

POSE GENERATION METHOD AND APPARATUS, ELECTRONIC DEVICE AND STORAGE MEDIUM

TECHNICAL FIELD

[0001] The disclosure relates to the technical field of computer information, and further relates to, but is not limited to, a pose generation method and apparatus, an electronic device and a storage medium.

BACKGROUND

[0002] Currently, in some scenarios, there is a need to simulate poses based on music. For example, in virtual scenarios such as games or animations, poses such as dance or rhythmic gymnastics are simulated following the currently playing music. However, in the related art, according to the given music, the generated poses are inconsistent with the rhythm of the music, and the generated poses are not smooth, so that it is difficult to produce satisfactory poses.

SUMMARY

[0003] The disclosure provides a pose generation method and apparatus, an electronic device and a storage medium.

[0004] According to a first aspect of an embodiment of the disclosure, a pose generation method is provided. The method includes the following operations. A temporally associated predicted pose feature sequence is acquired, the predicted pose feature sequence being generated at least according to an audio feature. A quantized pose feature matching each predicted pose feature in the predicted pose feature sequence is determined from a preset pose feature library. A spatial association relationship of respective pose features in the predicted pose feature sequence is determined. The

quantized pose feature matching the predicted pose feature is combined according to the spatial association relationship of the respective pose features and a temporal association relationship of the respective pose features, to obtain a target pose feature sequence.

[0005] According to a second aspect of an embodiment of the disclosure, a pose generation apparatus is provided. The apparatus includes an acquisition module, a first determination module, a second determination module, and a combination module. The acquisition module is configured to acquire a temporally associated predicted pose feature sequence, and the predicted pose feature sequence is generated at least according to an audio feature. The first determination module is configured to determine a quantized pose feature matching each predicted pose feature in the predicted pose feature sequence from a preset pose feature library. The second determination module is configured to determine a spatial association relationship of respective pose features in the predicted pose feature sequence. The combination module is configured to combine the quantized pose feature matching each predicted pose feature according to the spatial association relationship of the respective pose features and a temporal association relationship of the respective pose features, to obtain a target pose feature sequence.

[0006] According to a third aspect of an embodiment of the disclosure, an electronic device is provided. The electronic device includes a processor and a memory configured to store a computer program executable on the processor. The processor is configured to operate the computer program to perform any of the above pose generation method.

[0007] According to a fourth aspect of an embodiment of the disclosure, a computer storage medium is provided. The computer storage medium stores a computer program. Any of the above pose generation method is implemented when the computer program is performed by a processor.

[0008] The technical solutions provided by the embodiments of the disclosure may include the following beneficial effects.

[0009] In the embodiments of the disclosure, the quantized pose features respectively matching the temporally associated predicted pose features are determined from the preset pose feature library. The quantized pose feature matching the predicted pose

feature is combined according to the spatial association relationship of the respective pose features and the temporal association relationship of the respective pose features, to obtain the target pose feature sequence.

[0010] During the generation of the target pose feature sequence, the temporal association relationship of the respective pose features and the spatial association relationship both can be taken into consideration. Therefore, the standard of the generated target pose feature sequence can be improved, and the consistency between beats and audio rhythms of the target pose feature sequence can be enhanced.

[0011] It should be understood that, the above general description and the following detailed description are merely exemplary and explanatory, and cannot limit the disclosure.

BRIEF DESCRIPTION OF THE DRAWINGS

[0012] Drawings are incorporated into the specification and constitute a part of the specification. The drawings illustrate embodiments in accordance with the disclosure and serve to understand the principles of the disclosure together with the specification.

[0013] FIG. 1 is a flowchart of a pose generation method according to an embodiment of the disclosure.

[0014] FIG. 2 is a schematic diagram of a pose generation method according to an embodiment of the disclosure.

[0015] FIG. 3 is a schematic diagram of a feature processing method according to an embodiment of the disclosure.

[0016] FIG. 4 is a schematic diagram of a pose prediction method according to an embodiment of the disclosure.

[0017] FIG. 5a is a schematic diagram I of an association degree determination method according to an embodiment of the disclosure.

[0018] FIG. 5b is a schematic diagram II of the association degree determination method according to an embodiment of the disclosure.

[0019] FIG. 5c is a schematic diagram III of the association degree determination method according to an embodiment of the disclosure.

[0020] FIG. 6 is a schematic diagram of a reward determination method according to an embodiment of the disclosure.

[0021] FIG. 7 is a schematic diagram showing a target pose according to an embodiment of the disclosure.

[0022] FIG. 8 is a schematic structural diagram of a target pose apparatus according to an embodiment of the disclosure.

[0023] FIG. 9 is a schematic structural diagram of an electronic device according to an embodiment of the disclosure.

DETAILED DESCRIPTION

[0024] The exemplary examples will be described in detail here, and examples thereof are shown in the drawings. When the following description refers to the drawings, unless otherwise indicated, the same numbers in different drawings indicate the same or similar elements. The implementations described in the following exemplary examples do not represent all implementations consistent with the disclosure. On the contrary, the implementations are merely examples of apparatuses and methods consistent with some aspects of the disclosure as detailed in the appended claims.

[0025] An embodiment of the disclosure provides a pose generation method. The method is applicable to an electronic device. Here, the electronic device may include a terminal device, such as a mobile terminal, a fixed terminal or a vehicle terminal, and the like. The mobile terminal may include devices, such as a mobile phone, a tablet computer, a notebook computer or a wearable device and the like, or may include a smart home device, such as a smart speaker. The fixed terminal may include a desktop

computer or a smart television and the like. The vehicle terminal may include a front-end device of a vehicle monitoring management system, or may further be known as a Telematics Control Unit (TCU), such as a vehicle-computer terminal. Functions implemented by the method are applicable to a computer system composed of terminals and/or servers. Here, the terminals may be thin clients, thick clients, handheld or laptop devices, systems based on microprocessors, set-top boxes, programmable consumer electronics, network personal computers, small computer systems, or the like. The servers may be server computer systems, small computer systems, large computer systems, and distributed cloud computing technology environment including any of the above systems.

[0026] FIG. 1 is a flowchart of a pose generation method according to an exemplary embodiment. As shown in FIG. 1, the method mainly includes the following steps.

[0027] At S101, a temporally associated predicted pose feature sequence is acquired. The predicted pose feature sequence is generated at least according to an audio feature.

[0028] At S102, a quantized pose feature matching each predicted pose feature in the predicted pose feature sequence is determined from a preset pose feature library.

[0029] At S103, a spatial association relationship of respective pose features in the predicted pose feature sequence is determined.

[0030] At S104, the quantized pose feature matching each predicted pose feature is combined according to the spatial association relationship of the respective pose features and a temporal association relationship of the respective pose features, to obtain a target pose feature sequence.

[0031] In an embodiment of the disclosure, the pose feature may be understood as data used to characterize pose attributes. The pose feature may include three-dimensional coordinates of each joint point of an object (for example, a virtual character) that generates poses, or a pose vector characterizing the poses. One pose feature corresponds to one pose. The pose feature may be expressed in the form of a vector or an index value. In some embodiments, the electronic device may set the corresponding index value in advance according to different poses. For example, an index value corresponding to a

hand rising pose is 1, an index value corresponding to a kicking pose is 2, an index value corresponding to a squat pose is 3, and the like, no limitation is made in the disclosure. Different pose features may represent different poses. For example, a first pose feature may represent stand at attention, a second pose feature may represent stand at ease, and a third pose feature may represent straddling.

[0032] The pose feature sequence may be understood as a set of a plurality of pose features. For example, the pose feature sequence includes three pose features. A first pose feature is to rise hands, a second pose feature is to bow, and a third pose feature is to rise feet. Temporal association may be understood as the association of the pose features based on time sequence, and the pose features in the pose feature sequence may have an association relationship in time sequence. For example, a pose represented by the first pose feature occurs before. A pose represented by the second pose feature occurs after the pose represented by the first pose feature. A pose represented by the third pose feature occurs after the pose represented by the second pose feature.

[0033] In an embodiment of the disclosure, the electronic device may at least generate the predicted pose feature sequence according to the audio feature. The audio feature may be understood as data for characterizing the audio attributes. The audio feature may include one or more of data, such as a sampling frequency, a bit rate, the number of channels, and a frame rate. The predicted pose feature sequence may be understood as a pose feature sequence corresponding to information such as the audio feature and generated according to the information such as the audio feature. The electronic device may generate the predicted pose feature sequence by using a pose prediction model. The pose prediction model may be understood as a trained neural network that obtains a corresponding predicted pose feature sequence with an audio feature as the input. For example, the electronic device may generate a predicted pose feature sequence corresponding to a virtual teacher according to an audio feature about the content of knowledge points, so that diversity of teaching content can be improved, and the interest of students in learning can be stimulated. The electronic device may further generate a predicted pose feature sequence matching music information according to an audio feature about the music information, and control a virtual anchor to display a dance pose corresponding to the predicted pose feature sequence, so that a manpower cost for

maintaining the virtual anchor can be reduced.

[0034] In a possible embodiment, the electronic device may further generate the predicted pose feature sequence according to the audio feature and a starting pose feature. The starting pose feature may be understood as a preset pose feature, and may include one or more pose features. The electronic device may further generate the predicted pose feature sequence according to data such as the audio feature and a scenario feature. The scenario feature may be understood as a usage scenario of the predicted pose feature sequence, for example, a teaching scenario, a customer service scenario, or a navigation scenario. A method that the electronic device acquires the temporally associated predicted pose feature sequence is not limited in the disclosure.

[0035] In an embodiment of the disclosure, after the electronic device acquires the predicted pose feature sequence, the quantized pose feature matching each predicted pose feature in the predicted pose feature sequence may be determined from the preset pose feature library. The preset pose feature library may be understood as a set including a plurality of quantized pose features. The electronic device may summarize and quantizing a plurality of standard and reasonable poses to limited, sparse, and indexable quantized pose features, and store the quantized pose features in the preset pose feature library.

[0036] In some embodiments, the electronic device may learn the pose feature library by using an unsupervised approach, so as to obtain the trained preset pose feature library. Each quantized pose feature in the preset pose feature library may correspond to one pose. For example, the preset pose feature library includes the quantized pose features represented by 512 different poses. The quantized pose feature may be understood as a high-dimensional feature with a sparse constraint. The electronic device may determine a matching relationship between the predicted pose feature and the quantized pose feature by determining a similarity between the predicted pose feature and the quantized pose feature. For example, there are 3 quantized pose features in the pose feature library. Through a preset similarity calculation strategy, the electronic device determines the similarity between a first predicted pose feature and the 3 quantized pose features to be respectively 0.5 (corresponding to the first quantized pose feature), 0.9 (corresponding to

the second quantized pose feature), and 0.4 (corresponding to the third quantized pose feature), and the similarity between a second predicted pose feature and the 3 quantized pose features to be respectively 0.3 (corresponding to the first quantized pose feature), 0.6 (corresponding to the second quantized pose feature), and 0.85 (corresponding to the third quantized pose feature). By using the predicted pose feature with the largest similarity matching the quantized pose feature as an example, the electronic device may determine that the first predicted pose feature matches the second quantized pose feature, and the second predicted pose feature matches the third quantized pose feature.

[0037] In a possible embodiment, the electronic device may determine the quantized pose feature matching the predicted pose feature from the preset pose feature library. A number correspondence between the predicted pose feature and the matched quantized pose feature may include the following. One predicted pose feature corresponds to one quantized pose feature, or one predicted pose feature corresponds to at least two quantized pose features. The temporal association relationship between the predicted pose features may be the same as or different from the temporal association relationship of the quantized pose features.

[0038] In an embodiment of the disclosure, after the quantized pose feature matching the predicted pose feature is determined, the spatial association relationship of respective predicted pose features in the predicted pose feature sequence may be determined. The spatial association relationship may be understood as an association relationship between positions, an association relationship between moving distances, an association relationship between moving directions or the like. For example, the spatial association relationship between the first predicted pose feature and the second predicted pose feature may be a mirror-image relationship. The electronic device may determine the spatial association relationship of the predicted pose features through the quantized pose features matching the predicted pose features.

[0039] For example, the electronic device determines a generation object of a pose represented by the first quantized pose feature by analyzing the first quantized pose feature so as to determine the three-dimensional coordinates of respective joint points of the body part of the object to be $\{(1, 2, 3), \dots\}$, and determines a generation object of a

pose represented by the second quantized pose feature by analyzing the second quantized pose feature so as to determine the three-dimensional coordinates of the respective joint points of the body part of the object to be $\{(4, 5, 6), \dots\}$. Therefore, through a relationship between the three-dimensional coordinates of the joint points, the spatial association relationship of the predicted pose features may be determined. Here, the relationship between the three-dimensional coordinates of the joint points may be understood as a relationship between same joint points in different quantized pose features. For example, three-dimensional coordinates of the elbow joint represented by the first quantized pose feature are (1, 2, 3). Three-dimensional coordinates of the elbow joint represented by the second quantized pose feature are (4, 6, 7). Therefore, the relationship between the three-dimensional coordinates of the joint points may be a relationship between the three-dimensional coordinates of the elbow joints represented by different quantized pose features.

[0040] In an embodiment of the disclosure, after the electronic device determines the spatial association relationship of the predicted pose features, the quantized pose feature matching the predicted pose feature may be combined according to the spatial association relationship of respective predicted pose features and the temporal association relationship of the respective predicted pose features, to obtain the target pose feature sequence. The target pose feature sequence may be understood as an optimal pose feature sequence to be generated by the object. The size of the target pose feature sequence may be the same or different from the size of the predicted pose feature sequence, which is not limited in the disclosure. For example, the electronic device determines the spatial association relationship (for example, a relationship between three-dimensional coordinates of joint points of a pose generation object in the predicted pose features) of respective predicted pose features, and the temporal association relationship (for example, a relationship between generation moments of respective predicted pose features) of respective predicted pose features. The generation moment of the predicted pose feature may be a moment at which the predicted pose feature is obtained.

[0041] The electronic device may combine the quantized pose feature matching the predicted pose feature. For example, the electronic device determines a generation moment of the first predicted pose feature as a first moment, determines a generation

moment of the second predicted pose feature as a second moment, and determines the first moment to be temporally adjacent to the second moment. The electronic device determines a first set of three-dimensional coordinates of joint points of the pose generation object in the first quantized pose feature and a second set of three-dimensional coordinates of joint points of the pose generation object in the second quantized pose feature by determining the first quantized pose feature matching the first predicted pose feature and the second quantized pose feature matching the second predicted pose feature. The electronic device determines information such as the spatially moving distance and the spatially moving direction between same joint points among the first set of three-dimensional coordinates and the second set of three-dimensional coordinates, so as to determine a pose of the pose generation object from the first moment to the second moment. The electronic device may control the pose generation object to stably transit from a pose at the first moment to a pose at the second moment. For example, the pose generated at the first moment (for example, the 5th second) is to rise the hand to a height of 1.1 m, and the pose generated at the second moment (for example, the 7th second) is to rise the hand to a height of 1.2 m. Then a pose generated at the 6th second may be determined as rising the hand to a height of 1.15 m.

[0042] In a possible embodiment, a pre-trained neural network may be used to determine the target pose feature sequence according to the audio feature and the starting pose feature. FIG. 2 is a schematic diagram of a pose generation method according to an exemplary embodiment. As shown in FIG. 2, music features m and starting pose features (or starting pose codes) p_0 may be inputted. The starting pose features p_0 may be composed of at least two portions, for example, a starting pose feature representing an upper half-body pose and a starting pose feature representing a lower half-body pose. The starting pose feature may be expressed by using an index number. For example, the starting pose feature is (14, 13), which may respectively express the upper half-body pose and the lower half-body pose. Then, a target pose prediction model (Actor-Critic Motion GPT) 201 is used to generate a predicted pose feature sequence (or pose code sequence). The predicted pose feature sequence may include p_0 being (14, 13), p_1 being (27, 68),..., p_k being (0, 7), and the like, where k is a positive integer. Then, according to the predicted pose feature sequence, the quantized pose feature sequence matching each

predicted pose feature in the predicted pose feature sequence may be determined from the preset pose feature library 202. The preset pose feature library may refer to a choreographic memory codebook. The preset pose feature library may include N sets of quantized pose features 0 to (N-1). Each set of quantized pose features may respectively express the upper half-body pose and the lower half-body pose. Then, a Convolutional Neural Network (CNN) decoder 203 may be used to generate the target pose feature sequence according to the quantized pose feature sequence. The target pose feature sequence may refer to generated dance.

[0043] In the embodiments of the disclosure, the quantized pose features respectively matching the temporally associated predicted pose features are determined from the preset pose feature library. The quantized pose feature matching the predicted pose feature is combined according to the spatial association relationship of the respective pose features and the temporal association relationship of the respective pose features, to obtain the target pose feature sequence. During the generation of the target pose feature sequence, the temporal association relationship of the respective pose features and the spatial association relationship both can be taken into consideration. Therefore, the standard of the generated target pose feature sequence can be improved, and the consistency between beats and audio rhythms of the target pose feature sequence can be enhanced.

[0044] In some embodiments, determining the quantized pose feature matching each predicted pose feature in the predicted pose feature sequence from the preset pose feature library includes the following operations.

[0045] Dimension reduction processing is respectively performed on each predicted pose feature to obtain a coding feature. A dimension of the encoded feature is the same as a dimension of each quantized pose feature in the preset pose feature library.

[0046] The quantized pose feature matching each predicted pose feature is determined from the preset pose feature library according to a distance between the encoded feature and each quantized pose feature in the preset pose feature library.

[0047] In the embodiments of the disclosure, the electronic device may respectively

perform dimension reduction processing on each predicted pose feature to obtain the encoded feature. The dimension of the encoded feature is the same as the dimension of each quantized pose feature in the preset pose feature library. In this way, the dimension of the encoded feature can be the same as the dimension of the quantized pose feature. Therefore, processing such as contrast and calculation can be performed. In addition, by reducing the dimension of the feature, calculation quantity of the follow-up processing can be reduced, and calculation efficiency can be enhanced. For example, the dimension of the predicted pose feature sequence may be $(T \times (J \times 3))$. T may represent the number (that is, a sequence length of the predicted pose feature sequence) of predicted pose features, J may represent the number of each joint point of the body part of the object, and 3 may represent the three-dimensional coordinates of each joint point. The dimension of the quantized pose feature sequence may be $(T' \times C)$. T' may represent the number (that is, a sequence length of the quantized pose feature sequence) of quantized pose features, and C may represent the dimension of the quantized pose feature.

[0048] In the embodiments of the disclosure, after the electronic device obtains the encoded feature, the quantized pose feature matching each predicted pose feature may be determined from the preset pose feature library according to the distance between the encoded feature and each quantized pose feature in the preset pose feature library. The distance between the encoded feature and each quantized pose feature may be understood as a similarity or an association degree between the features, and may be implemented by determining a vector distance between different features. For example, the electronic device may determine the similarity between different features by determining the vector distance, such as a Euclidean distance, a Manhattan distance, a Chebyshev distance, a Mahalanobis distance, between the encoded feature and the quantized pose feature. For example, the electronic device determines distances between a first encoded feature and the quantized pose features in the preset pose feature library to be respectively 1, 2, 3 and the like, and distances between a second encoded feature and the quantized pose features in the preset pose feature library to be respectively 3, 4, 5 and the like. By using a minimum distance as a rule that different features match each other, the first encoded feature may be determined to match the first quantized pose feature in the preset pose feature library, and the second encoded feature may be determined to match the third

quantized pose feature in the preset pose feature library.

[0049] In the embodiments of the disclosure, the encoded features of which dimensions are the same as the dimensions of the quantized pose features are obtained by respectively performing dimension reduction processing on the predicted pose features. Then, according to the distance between the encoded feature and each quantized pose feature in the preset pose feature library, the quantized pose features matching the predicted pose features can be simply and accurately determined from the preset pose feature library. Therefore, calculation efficiency can be enhanced, and the predicted pose features can be converted to pre-trained quantized pose features with sparse constraints.

[0050] In some embodiments, the determining the quantized pose feature matching each predicted pose feature from the preset pose feature library according to a distance between the encoded feature and each quantized pose feature in the preset pose feature library includes the following operations.

[0051] A distance between the to-be-processed encoded feature and each quantized pose feature in the preset pose feature library is determined.

[0052] Respective distances corresponding to the to-be-processed encoded feature are sorted to obtain a sorting result.

[0053] The quantized pose feature corresponding to the to-be-processed encoded feature is determined according to the sorting result.

[0054] The quantized pose feature matching each predicted pose feature is determined according to the quantized pose feature corresponding to each to-be-processed encoded feature.

[0055] In the embodiments of the disclosure, the electronic device may determine a distance between the to-be-processed encoded feature and each quantized pose feature in the preset pose feature library. For example, an encoded feature sequence includes 3 encoded features. The preset pose feature library includes 10 quantized pose features. The electronic device may first and respectively determine the distances between the first encoded feature and the 10 quantized pose features, such as 3, 5, 2, ..., and the like, then

determine the distances between the second encoded feature and the 10 quantized pose features, such as 7, 4, 6, ..., and the like, and finally determine the distances between the third encoded feature and the 10 quantized pose features, such as 1, 4, 3, ..., and the like.

[0056] After determining the distance between the to-be-processed encoded feature and each quantized pose feature in the preset pose feature library, the electronic device may sort the distances corresponding to the to-be-processed encoded features, and determine the quantized pose features corresponding to the to-be-processed encoded features according to the sorting result. For example, by sorting the distances of the same encoded feature based on a rule from the minimum to the maximum, the first encoded feature may be determined to correspond to (2, 3, 5, ...), the second encoded feature to correspond to (4, 6, 7, ...), the third encoded feature to correspond to (1, 3, 4, ...), and the like. By using the minimum distance as a matching condition, the first encoded feature may be determined to correspond to the third quantized pose feature in the preset pose feature library, the second encoded feature may be determined to correspond to the second quantized pose feature in the preset pose feature library, and the third encoded feature may be determined to correspond to the first quantized pose feature in the preset pose feature library.

[0057] After determining the quantized pose feature corresponding to the to-be-processed encoded feature, the electronic device may determine the quantized pose feature matching each predicted pose feature according to the quantized pose feature corresponding to each to-be-processed encoded feature. For example, the electronic device determines the first encoded feature to correspond to the third quantized pose feature, the second encoded feature to correspond to the second quantized pose feature, and the third encoded feature to correspond to the first quantized pose feature. The electronic device may determine the quantized pose feature matching the predicted pose feature according to a correspondence between the predicted pose feature and the encoded feature. The correspondence between the predicted pose feature and the encoded feature may be that one predicted pose feature corresponds to one encoded feature, or a plurality of predicted pose features correspond to one encoded feature. For example, the first predicted pose feature corresponds to the first encoded feature, and the second predicted pose feature corresponds to the second encoded feature. Then the predicted

pose feature corresponds to the third quantized pose feature, and the second predicted pose feature corresponds to the second quantized pose feature.

[0058] In the embodiments of the disclosure, by determining the distance between the to-be-processed encoded feature and each quantized pose feature in the preset pose feature library, the distances corresponding to the to-be-processed encoded features are sorted. Then, according to the sorting result, the quantized pose feature corresponding to the to-be-processed encoded feature may be determined. Therefore, the predicted pose features can be simply and accurately converted into the quantized pose features according to the quantized pose feature corresponding to each to-be-processed encoded feature.

[0059] In some embodiments, determining the spatial association relationship of respective pose features in the predicted pose feature sequence includes the following operations.

[0060] Relative position information between each key point of a target role and a preset standard point and displacement information of the target role are determined according to each quantized pose feature. The predicted pose feature sequence is generated according to the audio feature and the starting pose feature of the target role.

[0061] The spatial association relationship of the respective pose features is determined according to the relative position information and the displacement information.

[0062] In the embodiments of the disclosure, the electronic device determines the relative position information between each key point of the target role and the preset standard point and the displacement information of the target role according to each quantized pose feature. The electronic device may generate the predicted pose feature sequence according to the audio feature and the starting pose feature of the target role. The audio feature may include at least one rhythm. The starting pose feature may include at least one beat (may also be known as the rhythm). One audio rhythm may correspond to at least one pose beat. The electronic device may customize the starting pose feature, and randomly select one pose feature as the starting pose feature from a starting pose

feature set under a case that the predicted pose feature sequence is required to be generated according to the starting pose feature. The target role may be understood as the object generating poses, and may include a virtual image such as a virtual teacher. The electronic device may control the target role to display the poses according to the target pose feature sequence. The electronic device may use each of the joint points of the body part of the target role as a key point representing a pose attribute, for example, a shoulder joint, a hip joint, a knee joint, and the like. The electronic device may use moving information between the preset standard points as the displacement information of the target role. The preset standard point may be understood as any one of the key points, for example, the key point corresponding to the hip joint may be used as the preset standard point.

[0063] In a possible embodiment, the electronic device may input a first decoder through the quantized pose feature, to obtain the relative position information between each key point and the preset standard point. The relative position information may be understood as three-dimensional coordinate information between the key points. For example, the first key point is located 30 degrees from the preset standard point with a distance being 10 cm. The second key point is located 60 degrees from the preset standard point with a distance being 12 cm. The electronic device may input a second decoder through the quantized pose feature, to obtain the displacement information of the target role. The displacement information may be understood as a moving attribute between poses represented by different quantized pose features. For example, a pose represented by the second quantized pose feature is moved 20 cm leftwards relative to a pose represented by the first quantized pose feature. That is to say, the spatial association relationship of the respective pose features may be expressed as the relative position information between each key point of the target role and the preset standard point and displacement information of the target role.

[0064] In a possible embodiment, the pre-trained neural network (that is, a target feature processing model) may be used to determine the quantized pose feature matching each predicted pose feature in the predicted pose feature sequence from the preset pose feature library, and determine the spatial association relationship of the respective pose features in the predicted pose feature sequence. FIG. 3 is a schematic diagram of a

feature processing method according to an exemplary embodiment. As shown in FIG. 3, the predicted pose feature sequence may be inputted. The predicted pose feature sequence may include an upper half-body pose and a lower half-body pose. The predicted pose feature in the predicted pose feature sequence may be expressed as P . A dimension of the predicted pose feature sequence may be $T \times (J \times 3)$, where T may represent a length of the sequence, J may represent the number of joint points of the body part of the object generating the poses, and 3 may represent the three-dimensional coordinates of each joint point. A pose encoder E 301 is used to perform dimension reduction processing on the predicted pose features respectively, to obtain the encoded feature sequence. The encoded feature may be expressed as e . A dimension of the encoded feature sequence may be $T' \times C$, where T' may represent a length of the sequence, and C may represent the number of dimensions of the features. Then, according to the encoded feature sequence, the quantized pose feature sequence matching each predicted pose feature in the predicted pose feature sequence is determined from the preset pose feature library 302 (choreographic memory codebook). The quantized pose feature may be expressed as eq . The dimension of the quantized pose feature sequence may be $T' \times C$. Then, according to the quantized pose feature sequence, a pose decoder D_p 303 and a global velocity decoder D_v 304 may be respectively utilized to determine the relative position information (which may be expressed as $\hat{p}$) between each key point of the target role and the preset standard point and the displacement information (which may be expressed as $\hat{v}$) of the target role. A dimension of a vector feature of the relative position information may be $T \times (J \times 3)$, and a dimension of a vector feature of the displacement information may be $T \times 3$.

[0065] In the embodiments of the disclosure, the relative position information between each key point of a target role and a preset standard point and the displacement information of the target role are determined according to each quantized pose feature. The predicted pose feature sequence is generated according to the audio feature and the starting pose feature of the target role. Therefore, according to the relative position information and the displacement information, the spatial association relationship of the respective pose features can be rapidly and accurately determined.

- [0066] In some embodiments, the method further includes the following operations.
- [0067] A sample pose feature sequence is acquired.
- [0068] A model parameter of a starting feature processing model and the quantized pose feature in a starting pose feature library of the starting feature processing model are updated according to respective sample pose features in the sample pose feature sequence and a preset spatial association relationship of the respective sample pose features, to obtain an updating result.
- [0069] A target feature processing model having the preset pose feature library is obtained based on the updating result.
- [0070] Determining the quantized pose feature matching each predicted pose feature in the predicted pose feature sequence from the preset pose feature library includes the following operation.
- [0071] The quantized pose feature matching each predicted pose feature in the predicted pose feature sequence is determined from the preset pose feature library of the target feature processing model.
- [0072] In the embodiments of the disclosure, the sample pose feature sequence may be understood as a manually labeled pose feature sequence, and may be used for training the starting feature processing model. The model parameter in the starting feature processing model may be set randomly. During training, a loss value may be calculated according to a predefined loss function to update the model parameter. The electronic device may extract the sample pose feature sequence from Internet, or select the sample pose feature sequence from a preset dataset (such as a dance movement dataset). A preset spatial association relationship of the respective sample pose features may be understood as an association relationship between respective manually labeled sample pose features. The starting pose feature library may be understood as an untrained pose feature library. The starting pose feature library may be a part of the starting feature processing model, or may refer to two different parts, which can commonly implement the processing of the spatial association relationship of the respective pose features according to the predicted pose feature sequence. The pose features in the starting pose feature library may be

customized. During the updating of the model parameter of the starting feature processing model, the updating of the pose features is further included.

[0073] The electronic device may update the model parameter of the starting feature processing model and the quantized pose feature in the starting pose feature library of the starting feature processing model according to respective sample pose features in the sample pose feature sequence and the preset spatial association relationship of the respective sample pose features, and then obtains the target feature processing model having the preset pose feature library based on the updating result. The target feature processing model may be understood as a trained feature processing model. The preset pose feature library may be understood as a trained pose feature library. After the electronic device obtains the target feature processing model having the preset pose feature library, during actual use of the model, the quantized pose feature matching each predicted pose feature in the predicted pose feature sequence may be determined from the preset pose feature library of the target feature processing model. The spatial association relationship of the respective pose features may be determined by using the target feature processing model.

[0074] In the embodiments of the disclosure, the sample pose feature sequence is acquired. Then, the model parameter of the starting feature processing model and the quantized pose feature in the starting pose feature library of the starting feature processing model may be updated according to respective sample pose features in the sample pose feature sequence and the preset spatial association relationship of the respective sample pose features. Therefore, the starting feature processing model can be rapidly and accurately trained into the target feature processing model, and the efficiency and accuracy of model training can be enhanced.

[0075] In some embodiments, updating the model parameter of the starting feature processing model and the quantized pose feature in the starting pose feature library of the starting feature processing model according to respective sample pose features in the sample pose feature sequence and the preset spatial association relationship of the respective sample pose features, to obtain the updating result includes the following operations.

- [0076] A predicated spatial association relationship of the respective sample pose features is obtained using the starting feature processing model.
- [0077] A first loss value is determined according to the predicted spatial association relationship and the preset spatial association relationship.
- [0078] Dimension reduction processing is performed on each sample pose feature to obtain a sample encoded feature.
- [0079] A second loss value is determined according to the sample encoded feature.
- [0080] A first target loss value is obtained according to the first loss value and the second loss value.
- [0081] The model parameter of the starting feature processing model and the quantized pose feature in the starting pose feature library of the starting feature processing model are updated using first target loss value, to obtain the updating result.
- [0082] In the embodiments of the disclosure, during the training of the starting feature processing model, the electronic device may obtain the predicated spatial association relationship of the respective sample pose features using the starting feature processing model. The predicted association relationship may be understood as an association relationship obtained according to the preset model parameter. The electronic device may determine the first loss value according to the predicted spatial association relationship and the preset spatial association relationship. For example, the electronic device may determine the first loss value by means of a calculation strategy of a preset first loss function according to a difference between the predicted association relationship and the preset association relationship. The electronic device may perform dimension reduction processing on each sample pose feature to obtain the sample encoded feature. The sample encoded feature may be understood as the encoded feature obtained according to the preset model parameter. The electronic device may determine the second loss value by processing the sample encoded feature according to a calculation strategy of a second loss function. For example, the electronic device may determine a norm value corresponding to the sample encoded feature, and the obtained norm value as the second loss function.

[0083] After determining the first loss value and the second loss value, the electronic device may obtain the first target loss value according to the first loss value and the second loss value. The first target loss value may be understood as a loss value determined during the primary training of the starting feature processing model. For example, the electronic device determines the first target loss value according to weighted summation on the first loss value and the second loss value, or according to a ratio of the first loss value to the second loss value. After the first target loss value is determined, the electronic device may update the model parameter of the starting feature processing model and the pose feature in the starting pose feature library of the starting feature processing model using the first target loss value. For example, the model parameter is adjusted from t1 to t2. The pose feature is adjusted from e1 to e2.

[0084] In the embodiments of the disclosure, different loss values can be obtained according to different calculation strategies. Therefore, target loss values can be accurately determined, and the efficiency and accuracy of model training can be enhanced.

[0085] In some embodiments, determining the second loss value according to the sample encoded feature includes the following operations.

[0086] A quantized pose feature matching the sample pose feature is determined from the starting pose feature library according to the sample encoded feature.

[0087] The second loss value is determined according to the sample encoded feature and the quantized pose feature matching the sample pose feature.

[0088] In the embodiments of the disclosure, the electronic device may determine a quantized pose feature matching the sample pose feature from the starting pose feature library according to the sample encoded feature. The electronic device may determine the quantized pose feature matching the sample pose feature by determining a distance between the sample encoded feature and the quantized pose feature in the starting pose feature library. For example, the first sample encoded feature matches the first quantized pose feature in the starting pose feature library. The second sample encoded feature matches the second quantized pose feature in the starting pose feature library.

[0089] After the quantized pose feature matching the sample pose feature is determined from the starting pose feature library, the electronic device may determine the second loss value according to the sample encoded feature and the quantized pose feature matching the sample pose feature. For example, the electronic device may determine a distance between the sample encoded feature and the quantized pose feature, and uses the distance as the second loss value. In the disclosure, a manner of determining the second loss value according to the sample encoded feature and the quantized pose feature matching the sample pose feature is not limited.

[0090] In the embodiments of the disclosure, by determining the second loss value through the sample encoded feature and the quantized pose feature matching the sample pose feature, the model parameter can be rapidly and accurately updated.

[0091] In some embodiments, determining the second loss value according to the sample encoded feature and the quantized pose feature matching the sample pose feature includes the following operations.

[0092] A first loss sub-value is determined according to a difference between the sample encoded feature and the sample quantized pose feature.

[0093] A second loss sub-value is determined according to an opposite value of the difference value between the sample encoded feature and the sample quantized pose feature.

[0094] The second loss value is determined according to the first loss sub-value, the second loss sub-value, and their respective weight coefficients.

[0095] In the embodiments of the disclosure, the electronic device may determine the first loss sub-value according to the difference between the sample encoded feature and the sample quantized pose feature, and determine the second loss sub-value according to the opposite value of the difference between the sample encoded feature and the sample quantized pose feature. For example, the difference between the sample encoded feature and the sample quantized pose feature is 2, and the opposite value of the difference between the sample encoded feature and the sample quantized pose feature is -2. The electronic device determines the second loss value according to the first loss sub-value,

the second loss sub-value, and their respective preset weight coefficient. For example, the preset weight coefficient corresponding to the first loss sub-value is 0.3. The preset weight coefficient corresponding to the second loss sub-value is 0.7. According to the preset calculation strategy, the weighted first loss sub-value and second loss sub-value are processed, so as to obtain the second loss value.

[0096] In a possible embodiment, a through which the electronic device determines the quantized pose feature matching each predicted pose feature in the predicted pose feature sequence from the preset pose feature library may be shown as follows.

$$[0097] \quad e_{q,i} = \arg \min_{z_j \in Z} \|e_j - z_j\| \quad (1).$$

[0098] In the formula (1), e_j may represent the predicted pose feature, z_j may represent the quantized pose feature in the preset pose feature library, and $e_{q,j}$ may represent the quantized pose feature matching the predicted pose feature. Z may represent the preset pose feature library, j may represent the number of the current predicted pose feature, and $\arg \min \| \|$ may represent a minimum distance between the predicted pose feature and the quantized pose feature.

[0099] In a possible embodiment, a through which the electronic device determines the first loss value according to the predicted spatial association relationship and the preset spatial association relationship may be shown as follows.

$$[00100] \quad L_{rec} = \|\hat{P} - P\|_1 + \alpha_1 \|\hat{P}' - P'\|_1 + \alpha_2 \|\hat{P}'' - P''\|_1 \quad (2).$$

[00101] In the formula (2), L_{rec} may represent the first loss value, $\hat{P}$ may represent the predicted association relationship, P may represent the preset association relationship, $\hat{P}'$ and P' may represent first-order derivatives, $\hat{P}''$ and P'' may represent second-order derivatives, $\| \|_1$ may represent a distance between the features, and α_1 and α_2 may represent custom parameters.

[00102] In a possible embodiment, a through which the electronic device determines the first target loss value may be shown as follows.

[00103]
$$L_{VQ} = L_{rec}(\hat{P}, P) + \|sg[e] - e_q\| + \beta \|e - sg[e_q]\| \quad (3).$$

[00104] In the formula (3), L_{VQ} may represent the first target loss value, L_{rec} may represent the first loss value, e may represent the sample encoded feature, e_q may represent the sample quantized pose feature. $Sg[]$ may represent a stop gradient, and may be understood that the gradient needs to be calculated during the forward propagation during model training, and the gradient does not need to be calculated during the backward propagation. $\| \|$ may represent the distance between the features. By setting the distance between the sample encoded feature and the sample quantized pose feature, the sample encoded feature approaches close to the sample quantized pose feature, while the sample quantized pose feature approaches close to the sample encoded feature. Therefore, training efficiency can be enhanced. β may represent the custom parameter. In a possible embodiment, a first item, which is the first loss value (also known as reconstruction loss), in L_{VQ} may be used to train the encoder and the decoder in the neural network. A second item, which may also be known as codebook loss, in L_{VQ} may further be used to train the starting pose feature library in the neural network. A third item, which may also be known as commitment loss, in L_{VQ} may further be used to train the encoder in the neural network, to prevent the output of the encoder from jumping repeatedly between the sample quantized pose features. The target feature processing model may also be known as choreographic memory model. The target feature processing model may be understood as a neural network trained based on Vector Quantized Variational AutoEncoder (VQ-VAE) framework.

[00105] In the embodiments of the disclosure, the second loss sub-value is determined through information of the difference between the sample encoded feature and the sample quantized pose feature, so that the similarity of the sample encoded feature and the quantized pose feature can be enhanced. Therefore, standard poses in the preset pose feature library can be more in line with actual specifications.

[00106] In some embodiments, acquiring the temporally associated predicted pose feature sequence includes the following operations.

[00107] The predicted pose feature temporally aligned with each audio sub-feature in

the audio feature is determined according to the each audio sub-feature and a starting pose feature of a target role. Respective audio sub-features in the audio feature are temporally associated.

[00108] The predicted pose feature sequence is obtained according to each predicted pose feature.

[00109] In the embodiments of the disclosure, the electronic device may divide the audio feature to obtain at least two audio sub-features. For example, the electronic device may divide the audio feature according to manners such as a preset pitch level, a tone range or a time interval, to obtain the temporally associated audio sub-features. For example, the first audio sub-feature is a first frame audio, and the second audio sub-feature is a second frame audio. The electronic device may produce the first predicted pose feature according to the first audio sub-feature and the starting pose feature, and then produce the second predicted pose feature according to the second audio sub-feature and the first predicted pose feature. In this way, there is a temporal association relationship between various generated predicted pose features. In a possible embodiment, the electronic device may further produce the first predicted pose feature according to the first audio sub-feature and the starting pose feature, and then produce the second predicted pose feature according to the first audio sub-feature, the second audio sub-feature, the starting pose feature, and the first predicted pose feature.

[00110] After the electronic device determines the predicted pose feature temporally aligned with the audio sub-feature, all of the predicted pose features corresponding to all of the audio sub-features are obtained. At least two predicted pose features may be combined according to the temporal association relationship of respective audio sub-features and an alignment relationship between the predicted pose feature and the audio sub-feature, so as to obtain the predicted pose feature sequence. For example, the predicted pose features at the first row in the predicted pose feature sequence are predicted pose features corresponding to the first audio sub-features. The predicted pose features at the second row in the predicted pose feature sequence are predicted pose features corresponding to the second audio sub-features.

[00111] In the embodiments of the disclosure, the predicted pose feature temporally

aligned with each audio sub-feature in the audio feature is determined according to each audio sub-feature and the starting pose feature. Therefore, the predicted pose feature sequence may be obtained according to each predicted pose feature. In this way, the predicted pose feature sequence can be rapidly and accurately obtained.

[00112] In some embodiments, determining the predicted pose feature temporally aligned with each audio sub-feature in the audio feature according to the each audio sub-feature and the starting pose feature of the target role includes the following operations.

[00113] The starting pose feature is split according to part information of the target role to obtain at least two starting pose sub-features.

[00114] An association degree between each audio sub-feature and each starting pose sub-feature is determined. The association degree at least characterizes an association degree between each audio sub-feature and each starting pose sub-feature in terms of content and time.

[00115] The predicted pose feature is determined according to the association degree between each audio sub-feature and each starting pose sub-feature.

[00116] In the embodiments of the disclosure, the electronic device may split the starting pose feature according to part information of the target role to obtain at least two starting pose sub-features. The part information may be understood as information of a part of the target role that performs a pose, for example, the target role is a virtual character. The part information may be the whole body, that is, a whole body pose represented by the starting pose feature. The whole body pose can be split into an upper half-body pose and a lower half-body pose, or a facial pose and a limb pose, and the like. The electronic device may split the starting pose feature to obtain at least two starting pose sub-features. For example, the first starting pose sub-feature represents the upper half-body pose, and the second starting pose sub-feature represents the lower half-body pose. In this way, more detailed whole body pose is accurately generated later, so that more diverse poses can be combined.

[00117] After obtaining at least two starting pose sub-features, the electronic device may determine the association degree between each audio sub-feature and each starting

pose sub-feature. The association degree includes at least an association degree between each audio sub-feature and each starting pose sub-feature in terms of content and time. During the generation of the predicted pose feature, according to the association relationship between each audio sub-feature and each audio sub-feature itself, the association relationship between each audio sub-feature and each starting pose sub-feature, and the association relationship between each starting pose sub-feature and itself, the generation of each predicted pose feature can be smoother, and the consistency with the audio rhythm can be improved. For example, there are 10 audio sub-features. The starting pose sub-feature includes 3 upper half-body starting pose features and 3 lower half-body starting pose features. One upper half-body starting pose feature and one lower half-body starting pose feature are combined to form a complete set of starting pose features.

[00118] In a possible embodiment, the electronic device may determine the association degree of the first audio sub-feature according to the association degree between the first audio sub-feature and the first audio sub-feature itself, an association relationship between the first audio sub-feature and the first upper half-body starting pose feature, and an association relationship between the first audio sub-feature and the first lower half-body starting pose feature. The association degree may include information such as feature information of the audio sub-feature and the starting pose sub-feature itself, and position information related to the audio sub-feature and the starting pose sub-feature. The electronic device may determine an association degree of the second audio sub-feature according to the association degree between the second audio sub-feature and the second audio sub-feature itself, an association relationship between the second audio sub-feature and the second upper half-body starting pose feature, an association relationship between the second audio sub-feature and the first upper half-body starting pose feature, an association relationship between the second audio sub-feature and the second lower half-body starting pose feature, and an association relationship between the second audio sub-feature and the first lower half-body starting pose feature. A manner for determining the association degrees of the second upper half-body starting pose feature and the second lower half-body starting pose feature is the same as a manner for determining the association degree of the second audio sub-feature.

[00119] After determining the association degree between each audio sub-feature and each starting pose sub-feature, the electronic device may determine the predicted pose feature according to the association degree between each audio sub-feature and each starting pose sub-feature. For example, the electronic device may determine the association degree according to a preset natural language processing (NLP) layer (Transformer layer), so as to determine the predicted pose feature.

[00120] In a possible embodiment, the pre-trained neural network (also known as the target pose prediction model) may be used to determine the predicted pose feature sequence according to the audio feature and the starting pose feature. FIG. 4 is a schematic diagram of a pose prediction method according to an exemplary embodiment. As shown in FIG. 4, a splicing feature sequence may be obtained by splicing the audio feature and the starting pose feature through feature embedding. The starting pose feature may include an upper/lower half-body sequence. For example, a first upper half-body starting pose is p_0^u (14), a second upper half-body starting pose is p_1^u (27), a first lower half-body starting pose is p_0^l (13), and a second lower half-body starting pose is p_1^l (68). A portion in the splicing feature sequence represent the audio feature may be expressed as m . A portion of a feature represented by an upper half-body pose may be expressed as u . A portion of a feature represented by a lower half-body pose may be expressed as l . A dimension of a splicing feature may be $3 \times T'$. Then, a position feature (positional embedding) is combined with the splicing feature. The position feature may represent position information of each item in the splicing feature. Then a combined result is inputted into a state network f_s 401 to obtain a state feature sequence (S). A second portion in the state feature sequence may represent a state feature corresponding to the upper half-body pose, and a third portion may represent a state feature corresponding to the lower half-body pose.

[00121] In a possible embodiment, a policy making network f_a 402 may be used to obtain a candidate pose feature sequence and a probability value (action probability) corresponding to each candidate pose feature. A dimension of a feature sequence (a) of the probability value may be $(3 \times T') \times N$. The candidate pose feature corresponding to the upper half-body pose may be expressed as a^u . The candidate pose feature

corresponding to the lower half-body pose may be expressed as a^l . Then, a selector (top-1 selection) may be used to determine the predicted sample pose feature from a plurality of candidate pose feature sequences. For example, the first upper half-body pose is $\hat{p}_1^u$ (27), the second upper half-body pose is $\hat{p}_2^u$ (33), the first lower half-body pose is $\hat{p}_1^l$ (68), and the second lower half-body pose is $\hat{p}_2^l$ (68). After the predicted sample pose feature sequence is determined, the quantized pose feature sequence matching each predicted pose feature in the predicted pose feature sequence may be determined from the preset pose feature library 404 (choreographic memory codebook). By using the CNN decoders 405, the target pose feature sequence (generated dance) is generated according to the quantized pose feature sequence.

[00122] In a possible embodiment, after the state feature sequence is obtained, a critic value network f_v 403 may be used to determine a critic value corresponding to each state feature. A dimension of a feature sequence represented by the critic value may be $(3 \times T') \times 1$. V^u represents the critic value corresponding to the feature represented by the upper half-body pose. V^l represents the critic value corresponding to the feature represented by the lower half-body pose. By combining V^u and V^l , the overall critic value (V) of the pose is obtained. Beat align/half-body consistency reward 406 may be performed according to the target pose feature sequence and music beats, so as to determine a reward R . According to the reward and the critic value, Time Difference error (TD-error) ε is determined. A cross-entropy loss L_{CE} may be determined according to the probability value. An actor-critic loss L_{AC} is determined according to the probability value and the TD-error. A value loss L_V is determined according to the TD-error. Finally, an initialized neural network may be trained according to the cross-entropy loss, the actor-critic loss, and the value loss, so as to obtain a trained neural network (which is the target pose prediction model). The target pose prediction model may also be known as a dance pose generation model (or actor-critic Generative Pre-trained Transformer, GPT), and may be understood as a neural network trained based on a reinforcement learning-based GPT framework.

[00123] In a possible embodiment, the state network includes a plurality of NLP layers (transformer layers). The NLP layers in the state network may include a horizontal

normalization layer, a cross conditional causal attention layer, a full connection layer (linear layer), an activation layer (GELU), a filter layer (dropout layer), and the like. The policy making network includes a plurality of NLP layers and linear layers or regression layers (Softmax). The critic value network includes a plurality of NLP layers and linear layers.

[00124] FIG. 5a, FIG. 5b, and FIG. 5c are schematic diagrams of different methods for determining an association degree according to an exemplary embodiment. FIG. 5a may show a full attention manner. An association degree between each feature and all of the features in the sequence is determined. FIG. 5b may show a causal attention manner. An association degree between a first feature in a feature sequence and the first feature is determined. An association degree between a second feature and the first feature and the second feature is determined. An association degree between a third feature and the first feature, the second feature and the third feature is determined. An area of FIG. 5c may show a cross-conditional causal attention manner. For a music feature in the feature sequence, an association degree between a first music feature and the first music feature, a first upper half-body pose feature and a first lower half-body pose feature is determined. An association degree between a second music feature and the first music feature, the second music feature, the first upper half-body pose feature, a second upper half-body pose feature, the first lower half-body pose feature and a second lower half-body pose feature is determined. For different types of feature such as the upper half-body pose features and the lower half-body pose features in the feature sequence, the manner for determining an association relationship is the same. In this embodiment, the cross-conditional causal attention manner may be used to determine the association degree between the features. Therefore, an upper half-body pose and a lower half-body pose can both be taken into consideration during the determination of the association relationship. In this way, the inconsistency between the upper/lower half-body poses cannot occur.

[00125] In the embodiments of the disclosure, the starting pose feature is split according to the part information of the target role, so as to obtain at least two starting pose sub-features. Then, the association degree between each audio sub-feature and each starting pose sub-feature may be determined. In this way, the predicted pose feature can

be determined according to the association degree between each audio sub-feature and each starting pose sub-feature. Therefore, the accuracy of generating the predicted pose feature can be enhanced.

[00126] In some embodiments, the method further includes the following operations.

[00127] A sample audio feature and a sample starting pose feature of a sample role are acquired.

[00128] The sample starting pose feature is split according to part information of the sample role to obtain at least two sample starting pose sub-features.

[00129] A predicted sample pose feature sequence is determined according to each sample audio sub-feature in the sample audio feature and each sample starting pose sub-feature.

[00130] A starting pose prediction model is trained based on each predicted sample pose feature in the predicted sample pose feature sequence and each preset sample pose feature, to obtain a target pose prediction model.

[00131] Acquiring the temporally associated predicted pose feature sequence includes the following operation.

[00132] The temporally associated predicted pose feature sequence is acquired using the target pose prediction model.

[00133] In the embodiments of the disclosure, the sample audio feature may be understood as manually labeled audio information. A pose represented by the sample starting pose feature is generated by the sample role. The sample starting pose feature may be understood as the manually labeled pose feature sequence, and may be used for training the starting pose prediction model. The model parameter in the starting pose prediction model may be set randomly. During training, the model parameter may be updated according to the loss value determined by the predefined loss function. The electronic device may extract the sample audio feature from Internet, and the sample starting pose feature corresponding to the sample audio feature. The electronic device

may split the sample starting pose feature according to the part information of the sample role to obtain at least two sample starting pose sub-features. For example, the sample starting pose feature is divided into the sample starting pose sub-feature representing the upper half-body pose and the sample starting pose sub-feature representing the lower half-body pose.

[00134] In a possible embodiment, the electronic device may determine the predicted sample pose feature sequence using the starting pose prediction model according to each sample audio sub-feature and each sample starting pose sub-feature. The electronic device determines a loss value corresponding to a preset loss function according to each predicted sample pose feature and each preset sample pose feature, and updates the model parameter of the starting pose prediction model according to the loss value to obtain a trained target pose prediction model. After obtaining the trained target pose prediction model, the electronic device may acquire the temporally associated predicted pose feature sequence using the target pose prediction model according to the audio feature and the starting pose feature.

[00135] In the embodiments of the disclosure, the sample audio feature and the sample starting pose feature of the sample role are acquired. Then, the sample starting pose feature may be split to obtain at least two sample starting pose sub-features. Then the predicted sample pose feature sequence is determined. In this way, the starting pose prediction model can be trained based on each predicted sample pose feature and each preset sample pose feature, so as to obtain the target pose prediction model. Therefore, the accuracy of model training can be enhanced.

[00136] In some embodiments, training the starting pose prediction model based on each predicted sample pose feature in the predicted sample pose feature sequence and each preset sample pose feature, to obtain the target pose prediction model includes the following operations.

[00137] Candidate pose features and a probability value corresponding to each candidate pose feature are obtained according to each sample audio sub-feature and each sample starting pose sub-feature. The probability value is used for determining the predicted sample pose feature from the candidate pose features.

- [00138] A third loss value is determined according to a number of the sample audio sub-features, the preset sample pose feature and the probability value.
- [00139] A fourth loss value is determined according to the number of the sample audio sub-features, the predicted sample pose feature and the probability value.
- [00140] A second target loss value is obtained according to the third loss value and the fourth loss value.
- [00141] A model parameter of the starting pose prediction model is updated using the second target loss value, so as to obtain the target pose prediction model.
- [00142] In the embodiments of the disclosure, during the training of the target pose prediction model, the electronic device may obtain the candidate pose features and the probability value corresponding to each candidate pose feature according to each sample audio sub-feature and each sample starting pose sub-feature. The probability value is used for determining the predicted sample pose feature from the candidate pose features. The candidate pose features may be understood as a plurality of pose features that may be selected later according to the current sample audio sub-feature and the sample starting pose sub-feature. Each candidate pose feature may be used as the predicted sample pose feature. In order to cause the audio rhythm to be consistent with the pose beat, the candidate pose feature with the highest probability that the audio rhythm is consistent with the pose beat is required to be determined from the plurality of candidate pose features. For example, under a case that the current predicted sample pose feature is generated, the probability value corresponding to the first candidate pose feature is 0.8, the probability value corresponding to the second candidate pose feature is 0.3, and the probability value corresponding to the third candidate pose feature is 0.7. Then the electronic device may use the first candidate pose feature as the current predicted sample pose feature.
- [00143] After determining the probability value corresponding to the candidate pose feature, the electronic device may determine the third loss value according to the number of the sample audio sub-features, the preset sample pose feature and the probability value. The third loss value may be understood as a part of the loss value of the starting

pose prediction model calculated according to the preset third loss function determination strategy, and may also be understood as a difference between the predicted sample pose feature of the starting pose prediction model and the preset sample pose feature. The electronic device may determine the fourth loss value according to the number of the sample audio sub-features, the predicted sample pose feature and the probability value. The fourth loss value may be understood as a part of the loss value of the starting pose prediction model calculated according to the preset fourth loss function determination strategy, and may also be understood as a reward under a case that the starting pose prediction model selects the predicted sample pose feature.

[00144] The electronic device may obtain the second target loss value according to the third loss value and the fourth loss value, and update the model parameter of the starting pose prediction model to obtain the target pose prediction model. For example, the electronic device may obtain the second target loss value by performing weighted summation on the third loss value and the fourth loss value.

[00145] In the embodiments of the disclosure, the candidate pose features and the probability value corresponding to each candidate pose feature are obtained according to each sample audio sub-feature and each sample starting pose sub-feature. Then the third loss value may be determined according to the number of the sample audio sub-features, the preset sample pose feature and the probability value. The fourth loss value is determined according to the number of the sample audio sub-features, the predicted sample pose feature and the probability value. In this way, the second target loss value is obtained. The model parameter of the starting pose prediction model is updated. Therefore, target loss values can be accurately determined, and the efficiency and accuracy of model training can be enhanced.

[00146] In some embodiments, determining the third loss value according to the number of the sample audio sub-features, the preset sample pose feature and the probability value includes the following operations.

[00147] A difference between a vector feature of the probability value and the preset sample pose feature is determined.

[00148] The third loss value is determined according to the difference value between the vector feature of the probability value and the preset sample pose feature, and the number of the sample audio sub-features.

[00149] In the embodiments of the disclosure, the electronic device may determine the difference between the vector feature of the probability value and the preset sample pose feature, and then determine the third loss value according to the difference value between the vector feature of the probability value and the preset sample pose feature, and the number of the sample audio sub-features. For example, in a process that the electronic device successively determines the predicted sample pose feature, the differences between the vector feature of the probability value and the preset sample pose feature are respectively be 2, 1, 3, 4, 2... When the predicted sample pose feature corresponding to the third sample audio sub-feature is determined, the first difference, the second difference and the third difference may be accumulated, so that the third loss value is determined when the predicted sample pose feature corresponding to the current sample audio sub-feature is determined.

[00150] In the embodiments of the disclosure, by determining the difference between the vector feature of the probability value and the preset sample pose feature, the third loss value can be rapidly and accurately determined according to the difference value between the vector feature of the probability value and the preset sample pose feature, and the number of the sample audio sub-features. Therefore, the efficiency and accuracy of model training can be enhanced.

[00151] In some embodiments, determining the fourth loss value according to the number of the sample audio sub-features, the predicted sample pose feature and the probability value includes the following operations.

[00152] A difference between the predicted sample pose feature and a vector feature of the probability value is determined.

[00153] A direction feature of the sample role is determined according to each sample starting pose sub-feature.

[00154] The fourth loss value is determined according to the number of the sample

audio sub-features, the difference value between the predicted sample pose feature and the vector feature of the probability value, and the direction feature.

[00155] In the embodiments of the disclosure, the electronic device may determine the difference between the predicted sample pose feature and the vector feature of the probability value, and determine the direction feature of the sample role according to each sample starting pose sub-feature. The difference between the predicted sample pose feature and the vector feature of the probability value may be understood as a distance difference between different features. The direction feature may be understood as direction information of a pose generated by the sample role when the sample role generates the pose. For example, the direction feature of the whole body pose can be characterized as facing the front right, the direction feature of the upper half-body pose can be characterized as facing the left, and the direction feature of the lower half-body pose can be characterized as facing the right.

[00156] After determining the difference between the predicted sample pose feature and the vector feature of the probability value and the direction feature of the sample role, the electronic device may determine the fourth loss value according to the number of the sample audio sub-features, the difference value between the predicted sample pose feature and the vector feature of the probability value, and the direction feature. For example, in a process that the electronic device successively determines the predicted sample pose feature, the differences between the vector feature of the probability value and the preset sample pose feature are respectively be 3, 5, 6, 2, 4.... The direction feature may represent that the upper half-body pose is consistent with the lower half-body pose, or the upper half-body pose is inconsistent with the lower half-body pose. When the predicted sample pose feature corresponding to the third sample audio sub-feature is determined, the difference and the direction feature may be fused (for example, by multiplying). Then a first fusion result, the second fusion result and the third fusion result are accumulated, so that the fourth loss value is determined when the predicted sample pose feature corresponding to the current sample audio sub-feature is determined.

[00157] In the embodiments of the disclosure, by determining the difference between the predicted sample pose feature and the vector feature of the probability value and the

direction feature of the sample role, the fourth loss value can be rapidly and accurately determined according to the number of the sample audio sub-features, the difference value between the predicted sample pose feature and the vector feature of the probability value, and the direction feature. Therefore, the efficiency and accuracy of model training can be enhanced.

[00158] In a possible embodiment, a through which the electronic device obtains the candidate pose features and the probability value corresponding to each candidate pose feature according to each sample audio sub-feature and each sample starting pose sub-feature may be shown as follows.

$$[00159] \quad \begin{cases} \hat{p}_t^u = \arg \max_k P(z_k^u | m_{1..t}, p_{0..t-1}^u, p_{0..t-1}^l) \\ \hat{p}_t^l = \arg \max_k P(z_k^l | m_{1..t}, p_{0..t-1}^u, p_{0..t-1}^l) \end{cases} \quad (4).$$

[00160] In the formula (4), $\hat{p}_t^u$ may represent the probability value corresponding to the candidate pose feature determined by the upper half-body pose, $\hat{p}_t^l$ may represent the probability value corresponding to the candidate pose feature determined by the lower half-body pose, $P()$ may represent probability distribution, z_k^u and z_k^l may respectively represent the quantized pose features corresponding to the upper half-body pose and the lower half-body pose, k may represent the plurality of determined candidate pose features, and $m_{1..t}$ may represent the audio feature. $p_{0..t-1}^u$ may represent the starting pose feature corresponding to the upper half-body pose, $p_{0..t-1}^l$ may represent the starting pose feature corresponding to the lower half-body pose, and $\arg \max_k || \cdot ||$ may represent the predicted pose feature corresponding to the maximum reward.

[00161] In a possible embodiment, a through which the electronic device determines the third loss value according to the number of the sample audio sub-features, the preset sample pose feature and the probability value may be shown as follows.

$$[00162] \quad L_{CE} = \frac{1}{T'} \sum_{t=0}^{T'-1} \sum_{h=u,l} CrossEntropy(a_t^h, p_{t+1}^h) \quad (5).$$

[00163] In the formula (5), L_{CE} may represent the third loss value, and may further

known as a cross-entropy loss value. T' may represent the number of the sample audio sub-features. a_t^h may represent the probability value. h may represent the upper half-body pose u or the lower half-body pose l . t may represent a current processing number. p_{t+1}^h may represent the preset sample pose feature. $CrossEntropy()$ may represent the acquisition of the distance between different features.

[00164] In a possible embodiment, a through which the electronic device determines the fourth loss value according to the number of the sample audio sub-features, the predicted sample pose feature and the probability value may be shown as follows.

$$[00165] \quad v = v^u + v^l = f_v(s)_{T':2T'-1} + f_v(s)_{2T':3T'-1} \quad (6).$$

$$[00166] \quad \varepsilon_{0:T'-2} = r_{0:T'-2} + sg[v_{1:T'-1}] - v_{0:T'-2} \quad (7).$$

$$[00167] \quad L_{AC} = \frac{1}{T'-1} \sum_{t=0}^{T'-2} \left(\sum_{h=u,l} CrossEntropy(a_t^h, \hat{p}_{t+1}^h) \right) \cdot sg[\varepsilon_t] \quad (8).$$

$$[00168] \quad R_b(t) = \begin{cases} -1, \exists music \quad beat \wedge \nexists dance & beats \in P'_{td:(t+1)d} \\ 1, otherwise \end{cases} \quad (9).$$

$$[00169] \quad R_c(t) = \inf \{ \hat{R}_c(t) \}, t \in [dt, d(t+1)] \quad (10).$$

$$[00170] \quad \hat{R}_c(t) = \begin{cases} \langle n_{xz}^u(t), n_{xz}^l(t) \rangle, \langle n_{xz}^u(t), n_{xz}^l(t) \rangle < 0 \\ 1, otherwise \end{cases} \quad (11).$$

[00171] In formulas (6), (7), (8), (9), (10), and (11), v may represent the critic value corresponding to the predicted sample pose feature. v^u And v^l may respectively represent the critic values corresponding to the predicted sample pose feature represented by the upper half-body pose and the lower half-body pose. $f_v(s)$ may represent the critic value network. ε_0 may represent the TD-error. $r = R$ may represent the reward. $Sg[]$ may represent the stop gradient, and may be understood that the gradient needs to be calculated during the forward propagation, and the gradient does not need to be calculated during the backward propagation. L_{AC} may represent the fourth loss value, and

may further be known as an actor-critic loss value. T' may represent the number of the sample audio sub-features. a_t^h may represent the probability value. h may represent the upper half-body pose u or the lower half-body pose l . t may represent a current processing number. $\hat{p}_{t+1}^h$ may represent the predicted sample pose feature. $CrossEntropy()$ may represent the acquisition of the distance between different features. $R_b()$ may represent the first reward, and may also be known as a beat-align reward. $R_c()$ may represent the second reward, and may also be known as a compositional consistency reward. $\inf\{\}$ may represent a greatest lower bound. $\hat{R}_c()$ may represent a reference reward corresponding to the second reward. $n_{xz}^u()$ and $n_{xz}^l()$ may respectively represent the direction features corresponding to the upper half-body pose and the lower half-body pose relative to a xz plane.

[00172] In a possible embodiment, a calculation formula of the value loss may be shown as follows.

$$[00173] \quad L_v = \frac{1}{T' - 1} \|\varepsilon\|_2^2 \quad (12).$$

[00174] In the formula (12), L_v may represent the value loss. T' may represent the number of the sample audio sub-features. ε may represent the TD-error. $\|\cdot\|_2$ may represent a norm formula.

[00175] In a possible embodiment, FIG. 6 is a schematic diagram of a reward determination method according to an exemplary embodiment. An (a) area may represent the first reward. d may represent a duration corresponding to the sample audio sub-feature. A vertical dotted line may represent a music beat. A curve may represent a change trend of a pose. A solid point may represent a dance beat. There is a music beat point in the duration corresponding to the sample audio sub-feature. If there is a dance beat point, the first reward is 1. If there is a music beat point in the duration corresponding to the sample audio sub-feature, but there is no dance beat point, or if there is no music beat point in the duration corresponding to the sample audio sub-feature, the first reward is -1. FIG. (b) may represent the second reward. The direction

feature $n^u()$ of the upper half-body pose and the direction feature $n^l()$ of the lower half-body pose may be respectively determined according to the predicted sample pose feature corresponding to the upper half-body pose and the predicted sample pose feature corresponding to the lower half-body pose. Then the direction features $n_{xz}^u()$ and $n_{xz}^l()$ corresponding to the xz plane may be obtained. If an included angle between the direction feature $n_{xz}^u()$ and the direction feature $n_{xz}^l()$ does not belong to a same side, the second reward is 1. If the included angle between the direction feature $n_{xz}^u()$ and the direction feature $n_{xz}^l()$ belongs to the same side, the second reward is the included angle between the different direction features.

[00176] In a possible embodiment, FIG. 7 is a schematic diagram of a target pose according to an exemplary embodiment. p_0 and p_1 may represent the predicted pose features. In the decoded target pose sequence, in the case that there are consecutive repeated codes (for example, a first row p_0 and p_0 , and a second row p_1 and p_1) in the predicted pose feature, the generated target poses remain static, and in the case that there are different consecutive codes in the predicted pose feature (for example, the first row p_0 and p_1), smooth transition is achieved between the generated target poses.

[00177] In some embodiments, the pose generation method in the embodiments of the disclosure may be used for generating dance poses. For example, a target dance pose may be generated according to a target music and a starting pose of the target role. The generation of the target dance pose may include the following steps.

[00178] An audio feature of a target music and a starting pose feature of a target role are acquired.

[00179] The temporally associated predicted pose feature sequence is generated according to the audio feature of the target music and the starting pose feature of the target role. The predicted pose feature sequence includes predicted dance pose features.

[00180] A quantized pose feature matching each predicted dance pose feature is determined from the preset pose feature library.

[00181] A spatial association relationship of respective predicted dance pose features in the predicted pose feature sequence is determined.

[00182] The quantized pose feature matching each predicted dance pose feature is combined according to the spatial association relationship of respective predicted dance pose features and a temporal association relationship of the respective predicted dance pose features, to obtain the target pose feature sequence corresponding to the target role.

[00183] The target role is controlled to perform a target dance pose corresponding to the target pose feature sequence.

[00184] In the embodiments of the disclosure, the target music may include different styles of music such as pop, classical, jazz, rock, or folk. The starting pose feature of the target role may be determined through a preset default pose feature, or may be determined according to a current instruction of a user. The electronic device may use the audio feature of the target music and the starting pose feature of the target role as inputs, and obtain the target pose feature sequence corresponding to the target role according to the audio feature of the target music and the starting pose feature of the target role. Therefore, the target role may be controlled to perform the target dance pose corresponding to the target pose feature sequence. The target dance pose may be understood as the dance pose matching the form of the target music. The target music may include dances such as breaking, ballet jazz, house dance, or middle hip-hop. Different forms of the target music may correspond to different types of target dance poses. For example, by inputting the target music in the form of folk songs and the starting pose feature, and the target role may be controlled to perform dance poses of the folk dance type.

[00185] On the basis of the pose generation method provided in the foregoing embodiments, an embodiment of the disclosure further provides a pose generation apparatus.

[00186] FIG. 8 is a schematic structural diagram of a pose generation apparatus according to an embodiment of the disclosure. As shown in FIG. 8, the apparatus may include an acquisition module, a first determination module, a second determination

module, and a combination module.

[00187] The acquisition module 801 is configured to acquire a temporally associated predicted pose feature sequence. The predicted pose feature sequence is generated at least according to an audio feature.

[00188] The first determination module 802 is configured to determine a quantized pose feature matching each predicted pose feature in the predicted pose feature sequence from a preset pose feature library. The second determination module 803 is configured to determine a spatial association relationship of respective pose features in the predicted pose feature sequence. The combination module 804 is configured to combine the quantized pose feature matching each predicted pose feature according to the spatial association relationship of the respective pose features and a temporal association relationship of the respective pose features, to obtain a target pose feature sequence.

[00189] In some embodiments, the first determination module 802 is configured to: respectively perform dimension reduction processing on each predicted pose feature to obtain an encoded feature, where a dimension of the encoded feature is the same as a dimension of each quantized pose feature in the preset pose feature library; and determine the quantized pose feature matching each predicted pose feature from the preset pose feature library according to a distance between the encoded feature and each quantized pose feature in the preset pose feature library.

[00190] In some embodiments, the first determination module 802 is configured to: determine a distance between the to-be-processed encoded feature and each quantized pose feature in the preset pose feature library; sort respective distances corresponding to the to-be-processed encoded feature to obtain a sorting result; determine the quantized pose feature corresponding to the to-be-processed encoded feature according to the sorting result; and determine the quantized pose feature matching each predicted pose feature according to the quantized pose feature corresponding to each to-be-processed encoded feature.

[00191] In some embodiments, the second determination module 803 is configured to: determine relative position information between each key point of a target role and a

preset standard point and displacement information of the target role according to each quantized pose feature, where the predicted pose feature sequence is generated according to the audio feature and a starting pose feature of the target role; and determine the spatial association relationship of the respective pose features according to the relative position information and the displacement information.

[00192] In some embodiments, the apparatus 800 further includes: a third determination module, configured to acquire a sample pose feature sequence; an updating module, configured to update a model parameter of a starting feature processing model and the quantized pose feature in a starting pose feature library of the starting feature processing model according to respective sample pose features in the sample pose feature sequence and a preset spatial association relationship of the respective sample pose features, to obtain an updating result; and a fourth determination module, configured to obtain a target feature processing model having the preset pose feature library based on the updating result. The first determination module 802 is configured to: determine the quantized pose feature matching each predicted pose feature in the predicted pose feature sequence from the preset pose feature library of the target feature processing model.

[00193] In some embodiments, the updating module is configured to: obtain a predicated spatial association relationship of the respective sample pose features using the starting feature processing model; determine a first loss value according to the predicted spatial association relationship and the preset spatial association relationship; perform dimension reduction processing on each sample pose feature to obtain a sample encoded feature; determine a second loss value according to the sample encoded feature; obtain a first target loss value according to the first loss value and the second loss value; and update the model parameter of the starting feature processing model and the quantized pose feature in the starting pose feature library of the starting feature processing model using the first target loss value, to obtain the updating result.

[00194] In some embodiments, the updating module is configured to: determine the quantized pose feature matching the sample pose feature from the starting pose feature library according to the sample encoded feature; and determine the second loss value according to the sample encoded feature and the quantized pose feature matching the

sample pose feature.

[00195] In some embodiments, the updating module is configured to: determine a first loss sub-value according to a difference between the sample encoded feature and the sample quantized pose feature; determine a second loss sub-value according to an opposite value of the difference value between the sample encoded feature and the sample quantized pose feature; and determine the second loss value according to the first loss sub-value, the second loss sub-value, and their respective weight coefficients.

[00196] In some embodiments, the acquisition module 801 is configured to: determine the predicted pose feature temporally aligned with each audio sub-feature in the audio feature according to the each audio sub-feature and a starting pose feature of a target role, where respective audio sub-features in the audio feature are temporally associated; and obtain the predicted pose feature sequence according to each predicted pose feature.

[00197] In some embodiments, the acquisition module 801 is configured to: split the starting pose feature according to part information of the target role to obtain at least two starting pose sub-features; determine an association degree between each audio sub-feature and each starting pose sub-feature, where the association degree at least characterizes an association degree between each audio sub-feature and each starting pose sub-feature in terms of content and time; and determine the predicted pose feature according to the association degree between each audio sub-feature and each starting pose sub-feature.

[00198] In some embodiments, the apparatus 800 further includes: a fifth determination module, configured to acquire a sample audio feature and a sample starting pose feature of a sample role; a splitting module, configured to split the sample starting pose feature according to part information of the sample role to obtain at least two sample starting pose sub-features; a sixth determination module, configured to determine a predicted sample pose feature sequence according to each sample audio sub-feature in the sample audio feature and each sample starting pose sub-feature; and a seventh determination module, configured to train a starting pose prediction model based on each predicted sample pose feature in the predicted sample pose feature sequence and each preset sample pose feature, to obtain a target pose prediction model. The acquisition module

801 is configured to: utilize the target pose prediction model to acquire the temporally associated predicted pose feature sequence.

[00199] In some embodiments, the seventh determination module is configured to: obtain candidate pose features and a probability value corresponding to each candidate pose feature according to each sample audio sub-feature and each sample starting pose sub-feature, where the probability value is used for determining the predicted sample pose feature from the candidate pose features; determine a third loss value according to the number of the sample audio sub-features, the preset sample pose feature and the probability value; determine a fourth loss value according to the number of the sample audio sub-features, the predicted sample pose feature and the probability value; obtaining a second target loss value according to the third loss value and the fourth loss value; and update a model parameter of the starting pose prediction model using the second target loss value, so as to obtain the target pose prediction model.

[00200] In some embodiments, the seventh determination module is configured to: determine a difference between a vector feature of the probability value and the preset sample pose feature; and determine the third loss value according to the difference value between the vector feature of the probability value and the preset sample pose feature, and the number of the sample audio sub-features.

[00201] In some embodiments, the seventh determination module is configured to: determine a difference between the predicted sample pose feature and a vector feature of the probability value; determine a direction feature of the sample role according to each sample starting pose sub-feature; and determine the fourth loss value according to the number of the sample audio sub-features, the difference value between the predicted sample pose feature and the vector feature of the probability value, and the direction feature.

[00202] In some embodiments, the apparatus 800 further includes: an eighth determination module, configured to acquire an audio feature of a target music and a starting pose feature of a target role; a generation module, configured to generate the temporally associated predicted pose feature sequence according to the audio feature of the target music and the starting pose feature of the target role, the predicted pose feature

sequence including predicted dance pose features; a ninth determination module, configured to determine a quantized pose feature matching each predicted dance pose feature from the preset pose feature library; a tenth determination module, configured to determine a spatial association relationship of respective predicted dance pose features in the predicted pose feature sequence; an eleventh determination module, configured to combine the quantized pose feature matching each predicted dance pose feature according to the spatial association relationship of respective predicted dance pose features and a temporal association relationship of the respective predicted dance pose features, to obtain the target pose feature sequence corresponding to the target role; and a control module, configured to control the target role to perform a target dance pose corresponding to the target pose feature sequence.

[00203] The acquisition module 801, the first determination module 802, the second determination module 803, and the combination module 804 may be implemented based on a processor of an electronic device.

[00204] In addition, the functional modules in this embodiment may be integrated into one processing unit, or each unit may exist alone physically, or two or more than two units may be integrated into one unit. The above integrated unit can be implemented in the form of hardware, or can be implemented in the form of a software functional module.

[00205] When implemented in form of software function module and sold or used not as an independent product, the integrated unit may be stored in a computer-readable storage medium. Based on such an understanding, the technical solution of the embodiment substantially or parts making contributions to the conventional art or all or part of the technical solution may be embodied in form of software product, and the computer software product is stored in a storage medium, including a plurality of instructions configured to enable a computer device (which may be a personal computer, a server, a network device or the like) or a processor to execute all or part of the steps of the method in the embodiment. The storage medium includes: various media capable of storing program codes such as a U disk, a mobile Hard Disk Drive (HDD), a Read-Only Memory (ROM), a Random Access Memory (RAM), a magnetic disk or an optical disk.

[00206] Specifically, a program computer instruction corresponding to the pose generation method in this embodiment may be stored in a storage medium such as an optical disk, an HDD and a U disk, and the computer program instruction corresponding to the pose generation method in the storage medium is read or executed by an electronic device to implement the pose generation method of the foregoing embodiments.

[00207] Based on the same technical concept of the foregoing embodiments, referring to FIG. 9, an electronic device 90 provided in the embodiments of the disclosure is shown, which may include a memory 91, a processor 92, and a computer program stored on the memory 91 and executable on the processor 92.

[00208] The memory 91 is configured to store the computer program and data. The processor 92 is configured to execute the computer program stored in the memory to implement the pose generation method of the foregoing embodiments.

[00209] During a practical application, the memory 91 may be a volatile memory such as a RAM, or a non-volatile memory such as a ROM, a flash memory, an HDD or a Solid-State Drive (SSD), or a combination of the memories, and provides instructions and data for the processor 92.

[00210] The processor 92 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.

[00211] In some embodiments, functions or modules of the device provided in the embodiment of the disclosure may be configured to execute the method described in the above method embodiment and specific implementation thereof may refer to the descriptions about the method embodiment and, for simplicity, will not be elaborated herein.

[00212] The above description of the various embodiments tends to emphasize the differences between the various embodiments, and the identicalness or similarities can be referred to each other. For ease of brevity, details are not described here again.

[00213] The methods disclosed in each method embodiment provided by the disclosure can be combined arbitrarily without conflict to obtain a new method embodiment.

[00214] The characteristics disclosed in each product embodiment provided in the disclosure can be combined arbitrarily without conflict to obtain a new product embodiment.

[00215] The characteristics disclosed in each method or device embodiment provided in the disclosure can be combined arbitrarily without conflict to obtain a new method embodiment or device embodiment.

[00216] From the above descriptions about the implementation modes, those skilled in the art may clearly know that the method of the foregoing embodiments may be implemented in a manner of combining software and a necessary universal hardware platform, and of course, may also be implemented through hardware, but the former is a preferred implementation mode under many circumstances. Based on such an understanding, the technical solutions of the disclosure substantially or parts making contributions to the conventional art may be embodied in form of software product, and the computer software product is stored in a storage medium (for example, a ROM/RAM), a magnetic disk and an optical disk), including a plurality of instructions configured to enable a terminal (which may be a mobile phone, a server, an air conditioner, a network device or the like) to execute the method in each embodiment of the disclosure.

[00217] The embodiments of the disclosure are described above with reference to the accompanying drawings, but the disclosure is not limited to the above specific embodiments, which are merely illustrative rather than restrictive. Under the inspiration of the disclosure, those of ordinary skill in the art can also make many forms without departing from the scope of the disclosure and the protection scope of the claims, which all fall within the protection of the disclosure.

CLAIMS

1. A pose generation method, comprising:

acquiring a temporally associated predicted pose feature sequence, wherein the predicted pose feature sequence is generated at least according to an audio feature;

determining a quantized pose feature matching each predicted pose feature in the predicted pose feature sequence from a preset pose feature library;

determining a spatial association relationship of respective pose features in the predicted pose feature sequence; and

combining the quantized pose feature matching each predicted pose feature according to the spatial association relationship of the respective pose features and a temporal association relationship of the respective pose features, to obtain a target pose feature sequence.

2. The method of claim 1, wherein determining the quantized pose feature matching each predicted pose feature in the predicted pose feature sequence from the preset pose feature library comprises:

respectively performing dimension reduction processing on each predicted pose feature to obtain an encoded feature, wherein a dimension of the encoded feature is the same as a dimension of each quantized pose feature in the preset pose feature library; and

determining the quantized pose feature matching each predicted pose feature from the preset pose feature library according to a distance between the encoded feature and each quantized pose feature in the preset pose feature library.

3. The method of claim 2, wherein determining the quantized pose feature matching each

predicted pose feature from the preset pose feature library according to the distance between the encoded feature and each quantized pose feature in the preset pose feature library comprises:

determining a distance between the to-be-processed encoded feature and each quantized pose feature in the preset pose feature library;

sorting respective distances corresponding to the to-be-processed encoded feature to obtain a sorting result;

determining the quantized pose feature corresponding to the to-be-processed encoded feature according to the sorting result; and

determining the quantized pose feature matching each predicted pose feature according to the quantized pose feature corresponding to each to-be-processed encoded feature.

4. The method of any of claims 1 to 3, wherein determining the spatial association relationship of respective pose features in the predicted pose feature sequence comprises:

determining relative position information between each key point of a target role and a preset standard point and displacement information of the target role according to each quantized pose feature, wherein the predicted pose feature sequence is generated according to the audio feature and a starting pose feature of the target role; and

determining the spatial association relationship of the respective pose features according to the relative position information and the displacement information.

5. The method of any of claims 1 to 4, further comprising:

acquiring a sample pose feature sequence;

updating a model parameter of a starting feature processing model and a quantized

pose feature in a starting pose feature library of the starting feature processing model according to respective sample pose features in the sample pose feature sequence and a preset spatial association relationship of the respective sample pose features, to obtain an updating result; and

obtaining a target feature processing model having the preset pose feature library based on the updating result,

wherein determining the quantized pose feature matching each predicted pose feature in the predicted pose feature sequence from the preset pose feature library comprises:

determining the quantized pose feature matching each predicted pose feature in the predicted pose feature sequence from the preset pose feature library of the target feature processing model.

6. The method of claim 5, wherein updating the model parameter of the starting feature processing model and the quantized pose feature in the starting pose feature library of the starting feature processing model according to respective sample pose features in the sample pose feature sequence and the preset spatial association relationship of the respective sample pose features, to obtain the updating result comprises:

obtaining a predicated spatial association relationship of the respective sample pose features using the starting feature processing model;

determining a first loss value according to the predicted spatial association relationship and the preset spatial association relationship;

performing dimension reduction processing on each sample pose feature to obtain a sample encoded feature;

determining a second loss value according to the sample encoded feature;

obtaining a first target loss value according to the first loss value and the second loss value; and

updating the model parameter of the starting feature processing model and the quantized pose feature in the starting pose feature library of the starting feature processing model using the first target loss value, to obtain the updating result.

7. The method of claim 6, wherein determining the second loss value according to the sample encoded feature comprises:

determining a quantized pose feature matching the sample pose feature from the starting pose feature library according to the sample encoded feature; and

determining the second loss value according to the sample encoded feature and the quantized pose feature matching the sample pose feature.

8. The method of claim 7, wherein determining the second loss value according to the sample encoded feature and the quantized pose feature matching the sample pose feature comprises:

determining a first loss sub-value according to a difference value between the sample encoded feature and a sample quantized pose feature;

determining a second loss sub-value according to an opposite value of the difference value between the sample encoded feature and the sample quantized pose feature; and

determining the second loss value according to the first loss sub-value, the second loss sub-value, and their respective weight coefficients.

9. The method of any of claims 1 to 8, wherein acquiring the temporally associated predicted pose feature sequence comprises:

determining the predicted pose feature temporally aligned with each audio sub-feature in the audio feature according to the each audio sub-feature and a starting pose feature of a target role, wherein respective audio sub-features in the audio feature are temporally

associated; and

obtaining the predicted pose feature sequence according to each predicted pose feature.

10. The method of claim 9, wherein determining the predicted pose feature temporally aligned with each audio sub-feature in the audio feature according to the each audio sub-feature and the starting pose feature of the target role comprises:

splitting the starting pose feature according to part information of the target role to obtain at least two starting pose sub-features;

determining an association degree between each audio sub-feature and each starting pose sub-feature, wherein the association degree at least characterizes an association degree between each audio sub-feature and each starting pose sub-feature in terms of content and time; and

determining the predicted pose feature according to the association degree between each audio sub-feature and each starting pose sub-feature.

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

acquiring a sample audio feature and a sample starting pose feature of a sample role;

splitting the sample starting pose feature according to part information of the sample role to obtain at least two sample starting pose sub-features;

determining a predicted sample pose feature sequence according to each sample audio sub-feature in the sample audio feature and each sample starting pose sub-feature; and

training a starting pose prediction model based on each predicted sample pose feature in the predicted sample pose feature sequence and each preset sample pose feature, to obtain a target pose prediction model,

wherein acquiring the temporally associated predicted pose feature sequence comprises:

acquiring the temporally associated predicted pose feature sequence using the target pose prediction model.

12. The method of claim 11, wherein training the starting pose prediction model based on each predicted sample pose feature in the predicted sample pose feature sequence and each preset sample pose feature, to obtain the target pose prediction model comprises:

obtaining candidate pose features and a probability value corresponding to each candidate pose feature according to each sample audio sub-feature and each sample starting pose sub-feature, wherein the probability value is used for determining the predicted sample pose feature from the candidate pose features;

determining a third loss value according to a number of the sample audio sub-features, the preset sample pose feature and the probability value;

determining a fourth loss value according to the number of the sample audio sub-features, the predicted sample pose feature and the probability value;

obtaining a second target loss value according to the third loss value and the fourth loss value; and

updating a model parameter of the starting pose prediction model using the second target loss value, so as to obtain the target pose prediction model.

13. The method of claim 12, wherein determining the third loss value according to the number of the sample audio sub-features, the preset sample pose feature and the probability value comprises:

determining a difference value between a vector feature of the probability value and the preset sample pose feature; and

determining the third loss value according to the difference value between the vector feature of the probability value and the preset sample pose feature, and the number of the sample audio sub-features.

14. The method of claim 12, wherein determining the fourth loss value according to the number of the sample audio sub-features, the predicted sample pose feature and the probability value comprises:

determining a difference value between the predicted sample pose feature and a vector feature of the probability value;

determining a direction feature of the sample role according to each sample starting pose sub-feature; and

determining the fourth loss value according to the number of the sample audio sub-features, the difference value between the predicted sample pose feature and the vector feature of the probability value, and the direction feature.

15. The method of any of claims 1 to 14, further comprising:

acquiring an audio feature of a target music and a starting pose feature of a target role;

generating the temporally associated predicted pose feature sequence according to the audio feature of the target music and the starting pose feature of the target role, wherein the predicted pose feature sequence comprises predicted dance pose features;

determining a quantized pose feature matching each predicted dance pose feature from the preset pose feature library;

determining a spatial association relationship of respective predicted dance pose features in the predicted pose feature sequence;

combining the quantized pose feature matching each predicted dance pose feature

according to the spatial association relationship of respective predicted dance pose features and a temporal association relationship of the respective predicted dance pose features, to obtain the target pose feature sequence corresponding to the target role; and

controlling the target role to perform a target dance pose corresponding to the target pose feature sequence.

16. A pose generation apparatus, comprising:

an acquisition module, configured to acquire a temporally associated predicted pose feature sequence, wherein the predicted pose feature sequence is generated at least according to an audio feature;

a first determination module, configured to determine a quantized pose feature matching each predicted pose feature in the predicted pose feature sequence from a preset pose feature library;

a second determination module, configured to determine a spatial association relationship of respective pose features in the predicted pose feature sequence; and

a combination module, configured to combine the quantized pose feature matching each predicted pose feature according to the spatial association relationship of the respective pose features and a temporal association relationship of the respective pose features, to obtain a target pose feature sequence.

17. An electronic device, comprising a processor and a memory configured to store a computer program executable on the processor, wherein

the processor is configured to operate the computer program to perform the pose generation method of any of claims 1 to 15.

18. A computer storage medium, having a computer program stored thereon, wherein the pose generation method of any of claims 1 to 15 is implemented when the computer program is performed by a processor.