WHAT IS CLAIMED IS:

1. A method for human-computer interaction, comprising:

receiving information of at least one modality of a user;

recognizing intention information of the user and an emotional characteristic of the user corresponding to the intention information based on the information of the at least one modality;

determining reply information to the user based on the intention information;

selecting an emotional characteristic of a character to be fed back to the user based on the emotional characteristic of the user; and

generating a broadcast video of an animated character image corresponding to the emotional characteristic of the character based on the emotional characteristic of the character and the reply information, comprising:

generating a reply audio based on the reply information and the emotional characteristic of the character; and

obtaining the broadcast video of the animated character image corresponding to the emotional characteristic of the character based on the reply audio, the emotional characteristic of the character, and a pre-established animated character image model, comprising:

inputting the reply audio and the emotional characteristic of the character into a trained mouth shape driving model to obtain mouth shape data outputted from the mouth shape driving model;

inputting the reply audio and the emotional characteristic of the character into a trained expression driving model to obtain expression data outputted from the expression driving model;

driving the animated character image model based on the mouth shape data and the expression data to obtain a three-dimensional model action sequence;

rendering the three-dimensional model action sequence to obtain a video frame picture sequence; and

synthesizing the video frame picture sequence to obtain the broadcast video of the animated character image corresponding to the emotional characteristic of the character,

wherein the mouth shape driving model and the expression driving model are trained based

on a pre-annotated audio of a same person and audio emotion information obtained from the audio.

2. The method according to claim 1, wherein

the information of the at least one modality comprises image data and audio data of the user, and

the recognizing the intention information of the user and the emotional characteristic of the user corresponding to the intention information based on the information of the at least one modality comprises:

recognizing an expression characteristic of the user based on the image data of the user;

obtaining text information from the audio data;

extracting the intention information of the user based on the text information; and

obtaining the emotional characteristic of the user corresponding to the intention information based on the audio data and the expression characteristic.

3. The method according to claim 2, wherein the recognizing the intention information of the user and the emotional characteristic of the user corresponding to the intention information based on the information of the at least one modality further comprises:

obtaining the emotional characteristic of the user further from the text information.

4. The method according to claim 2, wherein the obtaining the emotional characteristic of the user corresponding to the intention information based on the audio data and the expression characteristic comprises:

inputting the audio data into a trained speech emotion recognition model to obtain a speech emotion characteristic outputted from the speech emotion recognition model;

inputting the expression characteristic into a trained expression emotion recognition model to obtain an expression emotion characteristic outputted from the expression emotion recognition model; and

performing weighted summation on the speech emotion characteristic and the expression emotion characteristic to obtain the emotional characteristic of the user corresponding to the intention information.

5. The method according to claim 1, wherein the information of the at least one modality comprises image data and text data of the user; and

the recognizing the intention information of the user and the emotional characteristic of the user corresponding to the intention information based on the information of the at least one modality comprises:

recognizing an expression characteristic of the user based on the image data of the user;

extracting the intention information of the user based on the text data; and

obtaining the emotional characteristic of the user corresponding to the intention information based on the text data and the expression characteristic.

6. An apparatus for human-computer interaction, comprising:

a receiving unit configured to receive information of at least one modality of a user;

a recognition unit configured to recognize intention information of the user and an emotional characteristic of the user corresponding to the intention information based on the information of the at least one modality;

a determination unit configured to determine reply information to the user based on the intention information;

a selection unit configured to select an emotional characteristic of a character to be fed back to the user based on the emotional characteristic of the user; and

a broadcasting unit configured to generate a broadcast video of an animated character image corresponding to the emotional characteristic of the character based on the emotional characteristic of the character and the reply information,

wherein the broadcasting unit is configured to: generate a reply audio based on the reply information and the emotional characteristic of the character;

wherein the broadcasting unit further comprises:

a video obtaining subunit configured to obtain the broadcast video of the animated character image corresponding to the emotional characteristic of the character based on the reply audio, the emotional characteristic of the character, and a pre-established animated character image model,

wherein the video obtaining subunit comprises:

a mouth shape driving module, configured to input the reply audio and the emotional characteristic of the character into a trained mouth shape driving model to obtain mouth shape data outputted from the mouth shape driving model,

an expression driving module, configured to input the reply audio and the emotional characteristic of the character into a trained expression driving model to obtain expression data outputted from the expression driving model,

a model driving module, configured to drive the animated character image model based on the mouth shape data and the expression data to obtain a three-dimensional model action sequence,

a picture obtaining module, configured to render the three-dimensional model action sequence to obtain a video frame picture sequence,

and a video obtaining module, configured to synthesize the video frame picture sequence to obtain the broadcast video of the animated character image corresponding to the emotional characteristic of the character;

wherein the mouth shape driving model and the expression driving model are trained based on a pre-annotated audio of the same person and audio emotion information obtained from the audio.

7. A system for human-computer interaction, comprising: a collection device, a display device, and an interaction platform connected to the collection device and the display device respectively; wherein

the collection device is configured to collect information of at least one modality of a user;

the interaction platform is configured to receive the information of the at least one modality of the user; recognize intention information of the user and an emotional characteristic of the user corresponding to the intention information based on the information of the at least one modality; determine reply information to the user based on the intention information; select an emotional characteristic of a character to be fed back to the user based on the emotional characteristic of the user; and generate a broadcast video of an animated character image corresponding to the emotional characteristic of the character based on the emotional characteristic of the character and the reply information, comprising:

generating a reply audio based on the reply information and the emotional characteristic of the character; and

obtaining the broadcast video of the animated character image corresponding to the emotional characteristic of the character based on the reply audio, the emotional characteristic

of the character, and a pre-established animated character image model, comprising:

inputting the reply audio and the emotional characteristic of the character into a trained mouth shape driving model to obtain mouth shape data outputted from the mouth shape driving model;

inputting the reply audio and the emotional characteristic of the character into a trained expression driving model to obtain expression data outputted from the expression driving model;

driving the animated character image model based on the mouth shape data and the expression data to obtain a three-dimensional model action sequence;

rendering the three-dimensional model action sequence to obtain a video frame picture sequence; and

synthesizing the video frame picture sequence to obtain the broadcast video of the animated character image corresponding to the emotional characteristic of the character,

wherein the mouth shape driving model and the expression driving model are trained based on a pre-annotated audio of a same person and audio emotion information obtained from the audio; and

the display device is configured to receive and play the broadcast video.

8. An electronic device, comprising:

one or more processors; and

a storage apparatus storing one or more programs thereon,

wherein the one or more programs, when executed by the one or more processors, cause the one or more processors to implement the method according to any one of claims 1-5.

9. A computer-readable medium, storing a computer program thereon, wherein the program, when executed by a processor, implements the method according to any one of claims 1-5.

10. A computer program product, comprising a computer program, wherein the computer program, when executed by a processor, implements the method according to any one of claims 1-5.