Publication number	WO2007098560 A1
Publication type	Application
Application number	PCT/AU2007/000260
Publication date	Sep 7, 2007
Filing date	Mar 2, 2007
Priority date	Mar 3, 2006
Publication number	PCT/2007/260, PCT/AU/2007/000260, PCT/AU/2007/00260, PCT/AU/7/000260, PCT/AU/7/00260, PCT/AU2007/000260, PCT/AU2007/00260, PCT/AU2007000260, PCT/AU200700260, PCT/AU7/000260, PCT/AU7/00260, PCT/AU7000260, PCT/AU700260, WO 2007/098560 A1, WO 2007098560 A1, WO 2007098560A1, WO-A1-2007098560, WO2007/098560A1, WO2007098560 A1, WO2007098560A1
Inventors	Byl Penny-Anne Kaye De, Christopher William Mills
Applicant	Byl Penny-Anne Kaye De, Christopher William Mills, The University Of Southern Queensland, Less «2 More »

Patent Citations (2), Non-Patent Citations (6), Events (3)

External Links: Patentscope, Espacenet

An emotion recognition system and method

WO 2007098560 A1

ABSTRACT

A system and method or recognizing emotions in an input data stream. The method commences when the emotion recognition system of the invention receives an input data stream from a user. The emotion recognition system then extracts an emotional state, in the form of an emotional state vector, from the input data stream. The emotion recognition system then updates a current control point using the emotional state vector, the current control point being a multidimensional vector in an affective space representing a current emotional state.

CLAIMS (OCR text may contain errors)

1. A method of recognizing emotions in an input data stream, the method including the steps of:

(i) receiving the input data stream; (ii) extracting an emotional state in the form of an emotional state vector from the input data stream; and

(iii) updating a current control point utilizing the emotional state vector extracted from the input data stream, the current control point being a point in a multidimensional affective space representing a current emotional state.

2. The method of claim 1 further including the steps of:

(iv) communicating the current control point to an avatar control module; and

(v) gesticulating an avatar under control of the avatar control module based upon the current control point.

3. The method of claim 1 further including the steps of:

(iv) communicating the control point together with textual and/or audio data forming part of the input data stream to an avatar control module; and (v) gesticulating an avatar under control of the avatar control module based upon the current control point, the gesticulation of the avatar being synchronized with the textual and/or audio data forming part of the input data stream.

4. The method of claim 1 , wherein the emotional state vector is continuously extracted from the input data stream such that the current control point is continuously updated.

5. The method of claim 1, wherein the affective space is in the form of a multidimensional space

defined by two or more orthogonal valence appraisals, each orthogonal valence appraisal being representative of a human emotion.

6. The method of claim 1 , wherein the affective space is in the form of multidimensional space defined by six orthogonal valence appraisals, each orthogonal valence appraisal corresponding to one of the human emotions for happiness, sadness, anger, fear, disgust and surprise.

7. The method of claim 5, wherein the current control point is updated by combing the emotional state vector extracted from the input data stream with the current control point in the multidimensional affective space.

8. The method of claim 5, wherein the current control point is updated by combing the emotional state vector extracted from the input data stream with the current control point in the multidimensional affective space using the formula:

$$CP_t = \frac{W_0 CP_{t-1} + W_1 R_t}{2}$$

where 0 < J^$_0$ < 1 , 0 < J^$_1$ < 1 and J^$_0$ +]y$_X$ = 1 and CP is a value of the control

point value at a time t, and R is a resulting neural network vector and W_0 and Wi are each predetermined weightings assigned to the current control point prior to the current control point being updated and the emotional state vector respectively.

9. The method of claim 7 further including the step of: (iv) calculating a dominant emotion of the current control point

10. The method of claim 9, wherein the dominant emotion is calculated by determining the Euclidean distance between the current control point and each of the orthogonal valence appraisals of the affective space.

11 The method of claim 1 further including the steps, prior to step (ii), of:

(a) determining the type of input data stream; and

(b) selecting an appropriate extraction method based upon the type of input data stream.

12. The method of claim 1 , wherein the input data stream is an audio input data stream type and step (ii) includes the sub-steps of:

(a) splicing the input data stream into one or more frames;

(b) capturing each frame using a digital signal processing application; (c) extracting one or more audio characteristics from each frame;

(d) averaging the audio characteristics for a subset of the captured frames, the subset forming an utterance;

(e) communicating the averaged audio characteristics of the utterance to a neural network, the neural network being pre-trained using a plurality of frames of pre-recorded audio data representative of one or more human emotions; and

(f) creating the emotional state vector based upon an output of the neural network.

13. The method of claim 12, wherein the method continuously creates an emotional state vector for each utterance in the audio input data stream.

14. The method of claim 1 , wherein the input data stream is a textual input data stream type and step (ii) includes the sub-steps of: (a) parsing the textual input data stream to identify emoticons; and

(b) creating the emotional state vector based upon the emotional type of the identified emoticons.

15. The method of claim 1 , wherein the input data stream is an affective state control data stream type which includes an emotional state vector representative of an emotion of a user, the emotional state vector representative of an emotion of the user being generated by the user selecting an emotional state on a display provided to the user.

16. The method of claim 15, wherein the display is in the form of an emotional compass having a plurality of indicia, each indicia being representative of an emotion of human.

17. The method of claim 16, wherein the relative size of each indicia provides an indication as to the location of the current control point in the multidimensional affective space.

18. The method of claim 2, wherein the avatar is gesticulated by selecting one or more of a plurality of avatar emotional accent modifications stored within an avatar data store, the one or more avatar emotional accent modifications being selected based upon the location of the current control point in the multidimensional affective space.

19. An emotion recognition system comprising: one or more data analysis modules configure to receive a data input stream and extract an emotional state vector therefrom; and an affective space state module configure to maintain a current control point in a multidimensional affective space, the current control point representative of an emotional state; wherein, the affective space state module is configured to update the location of the current control point in the multidimensional affective space based upon the emotional state vector extracted from the data input stream.

Dated this 2nd day of March 2007

THE UNIVERSITY OF SOUTHERN QUEENSLAND

By their Patent Attorneys

FISHER ADAMS KELLY

DESCRIPTION (OCR text may contain errors)

TITLE "AN EMOTION RECOGNITION SYSTEM AND METHOD"

FIELD OF THE INVENTION

The invention relates to an emotion recognition system and method for extracting emotional expression from an input data stream. In particular, although not exclusively, the invention relates to an emotion recognition system and method for extracting emotional expression from an input data stream supplied by a user of a virtual environment in order to automatically gesticulate an avatar embodying the user in the virtual environment. BACKGROUND TO THE INVENTION

Virtual embodiment is an important part of collaborative virtual environments. This embodiment is a critical dimension in ensuring such environments augment intelligent interactions by amplifying the cognitive processes of the users. This stimulates the advancement of virtual reality systems.

The goal of embodiment in virtual reality has been to establish a form of presence. Presence is the creation of a suspension of disbelief where the user willingly accepts the virtual environment that they are in as real and tangible.

Such a sense of presence can be quite unstable. A user immersed in a virtual environment could be snapped out of their suspension of disbelief by the program crashing, badly drawn graphics or unbelievable character behaviours in the same way that a loud bang might interrupt a daydream. The minimum level of acceptable presence occurs when users feel that a form, behaviour, or sensory experience in the virtual environment indicates the presence of another intelligent being. The two practical design problems in creating a sense of presence for intelligent interactions in virtual environments are:

1) using telecommunications to transport and display illusions of the other person; and 2) creating a virtual embodiment for social presence.

The first relates to the transmission and coordination of virtual environments in which geographically distant users can communicate in real time in the same virtual space with virtual embodiments that transmit verbal and non-verbal communication. The goal of the second is to create an artificial agent that can mimic morphology, motion and communication of behaviours on behalf of their controlling agent or user. Such an agent is widely known in virtual reality and computer game domains as an avatar.

Avatars are virtual representations of users in virtual environments. These avatars can be simple static pictures on web pages or animated two or three dimensional characters existing within a two or three dimensional virtual world. They act as puppets for the users to control. To evoke a sense of virtual presence in all users and generate a social community, it is not the story taking place or the avatar's appearance that is most important, but perceived realistic behaviour (believability) from the avatars. When a user feels presence in a virtual environment some of their perceptions, thought processes and emotional responses are identical to those found in human to human interaction.

Traditionally, avatars have been passive graphical objects acting only when directed to by their user. However, recently, interest in creating semi- autonomous avatars that can take minimal directions from their user has arisen as many avatar behaviours that produce non-verbal communications are far too complex for a user to control.

In addition, if the user is engaged in a conversation or making a presentation in a virtual environment, coordinating the actions of their avatar whilst orchestrating a visual presentation with text chat or Voice over IP (VoIP) becomes a difficult exercise. Either the user is inclined to have their avatar stand in the environment unanimated while the presentation is taking place or they decline to use the virtual environment all together. While the later is somewhat unattractive to creators of virtual environments, the former dramatically reduces a user's ability to communicate effectively with their virtual audience and reduces their presence.

As it is undesirable to provide a virtual environment system that produces avatar behaviours that are unrelated to the emotional state or attitudes of the user, current efforts to integrate levels of autonomy in user controlled avatars has included methods whereby the user has varying levels of control over the avatar.

One such method known in the art is the Demeanour Framework as discussed by Giles and Ballin (Gillies, M. & Ballin, D., 2004, Integrating Autonomous Behavior and User Control for Believable Agents, in Proceedings for Third International Joint Conference on Autonomous Agents and Multiagent Systems, Vol. 1., pp. 336-343). This framework provides the user with three levels of control; real-time, profiling and behaviour language.

The real-time control allows a user to control the actions of the avatar while they are in the virtual environment. Profiling is an offline tool that allows the user to customise the personality type of their avatar and thus have it auto- generate behaviour towards other avatars based on its personality. The behaviour language allows further low level definitions of the avatar's reactions to events and other avatars in the environment.

Another alternative to full direct control of an avatar is the concept of influence as discussed by VaIa (VaIa, M., Paiva, A. & Prada, R.: Tangible Influence: Towards a New Interaction Paradigm for Computer Games, ICEC- lnternational Conference on Entertainment Computing, Springer, 2004).

Rather than controlling the behaviour of an avatar through specific action instructions, an avatar is placed under the emotional influence of an external input device.

It is known in the art to use an input device in the form of a rag doll fitted with sensors. This device allows the user to inflict emotional states on the avatar by manipulating the pose of the rag doll. The rag doll can relay six emotions to the avatar; fear, disgust, joy, sadness, anger and surprise. As previously stated, a user in a virtual environment, busy collaborating with another user, has little time or inclination to give extraneous directions regarding semi-autonomous behaviour in real time to their avatar. However, having their avatar remain idle during this time affects their presence in the virtual environment and their social interaction with others. As emotions are a key constituent of face-to-face human interaction, it is desirable that they be an integrated part of avatar-to-avatar interactions and be used to semi-automate a user's avatar in order to improve communication by disambiguating the meaning of utterances and portray the user's mood to another.

Thus, it is desirable in a virtual environment to provide believability and augmented intelligence to ensure that users can express themselves as thoroughly in the virtual world as they do in the real world through speech, gestures and emotions.

This brings forth the problem of having the avatar recognise the emotional state of the user and behaving accordingly. One method known in the art of identifying a user's emotional state is via the use of bio-sensors attached to the body to measure bio-signals such as heart-rate, skin conductivity, temperature, respiration and others. This has been discussed, for example, by Haag (Haag A., Goronzy, S., Schaich, P., Williams, J., 2004, Emotion Recognition Using Bio-sensors: First Steps towards an Automatic System, Affective Dialogue Systems, Proceedings. Lecture Notes in Computer Science 3068 Springer, pp. 36-48). Using this technology, researchers have been able to achieve 96.6% accuracy of classifying a user's emotional state by analysing the readings with a neural network.

Another method known in the art uses facial recognition to transcribe the user's emotional expression from the user and onto the avatar. This system places markers on the users face at key points and a video captured image is analysed for the marker position. Key facial points are then translated onto a virtual avatars face thus replicating the user's own facial expression.

The problem with the preceding methods is that they require sensors to be attached to the body. These apparatus are not inconspicuous, require setup and calibration and do not come as standard peripheral computing devices. In addition, the facial recognition system is sensitive to rapid head and body movements.

It has been recognised in the art that a more ideal method of recognising a user's emotional state through their natural interaction within a virtual environment is through speech analysis. One known method analyses the linguistic content of utterances (i.e. the meaning of the words being spoken) for emotional keywords. However, this method has disadvantages in that when a user is already using their own voice to speak through the avatar the processing of speech to text and back again places a huge demand on processing, can be inaccurate and not achievable in real time.

Although much work has been done in this area in the past, further improvements are possible and desirable to bring together linguistic and verbal cue analysis of text and speech in a single system that can be applied to semi- automated avatars.

OBJECT OF THE INVENTION

It is an object of the invention to overcome or at least alleviate one or more of the above problems and/or provide the consumer with a useful or commercial choice. DISCLOSURE OF THE INVENTION

In one form, although it need not be the only or indeed the broadest form, the invention resides in a method of recognizing emotions in an input data stream, the method including the steps of: (i) receiving the input data stream; (ii) extracting an emotional state in the form of an emotional state vector from the input data stream; and

(iii) updating a current control point utilizing the emotional state vector extracted from the input data stream, the current control point being a point in a multidimensional affective space representing a current emotional state. Further features of the present invention will become apparent from the following detailed description.

BRIEF DESCRIPTION OF THE DRAWINGS

To assist in understanding the invention and to enable a person skilled in the art to put the invention into practical effect preferred embodiments of the invention will be described by way of example only with reference to the accompanying drawings, wherein:

FIG 1 shows a schematic of an emotion recognition system according to an embodiment of the invention;

FIG 2 shows a method of emotion recognition according to an embodiment of the invention;

FIG 3 shows the process involved in the step of extracting emotional states from an input data stream forming part of the method shown in FIG 2 in more detail;

FIG 4 shows the process involved in the step of extracting the emotional intent of an audio input data stream forming part of the method shown in FIG 3 in more detail;

FIG 5 shows the process involved in the step of extracting the emotional intent of a text input data stream forming part of the method shown in FIG 3 in more detail; FIG 6 shows a representation of an interface for communicating and displaying an affective space state of a user;

FIG 7 shows the process involved in the step of updating an emotional control point in an affective space state module shown in FIG 2 in more detail;

FIG 8 shows the process involved in the step of generating avatar control commands shown in FIG 2 in more detail; FIG 9 shows a table of example visemes for a mouthshape phoneme database;

FIG 10 shows a table of example gestures for a gesture database; and FIG 11 shows a table of example facial expressions selected from a facial expression database.

DETAILED DESCRIPTION OF THE INVENTION

As discussed above, emotions are an important part of face to face communication and become even more important in online collaborative environments where users communicating with one another cannot see each other but are embodied by virtual avatars.

The emotion recognition system and method of the invention provides an enhancement to a user's presence in a virtual environment and has particular application to the control of avatars in a virtual reality system capable of supporting collaboration and communication between users. The emotion recognition system and method of the invention uses textual, audio and direct commands provided by a user and converts these input streams into an appropriate emotional gesture to be expressed by an avatar in a virtual environment together with the content of the input data stream.

Once the emotional state of the user is determined, the system of the invention provides feedback of its analysis on a display device, such as a monitor, for the user to assess the system's accuracy. At this point the user can adjust this result as they desire. This input is then fed back into the emotion recognition system to allow the system to learn about the user's unique speech patterns to better analyse their utterances. In this way, the system adapts to the particular user. FIG 1 shows a schematic of an emotion recognition system 100 according to an embodiment of the invention. Emotion recognition system 100 comprises a voice analysis module 110 and a textual analysis module 120 each in communication with an affective space state module 130. Also shown in FIG 1 are data input devices 200 that provide data input streams 210 to the emotion recognition system 100. The input devices 200 are in the form of a digital file 201A containing an audio voice recording or a microphone 202A for capturing and digitizing a users voice and transmitting an audio input data stream 210A to the voice analyzer 110 of the emotion recognition system 100.

Additionally, the data input devices 200 may further comprise a digital file 201 B containing text and emoticons or a key board 202B for capturing characters typed by the user. The text and emoticons captured from the data input devices 201 B and 202B are transmitted in a text input data stream 210B to the text analysis module 120 of the emotion recognition system 100.

Furthermore, the data input devices 200 further comprise a computer mouse 201 C or other known computer peripheral devices such as a keyboard 202C for transmitting affective state control commands in an affective state control input data stream 210C to the affective state module 130 of the emotion recognition system 100. Furthermore, the affective state control data stream 210C provides a representation back to the user of the current affective emotional state of the user as determined by the emotion recognition system 100. This representation is preferably a graphical representation as will be discussed in greater detail below. The voice analysis module 110 of the emotion recognition system 100 receives the audio input data stream 210A and processes the audio characteristics of the audio input data stream 210 to a continuous stream of emotional states. This process will be discussed in more detail below.

The text analysis module 120 of the emotion recognition system 100 receives the text input data stream 210B in order that the emotional characteristics of the text input data stream 210B are extracted.

The affective space state module 130 is a multidimensional space representing multiple independent emotions across cognitive appraisals. The appraisals are measurable values used to plot the coordinates of an emotion in the space. Preferably, in the emotion recognition system 100 of the invention, the affective space state module is a multi dimensional space defined by six orthogonal valence appraisals directly linked to the six universal emotions of happiness, sadness, anger, fear, disgust and surprise. The dimensions are bound by 0 and 1 where 0 indicates no valence with that emotion and 1 full valence. Hence, the valence dimensions are represented by a vector in six dimensional space thus:

valencyForHappiness (1, 0, 0, 0, 0, 0) valencyForSadness (0, 1 , 0, 0, 0, 0) valencyForAnger (0, 0, 1 , 0, 0, 0) valencyForFear (0, 0, 0, 1 , 0, 0) valencyForDisgust (0, 0, 0, 0, 1 , 0) valencyForSurprise (0, 0, 0, 0, 0, 1)

A control point is used in the affective space state module 130 to dynamically represent the current emotional state of the user in the 6 dimensional emotional state vector based on data received from the voice analysis module 110, text analysis module 120 and input from the user via the affective state control data stream 210C. Also shown in FIG 1 are acoustic phoneme module 300, textual phoneme module 400, avatar control module 500, avatar profile database 600, animation database 700 and display module 800. The function of these modules will be discussed in greater detail below.

FIG 2 shows a method of emotion recognition 1000 according to an embodiment of the invention.

The emotion recognition system 100 receives an input data stream 210 from an input data device 200 (step 1100). As previously discussed, this input data stream 210 may be in the form of an audio input data stream 210A, a text input data stream 210B or an affective space state control input data stream 210C.

The emotion recognition system 100 then processes the input data stream 210 and continuously extracts the emotional data from the input data stream (step 1200).

The emotional recognition system 100 then updates the control point of the affective space state module 130 based on the emotional data extracted from the input data stream (step 1300). In this way, the emotion recognition system 100 maintains a dynamic state of the user's emotional state.

The textual and/or audio content from the input data stream 210 together with the current control point of the affective space state module 130 is communicated to the avatar control module 500 (step 1400). The avatar control module 500 then gesticulates the avatar and has the avatar relay the content of the textual content of the input data stream 200 in the virtual environment (step 1500).

Suitably, this process occurs in real time such that the avatar embodying the user relays the content of the input data stream and portrays an appropriate emotional gesticulation based on the control point of the affective state module 130 that is synchronized with the content of the input data stream 200.

FIG 3 shows the step 1200 of extracting emotional states from an input data stream 210 in more detail. The emotion recognition system 100 first determines the type of input data stream 210 that is being received (step 1210).

If the input data stream 210 is an audio input data stream 210A, then the voice analysis module 110 of the emotion recognition system 100 extracts the emotional intent in this input data stream 210A (step 1220).

If the input data stream 210 is a text input data stream 210B, then the text analysis module 120 of the emotion recognition system 100 extracts the emotional intent in this input data stream 210B (step 1230).

If the input data stream 210 is an affective space state control data stream 210C, then the affective space state module 130 processes this type of data stream 210C (step 1240). FIG 4 shows the process involved in the step (step 1220) of extracting the emotional intent of the audio input data stream 210A by the voice analysis module 110.

The audio input data 210A is spliced into frames of discreet time intervals by the voice analysis module 110 and are captured (step 1221). Preferably, the audio input data 210A is spliced into frames of 10 milliseconds in length and each frame is captured using a known digital signal processing application, such as FMOD API (www.fmod.org). which forms part of the voice analysis module 110.

The voice analysis module 110 then extracts the audio characteristics of each frame (step 1222). Suitably, these audio characteristics are the energy, pitch, down select and second pitch of the audio input data 210A.

These audio characteristics are then averaged for each frame that forms part of an utterance of the user (step 1223). An utterance may be a word or other similar vocalized signal. The audio characteristics are averaged using the following formula:

Mean Energy

$E_{meαn}$ =

$$\frac{1}{f}\times\sum_{i=1}^{f} energy_i$$

where f is the number of frames and energy is the amplitude of the waveform in frame /.

Mean Pitch

pitch

$$P_{mean} = \frac{1}{f}\times\sum_{i=1}^{f}.$$

where f is the number of frames and pitch is the frequency of the waveform in frame i. Mean Downselect Pitch

$$P_{mean\_down} = Max\left\{pitch_i \mid \forall\, pitch_i, pitch_i < P_{mean}\right\}$$

Second Mean Pitch f

J P- second mean $_{=}$ -$_{s}$ o *Σ $_{/=}$j pitck pitch≤P_{mean}

where s is the number of frames where the pitch is below P, mean

These averaged audio characteristics for each utterance are then input into a neural network of the voice analysis module 110 (step 1224). Suitably, this neural network is pre-trained using frames of prerecorded emotional voice data.

The neural network of the voice analysis module 110 then outputs an emotional state vector representing the user's current emotional state (step 1225). As discussed above, this emotional state vector is a six dimensional vector for each for the six main emotions. Alternatively, other Al techniques could be used to produce the emotional state vector of the user such as genetic algorithms, K-Nearest Neighbour calculations or Bayesian Inferencing. The process then continues as per step 1300 which is discussed in greater detail below.

FIG 5 shows the process involved in the step (step 1230) of extracting the emotional intent of the text input data stream 210B by the text analysis module 120. The text analysis module 120 parses the text input data stream 210B for emoticons (step 1231). When an emoticon is detected, the emotional intent of the text is communicated to the affective space state module 130 for processing as per step 1300 discussed in greater detail below. For example, the following emoticons have the designated emotional intent:

:) happy

:(sad

>:| angry

A suitable emotional state vector is constructed based on the emotional intent determined from the parsed text. Furthermore, the text analysis module 120 parses the text input data stream 210B to search for action keywords from the user for the purposes of avatar animation (step 1232). This data is extracted and communicated directly to the avatar control module 500 as will be discussed further below.

As previously discussed, the affective state control data stream 210C provides for two way communication between the user and the affective space state module 130. On the one hand, the affective state control data stream 210C provides a visual representation to the user of the current control point (that is the value of the current emotional state vector) that the emotion recognition system 100 has inferred as the users emotional state. On the other hand, it allows the user to provide affective state control commands to alter the current control point calculated by the affective space state module 130 and thus the emotional state vector representing the user's current emotional state.

This allows a user to train the neural network forming part of the voice analysis module 110 in order that the emotion recognition system 100 adapts and learns the users emotions based on the input data streams. FIG 6 shows a representation of an interface 2000 for communicating and displaying the affective space state of the user. The interface provides an emotional compass whereby the distance the control point is away from each pure emotion (i.e. a value of 1 on the scale of 0 to 1 for each of the values in the six dimensional emotional state vector) is represented by the relative size of the face 2100 for each of the six emotions. Furthermore, the interface 2000 provides an emotional compass 2200 that points to the user's most dominant current emotional state.

The user is able to directly control the location of the control point in the affective space state module by using a mouse 201 C, a keyboard 202C or the like to change the size of each of the faces 2100 representing an emotion to thereby change the intensity of the value of that particular emotion in their control point. These changes are communicated to the affective space state module 130 via the affective space state control data stream 210C.

FIG 7 shows the process of updating the control point in the affective space state module 130 (step 1300 in FIG 2) in more detail.

The affective space state module 130 combines the emotional state vector calculated by either the voice analysis module 110 or the text analysis module 120 with the current emotional control point of the affective space state module 130 (step 1310). The calculated emotional state vector is combined with current emotional control point vector based upon weighted averages. This allows the control point to move between discreet emotional states more naturally thus minimizing erratic unnatural emotional state changes. This updating of the control point is carried utilizing the following formula:

$$CP_t = \frac{W_0 CP_{t-1} + W_1 R_t}{2}$$

where $0 < W_0 \leq 1$

and O ≤ rø^ ≤ l

and $\psi_{0\,+}\ \psi_X = l$

and CP is the control point value at time t, and R is the resulting neural network vector.

W_0 and Wi are the weightings assigned to the previous control point and the calculated emotional state vector respectively.

These weightings are used to assign a bias during the calculation to control the emotional transition sensitivity. For example, if the calculated emotional state vector is considered a dramatic change to a current state of happiness and the result is sadness then the weightings may be set to $W_0 = 0.2$ and Wi = 0.8 so that the calculated emotional state vector has more influence over the calculation of the new control point than the old value. The dominant emotion of the new affective control point is then calculated

(step 1320). The Euclidian distances between the new control point and all discrete emotional points in the affective space are calculated. The closest discrete emotional point is thus determined as the dominant emotion for the user. The distance between the new control point and the dominant emotion point measures the intensity of that emotion. Furthermore, as the affective space is a continuum the distance between the control point and the other discrete emotions can be used to create an emotional blend. For example, a control point with the value [0.2, 0.004, 0.01. 0.99, 0.00, 0.00] would be recognized as the emotion fear as it is closest to that binary representation.

This data is used to update the emotional compass 2200 that points to the user's most dominant current emotional state. Furthermore, the intensity of all the non-dominant emotions for the new affective control point are calculated (step 1330). The intensity of the other emotional states are measured according to the difference between their associated vector value and the value 1 , thus:

L= m where / is the intensity of happiness, h, represented by the first value in the neural network result vector, R.

In order to use these values in the interface 2000 to portray to the user the calculated emotional state, the intensities are normalized with respect to their sum, thus:

r = ∑Λ[i]

(=1 and

NL = W)÷т

where T is the sum of the intensities and NI is the normalized intensity for happiness. Normalization is calculated for all intensities in the same manner.

FIG 8 shows the process involved in the step of generating avatar control commands (step 1400 shown in FIG 2 in more detail. In the event that the input data stream is an audio input data stream 210A, the content of this input data stream 210A is communicated to the acoustic phoneme module 300 for processing (step 1410A).

The Acoustic Phoneme Analysis Module 300 takes the user's original speech input and analyses it for phonetic information. This information is passed to the Avatar Controller 300 to synchronise the avatars mouth movements with the speech to give the illusion that the speech is actually coming from the avatar.

The Avatar Controller 300 uses the phonetic information to select from mouthshape phoneme maps database 710 in the animations database 700 the most appropriate stream of mouth movements to match the speech. The result is two streams of information being used to give the avatar the power of speech; the user's voice and a synchronised mouthshape animation the same length as the speech. The acoustic phoneme module 300 may be in the form of any known commercially available module or may be proprietary software. In the event that the input data stream is a text input data stream 210B, the content of this input data stream 210B is communicated to the textual phoneme module 400 for processing (step 1410B).

The textual phoneme module 400 accepts sentences and paragraphs of text in the text input data stream 210B from the user. This text is the chat message typed into the system. The text is analysed for its phonetic properties and mouthshape instructions are embedded in the text for processing by the

Avatar Controller 300.

Suitably, this module may be any known module in the art or be in the form of proprietary software. These engines in the art determine the statistically- best time alignment of a text script with a speech file. They determine information such as time stamps for phonemes, words, sentences, and user markers.

The time stamps can be embedded in the text, like tags in HTML, and interpreted by the Avatar Controller 300 to control the mouthshape animations. The avatar controller 300 then selects a gesture from the gesture database 720 of the animation database based upon the current affective control point/vector calculated previously (step 1420).

The avatar controller 300 then selects a facial expression from the facial expression database 730 of the animation database 700 based upon the current affective control point/vector calculated previously (step 1430).

Hence based on the content of the input data stream and the current affective control point, the avatar controller gesticulates the avatar as the avatar is communicating in real time whereby the emotional attitude of the avatar is synchronised with the content being communicated (step 1500). FIG 9 shows a table of example visemes for the mouthshape phoneme database

710. The mouthshape phoneme database 710 specifies the head models for mapping recognized phonemes to visemes. Each model specifies a series of bone structures that allow inverse kinematic techniques to smoothly calculate facial movements between each viseme. This way the animated talking head does not jump from viseme frame to viseme frame but moves along a continuum from one to the other.

FIG 10 shows a table of example gestures for the gesture database 720. For each gesture, there are a set of emotional accent modifications that can be applied. The table provides an example for the gesture of STANDING with a variety of SAD intensities applied. Emotional accents can be added to either the upper (torso & head) or lower (pelvis & legs) of the avatar body.

The intensity value indicates how much the accent modification applies to

the original gesture. The system and method of the invention uses the key frame sequences for intensities 0 and 1 to obtain the start and end points of the

accents and use inverse kinematics to determine all other accent postures in between. In this way there are theoretically an infinite number of poses. This method also allows for blending emotional states by using the 0 and 1 key frames for the accents of all pure emotions and using inverse kinematics to

determine the blended pose.

For example, if the avatar is currently standing and they are required to look sad at an intensity of 0.5. The system examines key frames 0 and 1 of the sad accent, calculates the kinematics to move the pose from 0 to 1 and stops

when it is half way there. FIG 1 1 shows a table of example facial expressions selected from the facial expression database 730. The facial expression database 730 contains facial expression models representing each of the six universal emotions. Each

model specifies a series of bone structures that allow inverse kinematic

techniques to smoothly calculate facial movements between each expression as well as integrate them with mouth movements.

Throughout the specification the aim has been to describe the invention without limiting the invention to any one embodiment or specific collection of

features. Persons skilled in the relevant art may realize variations from the specific embodiments that will nonetheless fall within the scope of the invention. It will be appreciated that various other changes and modifications may be made to the embodiment described without departing from the spirit and scope of the invention.

PATENT CITATIONS

Cited Patent	Filing date	Publication date	Applicant	Title
US20020194002 *	Jul 12, 2002	Dec 19, 2002	Accenture Llp	Detecting emotions using voice signal analysis
US20030137515 *	Sep 6, 2002	Jul 24, 2003	3Dme Inc.	Apparatus and method for efficient animation of believable speaking 3D characters in real time

* Cited by examiner

NON-PATENT CITATIONS

		Reference
1	*	CHUNLING MA, HELMUT PREDINGER, MITSURU ISHIZUKA: 'A

		Chat System Based on Emotion Estimation from Text and Embodied Conversational Messengers' ICEC 2005, pages 535 - 538
2	*	CHUNLING MA, HELMUT PREDINGER, MITSURU ISHIZUKA: 'Emotion Estimation and Reasoning Based on Affective Textual Interaction' ASCII 2005, pages 622 - 628
3	*	FENG YU ET AL.: 'Emotion Detection from Speech to Enrich Multimedia Content' ADVANCES IN MULTIMEDIA INFORMATION PROCESSING - PCM 2001: SECOND IEEE PACIFIC RIM CONFERENCE ON MULTIMEDIA BEJING, CHINA vol. 2195/2001, 24 October 2001 - 26 October 2001,
4	*	KWON O.-W. ET AL.: 'Emotion Recognition by Speech Signals' EUROSPEECH 2003, pages 125 - 128
5	*	TATO R. ET AL.: 'Emotional Space Improves Emotion Recognition' ICSLP 2002, pages 2029 - 2032
6	*	WONSEOK CHAE, YEJIN KIM, SUNG YONG SHIN: 'An Example-based Approach to Text-driven Speech Animation with Emotional Expressions' EUROGRAPHICS vol. 22, no. 3, 2003,

* Cited by examiner

REFERENCED BY

Citing Patent	Filing date	Publication date	Applicant	Title
US8149241	Dec 10, 2007	Apr 3, 2012	International Business Machines Corporation	Arrangements for controlling activities of an avatar
US8228170	Jan 10, 2008	Jul 24, 2012	International Business Machines Corporation	Using sensors to identify objects placed on a surface
US8379968	Dec 10, 2007	Feb 19, 2013	International Business Machines Corporation	Conversion of two dimensional image data into three dimensional spatial data for use in a virtual universe
US8386918	Dec 6, 2007	Feb 26, 2013	International Business Machines Corporation	Rendering of real world objects and interactions into a virtual universe

CLASSIFICATIONS

International Classification	G10L17/00, G10L21/00, G06F17/20, G06T15/70, G06T13/00

Cooperative Classification	G10L17/26, G06N3/004
European Classification	G10L17/26, G06N3/00L

LEGAL EVENTS

Date	Code	Event	Description
Apr 1, 2009	122		Ref document number: 07718539 Country of ref document: EP Kind code of ref document: A1
Sep 4, 2008	NENP		Ref country code: DE
Nov 7, 2007	121		