

METHOD AND APPARATUS FOR MOTION ORCHESTRATION OF DIGITAL HUMANS

This application claims priority to Chinese Patent Application No. 2023109570705, filed with China National Intellectual Property Administration on July 31, 2023 and entitled "METHOD AND APPARATUS FOR MOTION ORCHESTRATION OF DIGITAL HUMANS", which is incorporated herein by reference in its entirety.

TECHNICAL FIELD

One or more embodiments of this specification relate to the field of virtual digital humans, and in particular, to a method and an apparatus for motion orchestration of digital humans.

BACKGROUND

In the real world, people typically accompany speech with some body motions, and many subconscious motions are typically closely linked to semantics and the environment. As a form of nonverbal signals (nonverbal signals) in daily communication, body language constitutes an essential component of human interaction. When a speaker incorporates appropriate gestures, the accuracy of audience's comprehension can be enhanced.

In practical services, digital humans need to use body motions to assist in conveying information through language during speech delivery, explanation, live streaming, and interaction. Meanwhile, in current common application scenarios of digital humans, body motions typically occur simultaneously with speech/text. Therefore, how to present motions that comply with semantics during speech delivery of digital humans is crucial for natural digital human interaction and enhanced user experience.

SUMMARY

One or more embodiments of this specification describe a method and an apparatus for motion orchestration of digital humans, to match appropriate motions for a text delivered by a digital human.

According to a first aspect, a method for motion orchestration of digital humans is provided, including:

obtaining a first word sequence corresponding to a target text, and obtaining a first motion set, where the first motion set includes a plurality of motions, and any motion has a preset

trigger word and corresponding motion duration;

matching words in the first word sequence against the first motion set at least based on the trigger word to obtain a first matching result including alternative motion sets corresponding to the words; and

determining, from the first matching result based on positions of the words in the target text and the motion duration, a group of motion sequences that do not overlap in the temporal dimension to obtain a first motion sequence, where the first motion sequence is configured to indicate a digital human to make a corresponding motion when delivering the target text.

In a possible implementation, the matching words in the first word sequence against the first motion set to obtain a first matching result includes:

for any target word, matching the target word against the trigger word to obtain a first candidate motion set;

obtaining a second candidate motion set based on similarities between a semantic representation of the target word and motion representations of the motions in the first motion set; and

determining an alternative motion set corresponding to the target word based on the first candidate motion set and the second candidate motion set.

In a possible implementation, the obtaining a second candidate motion set based on similarities between a semantic representation of the target word and motion representations of the motions in the first motion set includes:

encoding the target word using a first text encoder to obtain a target semantic representation;

obtaining a motion representation obtained by encoding any motion in the first motion set using a motion encoder; and

categorizing motions whose motion representation has a similarity with the target semantic representation being greater than a first threshold into the second candidate motion set.

In a possible implementation, the first text encoder is configured to encode a text in a first language; the motion encoder is pre-trained jointly with a second text encoder suitable for a second language; and the first text encoder is trained in the following manner:

encoding a first training corpus in the first language using the first text encoder to obtain a first semantic representation;

encoding a second training corpus in the second language using the second text encoder to obtain a second semantic representation, where the second training corpus and the first training corpus have the same semantics; and

adjusting a value of a parameter in the first text encoder by maximizing a similarity between the first semantic representation and the second semantic representation.

In a possible implementation, the determining an alternative motion set corresponding to the target word based on the first candidate motion set and the second candidate motion set includes:

determining the alternative motion set based on a union of the first candidate motion set and the second candidate motion set; or

determining the alternative motion set based on an intersection of the first candidate motion set and the second candidate motion set.

In a possible implementation, the first word sequence includes a first word and a second word; and the determining, from the first matching result, a group of motion sequences that do not overlap in the temporal dimension to obtain a first motion sequence includes:

determining a first motion and a second motion from alternative motion sets corresponding to the first word and the second word, respectively;

determining start positions of the first word and the second word on a timeline based on a timestamp prediction model; and

determining, based on the start positions and motion duration, whether the first motion and the second motion overlap on the timeline; and if the first motion and the second motion overlap on the timeline, excluding one of the first motion and the second motion from the first motion sequence.

In a possible implementation, the excluding one of the first motion and the second motion from the first motion sequence includes:

excluding the motion corresponding to the word with the later start position from the first motion sequence.

In a possible implementation, the first motion and the second motion further have corresponding priorities; and the excluding one of the first motion and the second motion from the first motion sequence includes:

excluding the motion with the lower priority from the first motion sequence.

In a possible implementation, the method further includes:

obtaining a second motion set, where the second motion set includes a plurality of generic motions, and any generic motion has corresponding motion duration;

extracting words whose part-of-speech is a target part-of-speech from the target text to obtain a second word sequence;

randomly selecting a corresponding generic motion from the second motion set to obtain

a second matching result for any word in the second word sequence; and

determining a group of motion sequences from the second matching result based on positions of words in the second word sequence in the target text and motion duration of generic motions corresponding to the words and inserting the group of motion sequences into the first motion sequence to obtain a second motion sequence to enable motions in the second motion sequence not to overlap in the temporal dimension.

In a possible implementation, the target part-of-speech includes one or more of the following: a quantifier, a verbal noun, an adverbial verb, a common verb, a temporal noun, and a locative word.

In a possible implementation, the randomly selecting a corresponding generic motion from the second motion set to obtain a second matching result for any word in the second word sequence includes:

randomly selecting a generic motion from the second motion set to enable adjacent words in the second word sequence to correspond to different generic motions to obtain the second matching result.

In a possible implementation, the second word sequence includes a third word; and the determining a group of motion sequences from the second matching result and inserting the group of motion sequences into the first motion sequence to obtain a second motion sequence includes:

determining a start position of the third word on a timeline based on a timestamp prediction model;

determining a third motion corresponding to the third word based on the second matching result; and

determining, based on the start position and motion duration of the third motion, whether the third motion overlaps any motion in the first motion sequence on the timeline, and if the third motion does not overlap any motion in the first motion sequence on the timeline, adding the third motion to the first motion sequence.

According to a second aspect, an apparatus for motion orchestration of digital humans is provided, including:

a first obtaining unit, configured to: obtain a first word sequence corresponding to a target text, and obtain a first motion set, where the first motion set includes a plurality of motions, and any motion has a preset trigger word and corresponding motion duration;

a first matching unit, configured to match words in the first word sequence against the first motion set at least based on the trigger word to obtain a first matching result including

alternative motion sets corresponding to the words; and

a first determination unit, configured to determine, from the first matching result based on positions of the words in the target text and the motion duration, a group of motion sequences that do not overlap in the temporal dimension to obtain a first motion sequence, where the first motion sequence is configured to indicate a digital human to make a corresponding motion when delivering the target text.

In a possible implementation, the apparatus further includes:

a second obtaining unit, configured to obtain a second motion set, where the second motion set includes a plurality of generic motions, and any generic motion has corresponding motion duration;

a word extraction unit, configured to extract words whose part-of-speech is a target part-of-speech from the target text to obtain a second word sequence;

a second matching unit, configured to randomly select a corresponding generic motion from the second motion set to obtain a second matching result for any word in the second word sequence; and

a second determination unit, configured to determine a group of motion sequences from the second matching result based on positions of words in the second word sequence in the target text and motion duration of generic motions corresponding to the words and inserting the group of motion sequences into the first motion sequence to obtain a second motion sequence to enable motions in the second motion sequence not to overlap in the temporal dimension.

According to a third aspect, a computer-readable storage medium is provided, having a computer program stored thereon. The computer program, when executed by a computer, causes the computer to perform the method in the first aspect.

According to a fourth aspect, a computing device is provided, including a memory and a processor. The memory has executable code stored thereon. The processor, when executing the executable code, implements the method in the first aspect.

Embodiments of this specification provide a method and apparatus for motion orchestration of digital humans. An appropriate motion is matched for a text based on matching between the text and a preset trigger word of the motion and matching between text semantic encoding and motion encoding. In addition, to ensure coherence between different motions and prevent prolonged silence of a digital human, embodiments of this specification further insert appropriate generic motions between adjacent motions, thereby ensuring the diversity and coherence of motion sequences.

BRIEF DESCRIPTION OF THE DRAWINGS

To describe the technical solutions in multiple embodiments disclosed in this specification more clearly, the following briefly describes the accompanying drawings required for describing the embodiments. It is clear that the accompanying drawings in the following description show merely multiple embodiments disclosed in this specification, and a person of ordinary skill in the art may still derive other drawings from these accompanying drawings without creative efforts.

FIG. 1 is a schematic diagram of an implementation scenario of a method for motion orchestration of digital humans according to an embodiment;

FIG. 2 is a flowchart of a method for motion orchestration of digital humans according to an embodiment;

FIG. 3 is a flowchart of a method for matching a text and a motion according to an embodiment;

FIG. 4 is a schematic diagram of a coarse-grained orchestration phase of motions of a digital human according to an embodiment;

FIG. 5 is a schematic diagram of a timeline in a fine-grained orchestration phase according to an embodiment; and

FIG. 6 is a schematic block diagram of an apparatus for motion orchestration of digital humans according to an embodiment.

DETAILED DESCRIPTION

The solutions provided in this specification are described below with reference to the accompanying drawings.

A digital human, also known as a virtual avatar, is a digital character that is created using digital technology and that closely resembles a human appearance. In practical services, digital humans need to use body motions to assist in conveying information through language during speech delivery, explanation, live streaming, and interaction. In a current common digital human application scenario, body motions typically occur simultaneously with speech/text. Therefore, based on consistency requirements between semantics and motions of the digital human during speaking, a motion orchestration solution is required to enable a digital human to present motions that comply with semantics during speech delivery.

FIG. 1 is a schematic diagram of an implementation scenario of a method for motion orchestration of digital humans according to an embodiment. In an example of FIG. 1, input

data includes text data to be delivered by a digital human and a candidate motion set used for motion orchestration of the digital human. The solution generally includes a coarse-grained orchestration phase, a fine-grained orchestration phase, and a re-orchestration phase.

In the coarse-grained orchestration phase, an intersection/union of a motion set of rule-based matching and a motion set of semantic similarity-based matching is calculated to obtain an alternative motion set. The rule-based matching mainly considers whether a phrase obtained by tokenizing an input text is consistent with a trigger word configured for a motion in the candidate motion set. A matching motion is added to the motion set of rule-based matching, and a specific motion is inserted at a text position specified by a user. The semantic similarity-based matching considers a similarity between a text semantic representation obtained by tokenizing a text and a motion representation in the candidate motion set, and a motion whose similarity is greater than a specified threshold is added to the motion set of semantic similarity-based matching.

The coarse-grained orchestration phase only considers a degree of matching between a text and a motion. However, each motion in the candidate motion set has certain duration, and each character in the input text has a start time and an end time in output audio delivered by the digital human. Therefore, in the fine-grained orchestration phase, to prevent different motions from appearing at the same position of a timeline (that is, a motion overlap), timeline alignment is performed on motions and texts. That is, motions in the alternative motion set are sorted first based on start positions of motions and configured priorities, motions that have high priorities and that do not overlap each other are then kept, and other conflicting motions are deleted.

To ensure coherence between different motions and prevent prolonged silence of a digital human, in the re-orchestration phase, syntactic structure analysis is performed on the input text to obtain phrases of specified syntactic structures, and appropriate generic motions are selected from an input generic motion set, inserted at the positions of specified phrases, and sorted based on priorities configured for motions, thereby ensuring the diversity and coherence of motion sequences.

Specific implementation steps of the foregoing method for motion orchestration of digital humans are described below with reference to specific embodiments. FIG. 2 is a flowchart of a method for motion orchestration of digital humans according to an embodiment. The method may be performed by any platform, server, device cluster, or the like having computing and processing capabilities.

Step 202: Obtain a first word sequence corresponding to a target text, and obtain a first motion set, where the first motion set includes a plurality of motions, and any motion has a

preset trigger word and corresponding motion duration.

The first word sequence may be obtained by tokenizing the target text. Words in the sequence may be single characters, words, phrases or combinations thereof according to actual requirements. Each word in the first word sequence further includes position information, for example, index information, of the word in the target text. The first motion set includes a plurality of motions linked to the text and the environment. The trigger word of the motion is configured to indicate a digital human to make a corresponding motion when delivering a corresponding word.

In an embodiment, to reduce overall computing overheads, in consideration of that a person typically has the habit of making motions at words of specific parts-of-speech during speaking, words of parts-of-speech with low-frequency motions in the first word sequence are removed based on parts-of-speech. Only one or more words of parts-of-speech being a common noun, an adjective, a common verb, an adverbial adjective, an adverbial verb, a nominal adjective, a verbal noun, and an adverb are kept.

Step 204: Match words in the first word sequence against the first motion set at least based on the trigger word to obtain a first matching result including alternative motion sets corresponding to the words.

In an embodiment, the procedure of implementing Step 204 is shown in FIG. 3. FIG. 3 is a flowchart of a method for matching a text and a motion according to an embodiment.

Step 302: For any target word, match the target word against the trigger words corresponding to the motions in the first motion set to obtain a first candidate motion set.

In an embodiment, a user-defined motion set of specific motions inserted by a user at specified text positions is further obtained, and the user-defined motion set is added to the first candidate motion set.

Step 304: Obtain a second candidate motion set based on similarities between a semantic representation of the target word and motion representations of the motions in the first motion set.

In an embodiment, the target word is encoded using a first text encoder to obtain a target semantic representation; a motion representation obtained by encoding any motion in the first motion set using a motion encoder is obtained; and motions whose motion representation has a similarity with the target semantic representation being greater than a first threshold are categorized into the second candidate motion set.

The first text encoder is trained to enable a semantic representation generated by the first text encoder and a motion representation generated by the motion encoder to be located in the

same representation space, so that a similarity between the representations generated by the two encoders can be calculated.

In an embodiment, a second text encoder that has been pre-trained jointly with the motion encoder may exist, so that the representations generated by the two encoders are located in the same representation space. The second text encoder and the first text encoder encode different types of languages. For example, the first text encoder is configured to encode a text in a first language, and the second text encoder is configured to encode a text in a second language. In this case, parallel training corpora in the first language and the second language, that is, training corpora that have the same semantics and that are in the two languages, are obtained. A first training corpus in the first language is encoded using the first text encoder to obtain a first semantic representation. A second training corpus in the second language is encoded using the second text encoder to obtain a second semantic representation. The first training corpus and the second training corpus have the same semantics. A value of a parameter in the first text encoder is then adjusted by maximizing a similarity between the first semantic representation and the second semantic representation, so that the semantic representation generated by the first text encoder and the motion representation generated by the motion encoder are also located in the same representation space.

In a more specific embodiment, the used first language is Chinese, and the second language is English. The first text encoder is a Chinese BERT (Bidirectional Encoder Representation from Builders) encoder, and the second text encoder and the motion encoder are a motionCLIP text encoder and a motionCLIP motion encoder, respectively. In motionCLIP, a motion encoder is trained based on a CLIP (Contrastive Language-Image Pre-training, a contrastive language-image pre-training model) text encoder and a CLIP image encoder, so that the representation space of the motion encoder may be aligned with that of the CLIP text encoder. Because a motionCLIP text encoder only supports encoding in English, when a text delivered by the digital human is Chinese, an encoder supporting Chinese further needs to be trained to enable a representation space of the motionCLIP text encoder to be aligned with that of the motionCLIP motion encoder. Through the use of Chinese and English parallel training corpora, the Chinese BERT encoder is fine-tuned (fine-tuning) by maximizing a cosine similarity between an English representation obtained by the English motionCLIP text encoder and a Chinese representation obtained by the Chinese BERT encoder, so that a representation space of the Chinese BERT encoder is aligned with the representation space of the motionCLIP motion encoder. In other embodiments, the similarity between the English representation obtained by the English motionCLIP text encoder and the Chinese representation obtained by

the Chinese BERT encoder may alternatively be calculated using another similarity calculation method. This is not limited herein.

Step 306: Determine an alternative motion set corresponding to the target word based on the first candidate motion set and the second candidate motion set.

The alternative motion set is determined by calculating an intersection or a union of the first candidate motion set and the second candidate motion set. Another set operation may alternatively be performed on the two sets according to actual requirements, or the alternative motion set is formed by selecting a subset of motions from any set. This is not limited herein.

It should be noted that in some embodiments, because the first motion set has a limited variety of motion types, it is impossible to match a corresponding motion for every word in the first word sequence. In this case, alternative motion sets corresponding to some words are empty sets.

In other embodiments, the words in the first word sequence may alternatively be matched against the first motion set using another matching method. For example, most frequent motions that appear when a person speaks some specific words may be statistically analyzed, and the words are matched against motions based on statistical frequencies. This is not limited herein.

Step 202 and Step 204 may correspond to a coarse-grained orchestration phase in FIG. 1. A schematic diagram of the coarse-grained orchestration phase may be shown in FIG. 4.

Next, return to FIG. 2. Step 206: Determine, from the first matching result based on positions of the words in the target text and the motion duration, a group of motion sequences that do not overlap in the temporal dimension to obtain a first motion sequence, where the first motion sequence is configured to indicate a digital human to make a corresponding motion when delivering the target text.

The group of motion sequences that do not overlap in the temporal dimension may be determined in various manners. In an embodiment, for each word, one motion is selected from an alternative motion set corresponding to the set as a corresponding motion of the word. The corresponding motions of the words in the first word sequence may be sequentially sorted into an initial motion sequence. Some motions are then discarded from the initial motion sequence based on overlapping of the motions on a timeline, and are excluded from the first motion sequence, to finally form the first motion sequence without overlaps.

Two words are used as an example for description below. Specifically, the first word sequence includes a first word and a second word, a first motion and a second motion are determined from alternative motion sets corresponding to the first word and the second word,

respectively, and start positions of the first word and the second word on a timeline are then determined based on a timestamp prediction model. The timestamp prediction model may be one submodule in a text-to-speech model TTS (Text To Speech), or may be an independent model. This is not limited herein. After the start positions of the two words are obtained, the start positions of the corresponding words are set as start positions of the two motions, respectively. It is determined based on the start positions and motion duration whether the first motion and the second motion overlap on the timeline. If the first motion and the second motion overlap on the timeline, one of the first motion and the second motion is excluded from the initial motion sequence, that is, excluded from the first motion sequence. For example, the motion corresponding to the word with the earlier start position may be kept, and the motion with the later start position is excluded. Alternatively, when the first motion and the second motion further has corresponding priorities, the motion with the higher priority is kept, and the motion with the lower priority is excluded from the first motion sequence. Any two words that overlap on the timeline may be processed using the method in this embodiment.

In another embodiment, the motions corresponding to the words are first sorted based on a start time order as described above to obtain a motion sequence M. When the motions have preset priorities, the motions are sorted based on both a start time order and priorities to obtain a motion sequence M. The first motion sequence is then updated through multiple rounds of iteration. In every round, one motion is selected from the motion sequence M and added to the first motion sequence, and a plurality of motions are then removed from the motion sequence M. Multiple rounds of iteration are repeated until the motion sequence M is empty. Any round of iteration includes:

- selecting the foremost motion in the motion sequence M, and adding the motion to the first motion sequence;

- placing the motion on the timeline based on a start position of the motion; and

- selecting, based on the start position and motion duration of the motion, all motions in the motion sequence M that overlap the motion on the timeline, removing the motions from the motion sequence M, and then performing a next round of iteration.

One group of motion sequences that do not overlap in the temporal dimension may be determined using the method in this embodiment.

In a specific example, a schematic diagram of a timeline in a fine-grained orchestration phase shown in FIG. 5 includes a text sequence and a motion sequence. Duration of words and motions are represented by lengths of corresponding rectangles. An input text is "Hello, I am a digital human." The text is tokenized to obtain a word sequence "Hello", ",", "I", "am", "a",

"digital human", and ".", corresponding motions are matched for the words in the word sequence using the method as described above, and start positions corresponding to the motions are determined to forming a motion sequence M.

The foremost motion, that is, a motion 1 is then selected from the motion sequence M first. The motion 1 is added to the first motion sequence and placed on the timeline. It may be seen that duration of the motion 1 covers the words "Hello", ",", and "I". The motions corresponding to these words overlap the motion 1. In this case, the motions corresponding to these words are removed from the motion sequence M. Next, the foremost motion, that is, a motion 2, is selected from the remaining motions in the motion sequence M. The motion 2 is added to the first motion sequence and placed on the timeline. Duration of the motion 2 covers the words "is" and "a". In this case, the motions corresponding to these words are removed from the motion sequence M. The foremost motion, that is, a motion 3, is then selected from the remaining motions in the motion sequence M. The motion 3 is added to the first motion sequence and placed on the timeline, and the motions corresponding to the words "digital human" and "." are removed from the motion sequence M. In this case, the motion sequence M is empty, motion orchestration iteration is ended, and a first motion sequence is obtained: the motion 1, the motion 2, and the motion 3.

In other embodiments, a group of motion sequences that do not overlap in the temporal dimension may alternatively be determined from the first matching result using another method, provided that no overlap exists. For example, a midpoint of motion duration is placed at a start position of a corresponding word. In other words, the digital human has started to make the motion corresponding to the word before delivering the word. A group of motion sequences that do not overlap in the temporal dimension are then determined based on this placement manner.

Step 206 may correspond to the fine-grained orchestration phase in FIG. 1. A group of digital human motion sequences without conflicts in time may be obtained using the method for motion orchestration of digital humans shown in FIG. 2 to meet requirements of motion orchestration. However, in some embodiments, and matching motions cannot be found from the first motion set for all words in a text. As a result, a prolonged blank (that is, a silence motion) may exist between adjacent motions in the first motion sequences generated in coarse-grained orchestration and fine-grained orchestration, leading to undiversified motions and poor user experience of a digital human in a text delivery process.

To improve the diversity of motion sequences, in some possible implementations, the method further includes the following steps.

Step 208: Obtain a second motion set, where the second motion set includes a plurality of generic motions, and any generic motion has corresponding motion duration.

The generic motions are weak-semantic motions for which no trigger words are configured. These motions do not appear abrupt when paired with any delivered copy, and therefore are suitable for use as transitional motions between intermediate- and strong-semantic motions in the first motion set, thereby improving the diversity of motions in an entire motion orchestration sequence.

Step 210: Extract words whose part-of-speech is a target part-of-speech from the target text to obtain a second word sequence.

In an embodiment, a syntactic analysis is performed on the target text to extract the words whose part-of-speech is the target part-of-speech from the target text. The target part-of-speech includes one or more of the following: a quantifier, a verbal noun, an adverbial verb, a common verb, a temporal noun, and a locative word.

Step 212: Randomly select a corresponding generic motion from the second motion set to obtain a second matching result for any word in the second word sequence.

In an embodiment, to keep the digital human from repeating the same generic motion, a generic motion is randomly selected from the second motion set to enable adjacent words in the second word sequence to correspond to different generic motions to obtain the second matching result.

Step 214: Determine a group of motion sequences from the second matching result based on positions of words in the second word sequence in the target text and motion duration of generic motions corresponding to the words and insert the group of motion sequences into the first motion sequence to obtain a second motion sequence to enable motions in the second motion sequence not to overlap in the temporal dimension.

In an embodiment, the second word sequence includes a third word. A start position of the third word on a timeline is first determined based on a timestamp prediction model. A third motion corresponding to the third word is determined based on the second matching result. It is then determined based on the start position and motion duration of the third motion whether the third motion overlaps any motion in the first motion sequence on the timeline. If the third motion does not overlap any motion in the first motion sequence on the timeline, the third motion is added to the first motion sequence.

In other embodiments, a group of motion sequences may alternatively be determined from the second matching result and inserted into the first motion sequence using another method, provided that no overlap exists. For example, a midpoint of motion duration is placed at a start

position of a corresponding word, and a group of motion sequences that do not overlap the first motion sequence in the temporal dimension are determined based on this placement manner and inserted into the first motion sequence.

Step 208 to Step 214 may correspond to a re-orchestration phase in FIG. 1. A prolonged blank between adjacent motions can be prevented through Step 208 to Step 214.

One or more embodiments of this specification provide determining a motion sequence of a digital human during speech delivery of a text in a manner of performing coarse-grained orchestration-fine-grained orchestration-re-orchestration based on a preset motion library. This manner based on the preset motion library can ensure the quality of candidate motions, thereby avoiding low-quality motions and incoherent motions in an output motion sequence. The semantic consistency and controllability of a motion orchestration result is improved in a manner of manually presetting trigger words for motions. In addition, an inference method that measures a text representation and a motion representation further improves the semantic consistency of an orchestrated motion sequence and an input text. For the problem that certain duration in motions in the preset motion library leads to an overlap on a timeline of audio in an orchestrated motion sequence, a timeline alignment method via backward extension from a start position of a word is provided, so that the problem that motions are unusable due to overlaps during speech delivery is prevented while ensuring the semantic consistency between a motion sequence and a text. In consideration of the problem that a motion sequence obtained based on a semantic matching method is sparse, a syntactic analysis-based method is provided to insert a generic transitional motion in a specific syntactic structure, thereby improving the diversity of orchestrated motion sequences.

According to another aspect of embodiments, an apparatus for motion orchestration of digital humans is further provided. FIG. 6 is a schematic block diagram of an apparatus for motion orchestration of digital humans according to an embodiment. The apparatus may be deployed in any device, platform or device cluster having computing and processing capabilities. As shown in FIG. 6, the apparatus 600 includes:

- a first obtaining unit 601, configured to: obtain a first word sequence corresponding to a target text, and obtain a first motion set, where the first motion set includes a plurality of motions, and any motion has a preset trigger word and corresponding motion duration;

- a first matching unit 602, configured to match words in the first word sequence against the first motion set at least based on the trigger word to obtain a first matching result including alternative motion sets corresponding to the words; and

- a first determination unit 603, configured to determine, from the first matching result based

on positions of the words in the target text and the motion duration, a group of motion sequences that do not overlap in the temporal dimension to obtain a first motion sequence, where the first motion sequence is configured to indicate a digital human to make a corresponding motion when delivering the target text.

In some possible implementations, the apparatus 600 further includes:

a second obtaining unit 604, configured to obtain a second motion set, where the second motion set includes a plurality of generic motions, and any generic motion has corresponding motion duration;

a word extraction unit 605, configured to extract words whose part-of-speech is a target part-of-speech from the target text to obtain a second word sequence;

a second matching unit 606, configured to randomly select a corresponding generic motion from the second motion set to obtain a second matching result for any word in the second word sequence; and

a second determination unit 607, configured to determine a group of motion sequences from the second matching result based on positions of words in the second word sequence in the target text and motion duration of generic motions corresponding to the words and inserting the group of motion sequences into the first motion sequence to obtain a second motion sequence to enable motions in the second motion sequence not to overlap in the temporal dimension.

According to another aspect of embodiments, a computer-readable storage medium is further provided, having a computer program stored thereon. The computer program, when executed by a computer, causes the computer to perform the method described in any foregoing embodiment.

According to yet another aspect of embodiments, a computing device is further provided, including a memory and a processor. The memory has executable code stored thereon. The processor, when executing the executable code, implements the method described in any foregoing embodiment.

The embodiments of this specification are all described in a progressive manner, for same or similar parts in the embodiments, refer to these embodiments, and descriptions of each embodiment focus on a difference from other embodiments. Especially, an apparatus embodiment is basically similar to a method embodiment, and therefore is described briefly; for related parts, refer to partial descriptions in the method embodiment.

Embodiments of this specification are described above. Other embodiments fall within the scope of the appended claims. In some cases, the actions or steps recorded in the claims may

be performed in sequences different from those in the embodiments and an expected result may still be achieved. In addition, the processes depicted in the accompanying drawings are not necessarily performed in the specific order or successively to achieve an expected result. In some implementations, multitasking and parallel processing may be feasible or beneficial.

It should be noted that the relational terms herein such as first and second are used only to differentiate an entity or operation from another entity or operation, and do not require or imply any actual relationship or sequence between these entities or operations. Moreover, the terms "include," "comprise," and any variation thereof are intended to cover a non-exclusive inclusion. Therefore, in the context of a process, a method, an object, or an apparatus that includes a series of elements, the process, method, object, or apparatus not only includes such elements, but also includes other elements not specified expressly, or may include inherent elements of the process, method, object, or apparatus. Without more limitations, elements defined by the sentence "including one ..." does not exclude that there are still other same elements in the processes, methods, objects, or apparatuses.

A person of ordinary skill in the art may understand that all or a part of the steps of the embodiments may be implemented by hardware or a program instructing relevant hardware. The program may be stored in a computer-readable storage medium. The storage medium may be a read-only memory, a magnetic disk, an optical disc or the like.

The foregoing specific implementations further describe the objectives, technical solutions in detail, and beneficial effects of the present invention. It should be appreciated that the foregoing descriptions are merely specific implementations of the present invention, but are not intended to limit the protection scope of the present invention. Any modification, equivalent replacement, improvement, and the like made within the spirit and principle of the present invention should fall within the protection scope of the present invention.

CLAIMS

1. A method for motion orchestration of digital humans, comprising:

obtaining a first word sequence corresponding to a target text, and obtaining a first motion set, wherein the first motion set comprises a plurality of motions, and any motion has a preset trigger word and corresponding motion duration;

matching words in the first word sequence against the first motion set at least based on the trigger word to obtain a first matching result comprising alternative motion sets corresponding to the words; and

determining, from the first matching result based on positions of the words in the target text and the motion duration, a group of motion sequences that do not overlap in the temporal dimension to obtain a first motion sequence, wherein the first motion sequence is configured to indicate a digital human to make a corresponding motion when delivering the target text.

2. The method according to claim 1, wherein the matching words in the first word sequence against the first motion set to obtain a first matching result comprises:

for any target word, matching the target word against the trigger word to obtain a first candidate motion set;

obtaining a second candidate motion set based on similarities between a semantic representation of the target word and motion representations of the motions in the first motion set; and

determining an alternative motion set corresponding to the target word based on the first candidate motion set and the second candidate motion set.

3. The method according to claim 2, wherein the obtaining a second candidate motion set based on similarities between a semantic representation of the target word and motion representations of the motions in the first motion set comprises:

encoding the target word using a first text encoder to obtain a target semantic representation;

obtaining a motion representation obtained by encoding any motion in the first motion set using a motion encoder; and

categorizing motions whose motion representation has a similarity with the target semantic representation being greater than a first threshold into the second candidate motion set.

4. The method according to claim 3, wherein the first text encoder is configured to encode a text in a first language; the motion encoder is pre-trained jointly with a second text encoder

suitable for a second language; and the first text encoder is trained in the following manner:

encoding a first training corpus in the first language using the first text encoder to obtain a first semantic representation;

encoding a second training corpus in the second language using the second text encoder to obtain a second semantic representation, wherein the second training corpus and the first training corpus have the same semantics; and

adjusting a value of a parameter in the first text encoder by maximizing a similarity between the first semantic representation and the second semantic representation.

5. The method according to claim 2, wherein the determining an alternative motion set corresponding to the target word based on the first candidate motion set and the second candidate motion set comprises:

determining the alternative motion set based on a union of the first candidate motion set and the second candidate motion set; or

determining the alternative motion set based on an intersection of the first candidate motion set and the second candidate motion set.

6. The method according to claim 1, wherein the first word sequence comprises a first word and a second word; and the determining, from the first matching result, a group of motion sequences that do not overlap in the temporal dimension to obtain a first motion sequence comprises:

determining a first motion and a second motion from alternative motion sets corresponding to the first word and the second word, respectively;

determining start positions of the first word and the second word on a timeline based on a timestamp prediction model; and

determining, based on the start positions and motion duration, whether the first motion and the second motion overlap on the timeline; and if the first motion and the second motion overlap on the timeline, excluding one of the first motion and the second motion from the first motion sequence.

7. The method according to claim 6, wherein the excluding one of the first motion and the second motion from the first motion sequence comprises:

excluding the motion corresponding to the word with the later start position from the first motion sequence.

8. The method according to claim 6, wherein the first motion and the second motion further have corresponding priorities; and the excluding one of the first motion and the second motion from the first motion sequence comprises:

excluding the motion with the lower priority from the first motion sequence.

9. The method according to claim 1, further comprising:

obtaining a second motion set, wherein the second motion set comprises a plurality of generic motions, and any generic motion has corresponding motion duration;

extracting words whose part-of-speech is a target part-of-speech from the target text to obtain a second word sequence;

randomly selecting a corresponding generic motion from the second motion set to obtain a second matching result for any word in the second word sequence; and

determining a group of motion sequences from the second matching result based on positions of words in the second word sequence in the target text and motion duration of generic motions corresponding to the words and inserting the group of motion sequences into the first motion sequence to obtain a second motion sequence to enable motions in the second motion sequence not to overlap in the temporal dimension.

10. The method according to claim 9, wherein the target part-of-speech comprises one or more of the following: a quantifier, a verbal noun, an adverbial verb, a common verb, a temporal noun, and a locative word.

11. The method according to claim 9, wherein the randomly selecting a corresponding generic motion from the second motion set to obtain a second matching result for any word in the second word sequence comprises:

randomly selecting a generic motion from the second motion set to enable adjacent words in the second word sequence to correspond to different generic motions to obtain the second matching result.

12. The method according to claim 9, wherein the second word sequence comprises a third word; and the determining a group of motion sequences from the second matching result and inserting the group of motion sequences into the first motion sequence to obtain a second motion sequence comprises:

determining a start position of the third word on a timeline based on a timestamp prediction model;

determining a third motion corresponding to the third word based on the second matching result; and

determining, based on the start position and motion duration of the third motion, whether the third motion overlaps any motion in the first motion sequence on the timeline, and if the third motion does not overlap any motion in the first motion sequence on the timeline, adding the third motion to the first motion sequence.

13. An apparatus for motion orchestration of digital humans, comprising:

a first obtaining unit, configured to: obtain a first word sequence corresponding to a target text, and obtain a first motion set, wherein the first motion set comprises a plurality of motions, and any motion has a preset trigger word and corresponding motion duration;

a first matching unit, configured to match words in the first word sequence against the first motion set at least based on the trigger word to obtain a first matching result comprising alternative motion sets corresponding to the words; and

a first determination unit, configured to determine, from the first matching result based on positions of the words in the target text and the motion duration, a group of motion sequences that do not overlap in the temporal dimension to obtain a first motion sequence, wherein the first motion sequence is configured to indicate a digital human to make a corresponding motion when delivering the target text.

14. The apparatus according to claim 13, further comprising:

a second obtaining unit, configured to obtain a second motion set, wherein the second motion set comprises a plurality of generic motions, and any generic motion has corresponding motion duration;

a word extraction unit, configured to extract words whose part-of-speech is a target part-of-speech from the target text to obtain a second word sequence;

a second matching unit, configured to randomly select a corresponding generic motion from the second motion set to obtain a second matching result for any word in the second word sequence; and

a second determination unit, configured to determine a group of motion sequences from the second matching result based on positions of words in the second word sequence in the target text and motion duration of generic motions corresponding to the words and inserting the group of motion sequences into the first motion sequence to obtain a second motion sequence to enable motions in the second motion sequence not to overlap in the temporal dimension.

15. A computer-readable storage medium, having a computer program stored thereon, wherein the computer program, when executed by a computer, causes the computer to perform the method according to any one of claims 1 to 12.

16. A computing device, comprising a memory and a processor, wherein the memory has executable code stored thereon, and the processor, when executing the executable code, implements the method according to any one of claims 1 to 12.