information.
- [0091] In an embodiment, after the multi-round conversation for the target image ends, the headphone device may also collect other inquiry voice information and other images to start a new round of conversation.
- [0092] In the embodiment of the present disclosure, through the communication connection between the headphone device and the terminal device, the headphone device may be given the function of intelligent conversation, the efficiency of implementing the intelligent conversation is improved, and the user experience is improved.
- [0093] Fig. 4 is a schematic structural diagram of a headphone device provided by an embodiment of the present disclosure, and the headphone device carries a camera. The headphone device 40 may include the

following units: a collecting unit 41, configured to collect first inquiry voice information; an obtaining unit 42, configured to obtain a target image collected by the camera; a processing unit 43, configured to obtain first reply voice information of the first inquiry voice information, where the first reply voice information is information related to the target image, and the target image and the first inquiry voice information are processed by a pre-trained dialogue model to obtain the first reply voice information; and a playing unit 44, configured to play the first reply voice information.

[0094] In some embodiments, the headphone device further includes: a waking up unit (not shown), configured to, before the target image collected by the camera is obtained, wake up the camera in response to a wake-up operation of the camera. The obtaining unit 42 is further configured to control the woken-up camera to collect the target image.

[0095] In some embodiments, the obtaining unit 42 is further configured to determine a shooting intention of the first inquiry voice information by a recognition model, and control the camera to collect the target image according to the shooting intention.

[0096] In some embodiments, a dialogue model is deployed in the headphone device, and the processing unit 43 is further configured to input the target image and the first inquiry voice information into the pre-trained dialogue model for processing to obtain the first reply voice information of the first inquiry voice information.

[0097] In some embodiments, the processing unit 43 is further configured to send the target image and the first inquiry voice information to a terminal device, the terminal device is configured to process the target image and the first inquiry voice information by a pre-trained dialogue model to obtain the first reply voice information of the first inquiry voice information; and receive the first reply voice information sent by the terminal device.

[0098] In some embodiments, the headphone device includes: a first headphone and a second headphone, the first headphone carries a first camera, the second headphone carries a second camera, and the target image includes: a target image collected by the first camera and a target image collected by the second camera.

[0099] In some embodiments, when sending the target image to the terminal device, the processing unit 43 is specifically configured to: send the target image collected by the first camera to the terminal device by the first headphone; and send the target image collected by the second camera to the terminal device by the second headphone.

[0100] In some embodiments, when sending the target image to the terminal device, the processing unit 43 is specifically configured to: receive, by the first headphone, the target image collected by the second camera and sent by the second headphone; and send, by the first headphone, the target image collected by the first camera and the target image collected by the second camera to the terminal device.

[0101] In some embodiments, when receiving the first reply voice information sent by the terminal device, the processing unit 43 is specifically configured to: receive the first reply voice information sent by the

terminal device by the first headphone; and forward the first reply voice information to the second headphone by the first headphone.

[0102] In some embodiments, when receiving the first reply voice information sent by the terminal device, the processing unit 43 is specifically configured to: receive the first reply voice information sent by the terminal device by the first headphone; and receive the first reply voice information sent by the terminal device by the second headphone.

[0103] In some embodiments, after the first reply voice information is played, the collecting unit 41 is further configured to collect second inquiry voice information; the processing unit 43 is further configured to obtain second reply voice information of the second inquiry voice information, where the second reply voice information is information related to the target image, and the target image and the second inquiry voice information are processed by the dialogue model to obtain the second reply voice information; and the playing unit 44 is further configured to play the second reply voice information.

[0104] In order to implement the above embodiments, an embodiment of the present disclosure further provides a voice reply system, including the above headphone device and a terminal device. The headphone device is configured to collect first inquiry voice information, acquire a target image collected by a camera, and send the target image and the first inquiry voice information to a terminal device. The terminal device is configured to process the target image and the first inquiry voice information through a pre-trained dialogue model to obtain first reply voice information of the first inquiry voice information, and send the first reply voice information to the headphone device. The headphone device is configured to receive the first reply voice information sent by the terminal device, and play the first reply voice information.

[0105] In order to implement the above embodiments, an embodiment of the present disclosure further provides a computer-readable storage medium, where a computer-executable instruction is stored in the computer-readable storage medium, and when the processor executes the computer-executable instruction, the voice reply method according to any one of the above embodiments is implemented.

[0106] In order to implement the above embodiments, an embodiment of the present disclosure further provides a computer program product, including a computer program, where when the computer program is executed by a processor, the voice reply method according to any one of the above embodiments is implemented.

[0107] In order to implement the above embodiments, an embodiment of the present disclosure further provides an electronic device, including: a processor and a memory; the memory stores a computer-executable instruction; and the processor executes the computer-executable instruction stored in the memory to cause the processor to execute the voice reply method according to any one of the above embodiments.

[0108] Fig. 5 is a schematic structural diagram of an electronic device provided by an embodiment of the present disclosure, and the electronic device 50 may be a terminal device or a server. The terminal device may include, but is not limited to, a mobile terminal such as a mobile phone, a notebook computer, a digital broadcast receiver, a personal digital assistant (abbreviated as PDA), a tablet computer, a portable media

player (abbreviated as PMP), an in-vehicle terminal (such as an in-vehicle navigation terminal), etc., and a stationary terminal such as a digital TV, a desktop computer, etc. The electronic device shown in Fig. 5 is only an example, and should not impose any limitation on the function and scope of use of the embodiments of the present disclosure.

[0109] As shown in Fig. 5, the electronic device 50 may include a processing apparatus (such as a central processing unit, a graphics processor, etc.) 51, which may perform various appropriate actions and processing according to a program stored in a read-only memory (abbreviated as ROM) 52 or a program loaded from a storage apparatus 58 into a random access memory (abbreviated as RAM) 53. In the RAM 53, various programs and data required for operations of the electronic device 50 are also stored. The processing apparatus 51, the ROM 52 and the RAM 53 are connected to each other through a bus 54. An input/output (I/O) interface 55 is also connected to the bus 54.

[0110] Usually, the following apparatuses may be connected to the I/O interface 55: an input apparatus 56 including, for example, a touch screen, a touchpad, a keyboard, a mouse, a camera, a microphone, an accelerometer, a gyroscope, etc.; an output apparatus 57 including, for example, a liquid crystal display (abbreviated as LCD), a speaker, a vibrator, etc.; the storage apparatus 58 including, for example, a magnetic tape, a hard disk, etc.; and a communication apparatus 59. The communication apparatus 59 may allow the electronic device 50 to communicate wirelessly or wiredly with other devices to exchange data. Although Fig. 5 shows the electronic device 50 having various apparatuses, it should be understood that not all of the apparatuses shown here need to be implemented or provided. More or fewer apparatuses may be implemented or provided alternatively.

[0111] In particular, according to the embodiments of the present disclosure, the process described above with reference to the flowchart may be implemented as a computer software program. For example, the embodiments of the present disclosure include a computer program product, which includes a computer program carried on a computer-readable medium, and the computer program contains program code for performing the method shown in the flowchart. In such an embodiment, the computer program may be downloaded and installed from the network through the communication apparatus 59, or installed from the storage apparatus 58, or installed from the ROM 52. When the computer program is executed by the processing apparatus 51, the above-mentioned functions defined in the method of the embodiments of the present disclosure are executed.

[0112] It should be noted that the above computer-readable medium in the present disclosure may be a computer-readable signal medium, a computer-readable storage medium, or any combination of the two. The computer-readable storage medium may be, for example, but not limited to, an electrical, magnetic, optical, electromagnetic, infrared, or semiconductor system, apparatus or device, or any combination of the above. More specific examples of the computer-readable storage medium may include, but are not limited to: an electrical connection having one or more wires, a portable computer magnetic disk, a hard disk, a random access memory (RAM), a read-only memory (ROM), an erasable programmable read-only memory

(EPROM or flash memory), an optical fiber, a portable compact disc read-only memory (CD-ROM), an optical storage device, a magnetic storage device, or any suitable combination of the above. In the present disclosure, the computer-readable storage medium may be any tangible medium that contains or stores a program, and the program may be used by or in combination with an instruction execution system, apparatus or device. In the present disclosure, the computer-readable signal medium may include a data signal propagated on a baseband or as a part of a carrier wave, and computer-readable program code is carried in the data signal. This propagated data signal may take many forms, including but not limited to an electromagnetic signal, an optical signal, or any suitable combination of the above. The computer-readable signal medium may also be any computer-readable medium other than the computer-readable storage medium, and the computer-readable signal medium may send, propagate, or transmit the program used by or in combination with the instruction execution system, apparatus or device. The program code contained on the computer-readable medium may be transmitted by any suitable medium, including but not limited to: a wire, an optical cable, an RF (radio frequency), etc., or any suitable combination of the above.

[0113] The above computer-readable medium may be included in the above electronic device, or may exist alone without being assembled into the electronic device.

[0114] The above computer-readable medium carries one or more programs, and when the one or more programs are executed by the electronic device, the electronic device is caused to execute the method shown in the above embodiments.

[0115] The computer program code for performing the operations of the present disclosure may be written in one or more programming languages or a combination thereof. The above programming languages include object-oriented programming languages such as Java, Smalltalk, C++, and include conventional procedural programming languages such as "C" language or similar programming languages. The program code may be completely executed on a target user computer, partially executed on a target user computer, executed as an independent software package, partially executed on a target user computer and partially executed on a remote computer, or completely executed on a remote computer or server. In the case of involving a remote computer, the remote computer may be connected to the target user computer through any kind of network, including a local area network (abbreviated as LAN) or a wide area network (abbreviated as WAN), or it may be connected to an external computer (for example, connected by using Internet provided by an Internet service provider).

[0116] The flowcharts and block diagrams in the drawings illustrate the architecture, functionality, and operation of possible implementations of the systems, methods and computer program products according to various embodiments of the present disclosure. In this regard, each block in the flowchart or block diagrams may represent a module, program segment, or portion of code, which includes one or more executable instructions for implementing the specified logical function(s). It should also be noted that, in some alternative implementations, the functions noted in the block may occur in an order different from that noted in the drawings. For example, two blocks shown in succession may, in fact, be executed substantially

concurrently, or the blocks may sometimes be executed in a reverse order, depending upon the functionality involved. It will also be noted that each block of the block diagrams and/or flowchart, and combinations of blocks in the block diagrams and/or flowchart, may be implemented by special purpose hardware-based systems that perform the specified functions or operations, or combinations of special purpose hardware and computer instructions.

[0117] The units involved in the embodiments of the present disclosure may be implemented by software or hardware. Among them, the name of the unit does not constitute a limitation on the unit itself in some cases, for example, the first obtaining unit may also be described as "a unit for obtaining at least two internet protocol addresses".

[0118] The functions described above herein may be performed at least in part by one or more hardware logic components. For example, without limitation, exemplary types of hardware logic components that may be used include: field programmable gate array (FPGA), application specific integrated circuit (ASIC), application specific standard product (ASSP), system on chip (SOC), complex programmable logic device (CPLD) and more.

[0119] In the context of the present disclosure, a machine-readable medium may be a tangible medium that may contain or store a program for use by or in combination with an instruction execution system, apparatus or device. The machine-readable medium may be a machine-readable signal medium or a machine-readable storage medium. The machine-readable medium may include, but is not limited to, an electronic, magnetic, optical, electromagnetic, infrared, or semiconductor system, apparatus or device, or any suitable combination of the above. More specific examples of the machine-readable storage medium will include an electrical connection based on one or more wires, 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), an optical fiber, a portable compact disk read-only memory (CD-ROM), an optical storage device, a magnetic storage device, or any suitable combination of the above.

[0120] In a first aspect, according to one or more embodiments of the present disclosure, a voice reply method is provided, which is applied to a headphone device carrying a camera, where the method includes: collecting first inquiry voice information; obtaining a target image collected by the camera; obtaining first reply voice information of the first inquiry voice information, where the first reply voice information is information related to the target image, and the target image and the first inquiry voice information are processed by a pre-trained dialogue model to obtain the first reply voice information; and playing the first reply voice information.

[0121] According to one or more embodiments of the present disclosure, before the obtaining the target image collected by the camera, the method further includes: waking up the camera in response to a wake-up operation of the camera; the obtaining the target image collected by the camera includes: controlling the woken-up camera to collect the target image.

- [0122] According to one or more embodiments of the present disclosure, the obtaining the target image collected by the camera includes: determining a shooting intention of the first inquiry voice information by a recognition model; and controlling the camera to collect the target image according to the shooting intention.
- [0123] According to one or more embodiments of the present disclosure, a dialogue model is deployed in the headphone device, and the obtaining the first reply voice information of the first inquiry voice information includes: inputting the target image and the first inquiry voice information into the pre-trained dialogue model for processing to obtain the first reply voice information of the first inquiry voice information.
- [0124] According to one or more embodiments of the present disclosure, the obtaining the first reply voice information of the first inquiry voice information includes: sending the target image and the first inquiry voice information to a terminal device, the terminal device is configured to process the target image and the first inquiry voice information by the pre-trained dialogue model to obtain the first reply voice information of the first inquiry voice information; and receiving the first reply voice information sent by the terminal device.
- [0125] According to one or more embodiments of the present disclosure, the headphone device includes: a first headphone and a second headphone, the first headphone carries a first camera, the second headphone carries a second camera, and the target image includes: a target image collected by the first camera and a target image collected by the second camera.
- [0126] According to one or more embodiments of the present disclosure, the sending the target image to the terminal device includes: sending the target image collected by the first camera to the terminal device by the first headphone; and sending the target image collected by the second camera to the terminal device by the second headphone.
- [0127] According to one or more embodiments of the present disclosure, the sending the target image to the terminal device includes: receiving, by the first headphone, the target image collected by the second camera and sent by the second headphone; and sending, by the first headphone, the target image collected by the first camera and the target image collected by the second camera to the terminal device.
- [0128] According to one or more embodiments of the present disclosure, the receiving the first reply voice information sent by the terminal device includes: receiving the first reply voice information sent by the terminal device by the first headphone; and forwarding the first reply voice information to the second headphone by the first headphone.
- [0129] According to one or more embodiments of the present disclosure, the receiving the first reply voice information sent by the terminal device includes: receiving the first reply voice information sent by the terminal device by the first headphone; and receiving the first reply voice information sent by the terminal device by the second headphone.

[0130] According to one or more embodiments of the present disclosure, after the playing the first reply voice information, the method further includes: collecting second inquiry voice information; obtaining second reply voice information of the second inquiry voice information, where the second reply voice information is information related to the target image, and the target image and the second inquiry voice information are processed by the dialogue model to obtain the second reply voice information; and playing the second reply voice information.

[0131] In a second aspect, according to one or more embodiments of the present disclosure, a headphone device is provided, the headphone device carries a camera, and the headphone device includes: a collecting unit, configured to collect first inquiry voice information; an obtaining unit, configured to obtain a target image collected by the camera; a processing unit, configured to obtain first reply voice information of the first inquiry voice information, where the first reply voice information is information related to the target image, and the target image and the first inquiry voice information are processed by a pre-trained dialogue model to obtain the first reply voice information; and a playing unit, configured to play the first reply voice information.

[0132] In a third aspect, according to one or more embodiments of the present disclosure, a voice reply system is provided, which includes the headphone device of the second aspect and a terminal device, where: the headphone device is configured to: collect first inquiry voice information; obtain a target image collected by a camera, and send the target image and the first inquiry voice information to the terminal device; the terminal device is configured to process the target image and the first inquiry voice information by a pre-trained dialogue model to obtain first reply voice information of the first inquiry voice information, and send the first reply voice information to the headphone device; and the headphone device is configured to receive the first reply voice information sent by the terminal device, and play the first reply voice information.

[0133] In a fourth aspect, according to one or more embodiments of the present disclosure, an electronic device is provided, which includes: at least one processor and a memory; the memory stores a computer-executable instruction; and the at least one processor executes the computer-executable instruction stored in the memory to cause the at least one processor to execute the voice reply method according to the first aspect above and various possible designs of the first aspect.

[0134] In a fifth aspect, according to one or more embodiments of the present disclosure, a computer-readable storage medium is provided, where a computer-executable instruction is stored in the computer-readable storage medium, and when the processor executes the computer-executable instruction, the voice reply method according to the first aspect above and various possible designs of the first aspect is implemented.

[0135] In a sixth aspect, according to one or more embodiments of the present disclosure, a computer program product is provided, which includes a computer program, where when the computer program is

executed by a processor, the voice reply method according to the first aspect above and various possible designs of the first aspect is implemented.

[0136] The above description is only preferred embodiments of the present disclosure and an explanation of the technical principles employed. Those skilled in the art should understand that the scope of disclosure involved in the present disclosure is not limited to the technical solutions formed by the specific combination of the above technical features, and should also cover other technical solutions formed by any combination of the above technical features or their equivalent features without departing from the above disclosure concept. For example, a technical solution formed by replacing the above features with technical features with similar functions disclosed in the present disclosure (but not limited to).

[0137] In addition, although the operations have been described in a particular order, this should not be understood as requiring the operations to be performed in the particular order shown or in a sequential order. Under certain circumstances, multitasking and parallel processing may be advantageous. Likewise, although several specific implementation details are included in the above discussion, these should not be interpreted as limiting the scope of the present disclosure. Certain features described in the context of separate embodiments may also be implemented in combination in a single embodiment. Conversely, various features described in the context of a single embodiment may also be implemented in multiple embodiments individually or in any suitable sub-combination.

[0138] Although the subject matter has been described in a language specific to structural features and/or logical actions of the method, it should be understood that the subject matter defined in the appended claims is not necessarily limited to the specific features or actions described above. Conversely, the specific features and actions described above are only exemplary forms of implementing the claims.

CLAIMS

1. A headphone device, comprising:
a first headphone and a second headphone,
wherein the first headphone carries a camera and/or the second headphone carries a camera.
2. The headphone device of claim 1, wherein the headphone device comprises one of an in-ear headphone, a semi-in-ear headphone, an over-ear headphone, a wired headphone, and a true wireless headphone.
3. The headphone device of claim 1, wherein the headphone device is configured to collect a target image, the first headphone carries a first camera, and the second headphone carries a second camera; and
the target image comprises a target image collected by the first camera and a target image collected by the second camera.
4. The headphone device of claim 3, wherein the first headphone is configured to send the target image collected by the first camera to a terminal device, and the second headphone is configured to send an image collected by the second camera to the terminal device.
5. The headphone device of claim 3, wherein the first headphone is configured to receive the target image collected by the second camera and sent by the second headphone, and send the target image collected by the first camera and the target image collected by the second camera to a terminal device.
6. The headphone device of claim 1, wherein the headphone device is configured to collect a target image, the first headphone carries a first camera, and the target image comprises a target image collected by the first camera; or the second headphone carries a second camera, and the target image comprises a target image collected by the second camera.
7. The headphone device of claim 3 or 6, wherein the headphone device further comprises a microphone, the microphone is capable of collecting first inquiry voice information, and the target image and the first inquiry voice information are input by the headphone into a dialogue model that is pre-trained for processing.
8. The headphone device of claim 7, wherein the dialogue model is deployed on the headphone device;
or

a dialogue application is installed on the headphone device, the dialogue model is deployed in a cloud server, and the headphone device uploads the target image and the first inquiry voice information to the dialogue model of the cloud server through the dialogue application for processing.

9. The headphone device of claim 7, wherein the headphone device is in communication connection with a terminal device, and the terminal device is deployed with the dialogue model; or,

the terminal device is installed with a dialogue application, the dialogue model is deployed in a cloud server, and the terminal device uploads the target image and the first inquiry voice information to the dialogue model of the cloud server through the dialogue application for processing.

10. A voice reply method, wherein the voice reply method is applied to a headphone device carrying a camera, and the voice reply method comprises:

collecting first inquiry voice information;

obtaining a target image collected by the camera;

obtaining first reply voice information of the first inquiry voice information, wherein the first reply voice information is information related to the target image, and the target image and the first inquiry voice information are processed by a dialogue model that is pre-trained to obtain the first reply voice information; and

playing the first reply voice information.

11. The voice reply method of claim 10, wherein before obtaining the target image collected by the camera, the voice reply method further comprises:

waking up the camera in response to a wake-up operation of the camera; and

obtaining the target image collected by the camera comprises:

controlling the camera that is woken up to collect the target image.

12. The voice reply method of claim 10, wherein obtaining the target image collected by the camera comprises:

determining a shooting intention of the first inquiry voice information by a recognition model; and

controlling the camera to collect the target image according to the shooting intention.

13. The voice reply method of claim 10, wherein the dialogue model is deployed in the headphone device, and obtaining the first reply voice information of the first inquiry voice information comprises:

inputting the target image and the first inquiry voice information into the pre-trained dialogue model for processing to obtain the first reply voice information of the first inquiry voice information.

14. The voice reply method of claim 10, wherein obtaining the first reply voice information of the first inquiry voice information comprises:

 sending the target image and the first inquiry voice information to a terminal device, the terminal device is configured to process the target image and the first inquiry voice information by a pre-trained dialogue model to obtain the first reply voice information of the first inquiry voice information; and

 receiving the first reply voice information sent by the terminal device.

15. The voice reply method of any of claims 10 to 14, wherein the headphone device comprises: a first headphone and a second headphone, the first headphone carries a first camera, and the second headphone carries a second camera; and

 the target image comprises a target image collected by the first camera and a target image collected by the second camera.

16. The voice reply method of claim 15, wherein sending the target image to the terminal device comprises:

 sending the target image collected by the first camera to the terminal device by the first headphone; and

 sending the target image collected by the second camera to the terminal device by the second headphone.

17. The voice reply method of claim 15, wherein sending the target image to the terminal device comprises:

 receiving, by the first headphone, the target image collected by the second camera and sent by the second headphone; and

 sending, by the first headphone, the target image collected by the first camera and the target image collected by the second camera to the terminal device.

18. The voice reply method of claim 15, wherein receiving the first reply voice information sent by the terminal device comprises:

 receiving the first reply voice information sent by the terminal device by the first headphone; and
 forwarding the first reply voice information to the second headphone by the first headphone.

19. The voice reply method of claim 15, wherein receiving the first reply voice information sent by the terminal device comprises:

 receiving, by the first headphone, the first reply voice information sent by the terminal device; and
 receiving, by the second headphone, the first reply voice information sent by the terminal device.

20. The voice reply method of any of claims 10 to 14, wherein after playing the first reply voice information, the voice reply method further comprises:

collecting second inquiry voice information;

obtaining second reply voice information of the second inquiry voice information, wherein the second reply voice information is information related to the target image, and the target image and the second inquiry voice information are processed by the dialogue model to obtain the second reply voice information; and

playing the second reply voice information.

21. A headphone device, wherein the headphone device carries a camera, and the headphone device comprises:

a collecting unit, configured to collect first inquiry voice information;

an obtaining unit, configured to obtain a target image collected by the camera;

a processing unit, configured to obtain first reply voice information of the first inquiry voice information, wherein the first reply voice information is information related to the target image, and the target image and the first inquiry voice information are processed by a pre-trained dialogue model to obtain the first reply voice information; and

a playing unit, configured to play the first reply voice information.

22. A voice reply system, characterized by comprising the headphone device of claim 21 and a terminal device, wherein:

the headphone device is configured to: collect first inquiry voice information; obtain a target image collected by the camera, and send the target image and the first inquiry voice information to the terminal device;

the terminal device is configured to process the target image and the first inquiry voice information by a pre-trained dialogue model to obtain first reply voice information of the first inquiry voice information, and send the first reply voice information to the headphone device; and

the headphone device is configured to receive the first reply voice information sent by the terminal device, and play the first reply voice information.

23. An electronic device, comprising:

a processor and a memory,

wherein the memory stores a computer-executable instruction; and

the processor executes the computer-executable instruction stored in the memory to cause the processor to execute the voice reply method of any of claims 10 to 20.

24. A computer-readable storage medium, wherein a computer-executable instruction is stored in the computer-readable storage medium, and when a processor executes the computer-executable instruction, the voice reply method of any of claims 10 to 20 is implemented.

25. A computer program product, comprising a computer program, wherein when the computer program is executed by a processor, the voice reply method of any of claims 10 to 20 is implemented.