Multimodal Emotion Estimation and Emotional Synthesize for Interaction Virtual Agent

Minghao Yang, Jianhua Tao Tan, Hao Li, Kaihui Mu,

The National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences
mhyang@nlpr.ia.ac.cn, jhtao@nlpr.ia.ac.cn, hli@nlpr.ia.ac.cn, khm@ nlpr.ia.ac.cn

Abstract: In this study, we create a 3D interactive virtual character based on multi-modal emotional recognition and rule based emotional synthesize techniques. This agent estimates users' emotional state by combining the information from the audio and facial expression with CART and boosting. For the output module of the agent, the voice is generated by TTS (Text-to-Speech)system by freely given text. The synchronous visual information of agent, including facial expression, head motion, gesture and body animation, are generated by multi-modal mapping from motion capture database. A kind of high level behavior markerup language(hBML) which contains five keywords is used to drive the animation of virtual agent for emotional expression. Experiments show that this virtual character is considered natural and realistic in multimodal interaction environments.

Keywords: interactive virtual character, multi-modal, face animation, body movements, CART, boosting

1 Introduction

Virtual agents, as an exciting application in virtual reality and artificial intelligence, have been studied intensively for the last ten years. They are often utilized as interactive agents communicating with human to make the users understand the scene and designers' purposes. In some situations, they are used to make the virtual scene seem more real, or to create the core of the scene, the context, the background and to evolution of the stories in amusements[1], games[2], education[3], interactive virtual agents[4].

Recently, with the improvement of speech recognition and natural language process, topic-centric human-computer interactions have achieved great progress and widely applied in call center services, language learning and hotel booking [5-7], etc. In our daily life, we communicate with each other through audio, expression, gesture and body motion. All of these channels are closely interconnected and give a vivid interactive expression in case of a high degree of synchronism[8]. In human-computer interaction, there is also an urgent need for interactive agents which can communicate with users through all of above communicative channels to express their meanings, aims, mood and personalities[9]. Now it's still a big challenge to make virtual agents conversation or response like a real person in human-computer interactions, because of human's great sensitivity over the subtleties of emotional communication, especially in face to face human-computer conversation applications.

In this study, we aim to create a 3D interactive virtual character based on multi-modal emotional estimation and rule based emotional synthesize techniques. The remainder parts of this paper are organized as followings: the related works for multimodal emotion estimation and behavior controller technique are introduced in section 2; in section 3 and section 4, we will present the idea of emotion estimation and behavior markeup language in details respectively; and the experiments and conclusion are given in section 5 and section 6.

2 Related Works

2.1 Multimodal Emotion Recognition

Emotion recognition has been one of the most important issues in human computer interaction. The importance of automatic emotion recognition has been emphasized by many researchers [10-12]during the last two decades. Basically, audio and visual channels (mostly facial expressions) are the two that transmitting human emotions mostly. Extensive work has been done in the past to recognize emotions from only single channel [13] [14]. While in nature human interaction, the most two obvious two emotional channels are acoustic and expression[12]. Discussion on these two channels be great help to improve the emotional estimation in human-computer interactions.

Up to now, there are two kinds of approaches to fuse the bimodal information: the decision-level fusion and feature-level fusion methods. The former combines the results from acoustic classifier and visual classifier by rules [15], while the latter classifies the bimodal feature vectors combined from audio and visual channels into different emotions directly [12]. The decision-level fusion takes advantage of the assumption that audio and visual channels play different roles in human perception of different emotions. In [15], the authors determined the weighting matrix for audio and visual channels by subjective perception. However, whether the weights can be applied to other applications is in doubt. In [12], the authors used SVM model to obtain better performance in classifying emotions. However, they did not consider the different dominances of features.

2.2 Multimodal Agent Animation

For face expression, virtual agents can animate through embedding markup language in text to control the non-speech expression, such as the Avatar Markup Language (AML) [17] and Affective Presentation Markup Language (APML) [18]. Some systems are dominated by user mainly through audio or text input. BEAT[19] parses natural language and tags it with grammatical

information and dialog structure, which will be used for nonverbal behavior through a number of rules.

Face is the most expressive area of the body. Many of the face animation systems are based on the general recognized face expression systems, such as Facial Action Coding System (FACS)[20] and Emotional Wheel[21]. The emotional space in Emotional Wheel is represented by a disk defined by two dimensions: activation and evaluation. Similarity between two emotions is proportional to the angle that separates their positions on the wheel. In Greta[4], the type of emotions and expressions are based on a representation of beliefs and goals. The internal states of the agent are generated by Dynamic Belief Networks. Some systems [22] also use the dynamic expressions to create emotional models, in which the dynamics of emotions and moods are simulated over time.

In this study, we create a 3D interactive virtual character based on multi-modal emotional estimation and rule based emotional synthesize techniques. This agent estimates users' emotional state by combining the information from the audio and facial expression with CART and boosting method. For the output module of the agent, the agent's voice is generated by TTS (Text-to-Speech)system by freely given text information. The synchronous visual information for agent, including facial expression, head motion, gesture and body animation, are generated by multi-modal mapping from motion capture system. Experiments show that this virtual character is considered natural and realistic in multimodal interaction environments.

3 Bimodal Emotion Recognition with CART and Boosting

The input information of our emotion estimation module includes visual and audio features. The boosting-based framework with CART method is used to fuse the features from two input modules. Fig. 1 presents the total framework of our bimodal emotion recognition. In this section, we first introduce how the visual and audio features are selected and then, how CART and Boosting is used to emotion recognition will be given.

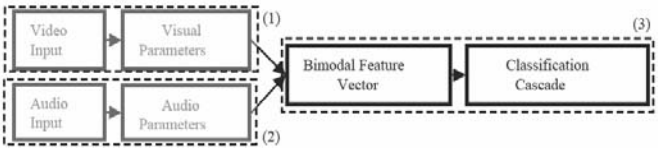

Figure 1 Emotion recognition with CART and Boosting

3.1 Visual Features

Visual module is to track facial features and calculate visual parameters. Point Distribution Models (PDM) is introduced here as a kind of high-level constraint in this study. We selected a small set of sample images from the large corpus as the training set, and marked 20 feature points (as shown in Fig. 2) in each image by hand. The training set is used to train the PDM.

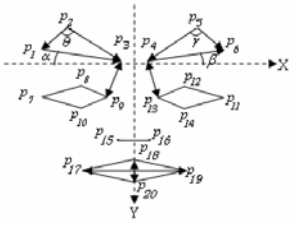

Figure 2 PDM model

The facial shape in each sample image is formulated as a vector by concatenating the coordinates of feature points, $X_i(x_i^1, y_i^1, x_i^2, y_i^2, \ldots x_i^N, y_i^N)$, where x_i^j and y_i^j are the two coordinates of j-th feature point in i-th image. N indicates the number of feature points. Here, N=20. In order to compare equivalent points from different shapes, an alignment is necessary. A modification of the Procrustes method described in[23] is used here to align all facial shapes before training the PDM. Then, we get aligned facial shapes. With the PDM definition as figure 2, we get visual features as formula (1)-formula (4)

$$\phi_1 = \frac{1}{2}(\theta + \gamma) \quad (1)$$

$$\phi_2 = \frac{1}{2}(\alpha + \beta) \quad (2)$$

$$d_1 = \frac{1}{2}(\overline{p_3 p_9} + \overline{p_4 p_{13}}) \quad (3)$$

$$d_2 = \overline{p_{17} p_{19}} \quad (4)$$

where θ, γ, α and β are angles defined in Fig.3. We also define two directions X and Y, which are parallel with the vectors $\overline{p_3 p_4}$ and $\overline{p_{18} p_{20}}$ respectively. $|\overline{p_3 p_9}|$, $|\overline{p_4 p_{13}}|$ and $|\overline{p_{17} p_{19}}|$ are length of vectors $\overline{p_3 p_9}$, $\overline{p_4 p_{13}}$ and $\overline{p_{17} p_{19}}$. Since the lip movements are synchronized with the speech content, we only choose one feature from lower face to reduce the influence of speech content.

3.2 Audio Features

For the audio features, some research has confirmed the following features to be useful for the emotional speech classification: utterance duration, F0 range, the maximum of F0s, the minimum of F0s, the mean of F0s, the mean of energy, and the mean of durations. To simulate the stress, we added some more parameters in our previous work [24]: the position of the maximum F0, the position of the minimum F0, the position of duration peak, the position of the minimum duration. Parameters describing laryngeal characteristics on voice quality were also taken into account. In [24], we analyzed the importance of all these audio features for emotional speech classification. The results showed that the mean of F0s, the maximum of F0s, F0 range, the mean of energy, the mean of durations and the position of the minimum F0 have much better "resolving power" for emotion perception than other parameters. They are selected as the audio feature vector for audio input module in this paper.

3.3 Bimodal Emotion Recognition

With the visual and audio features and the corresponding emotion labels, we get our bimodal emotion training set S. Let $S=\{(xi,yi)\}_{i=1}^{\Pi}$, where x_i the combined feature vector, y_i is the emotion label from six emotion classes $Y=\{0,1,2,3,4,5\}$, and Π is the total number of training samples. Fig.3 shows the configuration of the strong classifier. Each box represents a "weak" classifier, which is a CART model in our method. The strong classifier is a linear combination of each CART model [25]. T is the number of iterations.

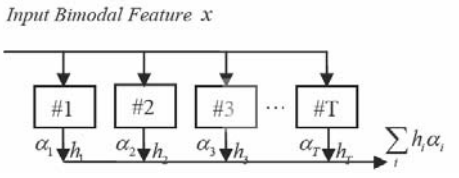

Figure 3. Bimodal emotion recognition

The scheme of our CART and boosting algorithm is showed as follows:

Step 1. Given M training samples $(x_1,y_1)\ldots(x_m, y_m)$, where $y_i \in Y$;

Step 2. Initialize

$$D_1(i,\ell_0,\ell_1)=\begin{cases}1/(M\cdot|y_i|\cdot|Y-y_i|), \text{ if } \ell_0\neq y_i \quad and \quad \ell_1=y_i;\\ 0, \qquad \text{else.}\end{cases}$$

Step 3. For t=1,2,….T, Train the weak classifier using distribution for D_t; Get the weak hypothesis h; Choose weight α_t and update

$$D_{t+1}(i,\ell_0,\ell_1)=\frac{D_t(i,\ell_0,\ell_1)\exp(\alpha_t(h_t(x_i,\ell_0)-h_t(x_i,\ell_1))/2)}{Z_t}$$

where Z_t is a normalization factor.

Step 4 Output the final hypothesis:

$$f(x,k)=\sum_{t=1}^{T}\alpha_t h_t(x,k)$$

where each instance $x \in X$ may belong to multiple classes in Y, is the number of classes x belongs to, and $|Y-y_i|$ is the number of the remaining classes.

In step 3, the samples assigned with larger weights will duplicate themselves in the current training set With the resample training set, we can get a new CART model and choose the $\alpha_t = 0.5*\ln((1+r_t)/(1-r_t))$, where r_t is a pseudo-loss of for this multi-class self-rated problem, obtained by

$$r_t=\sum_{t,\ell_0,\ell_1} D_t(i,\ell_0,\ell_1)[h_t(x_i,\ell_1)-h_t(x_i,\ell_0)]$$

4 Agent Animation

4.1 Audio driven emotional speech synthesis

Various kinds of methods have been proposed for lip-synch, including linear prediction analysis[26], artificial neural networks (ANNs) [27], hidden Markov models (HMMs) [28, 29], and head motion, including parameter-driven (rule-based) [30], data-driven head animation synthesis [31]. Based on the collaborative filtering[32] and k-nearest neighbors(kNN), we developed a lip-synch system[33]. This technique renders lip movements and makes it synchronized with the acoustic signals generated from Text-to-Speech(TTS).

On the point of head motion, speech is usually accompanied by head movements, which could be effectively described by visual prosody[31]. Like what introduced in [31], first, we cluster head movement patterns that show up in each of seven different emotional states performed by one actress, and investigate how they relate with text's features. In this method, patterns of head and facial movements are strongly correlated with the prosodic structure of the text. As lip-synch and head-motion are not key techniques for emotion expressions in our system, please refer to [33] for the relative details.

4.2 Body Movement

The Motion Analysis system is utilized to capture the movements of face and body movements. The typical animations and marker sets placed on the actors are shown in Fig. 4(a). Motion Builder receives the captured animating data from the Motion Analysis system and a skeletal based 3D character generated in 3DMax. Then the character model and animations are transferred into the Cal3d engine through 3DMax. Fig 4(b) lists some frames for happiness and anger actions on strong and weak states. The actions in a column belong to same emotional state, and they are different on speed and magnitude. Finally, we obtain 42 action units for all emotions.

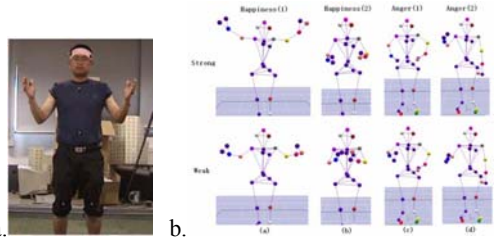

Figure 4 Emotional body movements

4.3 Agent Controller

The virtual character needs the control information coming from the Control Module to synthesize audio and visual speech, face expressions and body gestures. The form of the manipulation information is like this: <animation=* emotion=* pitchRatio=* speed=* volume=* > text. The "animation" field gives the index number of body gestures. The "emotion" field transfers the index number of face expressions, while "pitchRatio", "speed" and "volume" are used to adjust the synthesized audio speech. The whole relationship and how the multimodal channels are used to agent' animation synthesize is shown in Fig. 5.

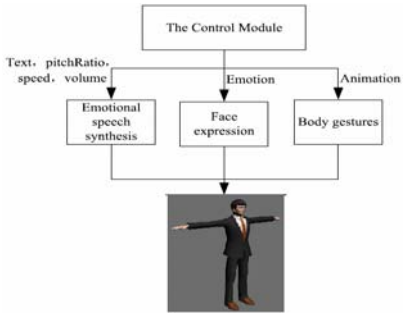

Figure 5 Multimodal animation synthesize

5 Experiments

5.1 Emotion Recognition

We evaluate the performance of emotion recognition methods by confusion matrix. In our work, the confusion matrix is a six by six matrix. The element in i-th row and j-th column in it represents the percentage of samples whose real emotion state is i while the estimated emotion state is j. The larger the elements in the diagonal, the better performance a recognition method will give. Results are shown in table 1.

	Neutral	Happiness	Sadness	Anger	Fear	Surprise
Neutral	*70%*	12%	4%	0%	8%	6%
Happiness	10%	*78%*	2%	2%	0%	8%
Sadness	22%	6%	*40%*	0%	30%	2%
Anger	0%	0%	0%	*94%*	0%	6%
Fear	20%	8%	32%	0%	*40%*	0%
Surprise	0%	4%	2%	32%	0%	*62%*

Confusion matrix by using only acoustic parameters

	Neutral	Happiness	Sadness	Anger	Fear	Surprise
Neutral	*72%*	0%	26%	0%	2%	0%
Happiness	40%	*40%*	0%	20%	0%	0%
Sadness	2%	18%	*80%*	0%	0%	0%
Anger	0%	0%	10%	*78%*	0%	12%
Fear	18%	0%	0%	0%	*50%*	32%
Surprise	10%	0%	2%	0%	28%	*60%*

Confusion matrix by using only visual parameters

	Neutral	Happiness	Sadness	Anger	Fear	Surprise
Neutral	*88%*	6%	6%	0%	0%	0%
Happiness	14%	*84%*	2%	0%	0%	0%
Sadness	0%	2%	*85%*	0%	13%	0%
Anger	0%	2%	0%	*96%*	0%	2%
Fear	0%	0%	5%	0%	*95%*	0%
Surprise	0%	1%	0%	6%	0%	*93%*

Confusion matrix of our boosting-based bimodal emotion recognition

Table 1 Bimodal emotion recognition confusion matrices

From the confusion matrices in table 1, it is clear that "sadness" and "fear" is easy to be confused with only audio features. Similar situations also happen to "surprise" and "anger". These confusions are caused by similarity in expressing emotions through only one channel, which matches the observation in works [15] [10] [16]. It is also clear that the confused emotion pairs caused by the audio channel are different from those caused by the visual channel.

Comparison between matrix 1,2 and matrix 3 shows that our boosting-based algorithm performs better than the rule-based algorithm. Intuitively, different persons may have different styles to express emotions, and different emotions may be expressed in different ways. The boosting framework in our method can capture the different dominance of every feature in bimodal feature vector statistically, not only the dominance of audio or visual channels, which is more precise than the previous rule-based methods.

5.2 Agent Animation

We build an emotional 3D talking agent with free Cal3D platform[88] on PC with 2.6G CPU and 2G RAM. We drive head and lip movement in real-time depending on visual prosody of input text by TTS system. Figure 6 presents several selected frames for our agent when he is speaking "it's sunny tomorrow" with neural emotion state.

Figure 6 Several selected frames for agent speaking "it's sunny tomorrow" with neural emotion state.

Subjective evaluation on emotion animation of 3D taking agent are given by Mean Opinion Score (MOS) scores, which is adopted for different emotion conversation on: 1) pure speech based conversation without any emotion recognition for users and without emotion output for agent; 2) Emotion based speech conversation with bimodal emotion recognition for users facial expressions and audio input. However, the virtual agent only presents random emotion output on body and head animation; 3) HCI conversation with our schema on bimodal emotion recognition and emotional controller for virtual agent.

For each user, the subjects were asked to score the expressivity of HCI conversation system on a five level mean opinion score (MOS) scale. The results show in figure 10. The average MOS scores of three sessions are 2.2 (Session (1)), 2.8 (session (2)), 3.8 (session (3)). The pure speech based HCI conversation without emotion analysis and emotion output is always considered unnatural. The speech based HCI conversation with random animation output gets a higher score than that of previous one. When the features synthesized with the multimodal emotion dialog management and emotional behavior control approach proposed in this paper is added, the HCI dialog shows a nature and realistic conversation procedure and improves the MOS by 1.0 points than that with random animation on body.

6 Conclusions and Future Works

In this paper, a virtual agent is designed to interact with users with appropriate expressions including speech, virtual speech, face expressions, head movements and body gestures. Though natural emotion based 3D interactive agent is achieved in our work, there are some improvements needed to be done. The first one is the emotional state of each sentence in hBML needs to be assigned by user, a work of predicting the emotional state from the output context should be added. Secondly, all the body and face action units in database are recorded in

advance, which could not modified by users at run-time. It needs to generate more action units from the fixed action in database by motion morph techniques in the future.

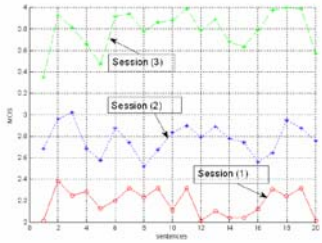

Figure 7 Subjective evaluations on emotion animation of 3D taking agent

References

[1] J. Lasseter, "Principles of traditional animation applied to 3d computer animation," ACM SIGGRAPH Computer Graphics, vol. 21, no. 4, pp. 35-44, 1987.

[2] I. Machado, A. Paiva, and R. Prada, "Is the wolf angry or... just hungry?," pp. 370-376, 2001.

[3] J. Gratch and S. Marsella, "Lessons from emotion psychology for the design of lifelike characters," Applied Artificial Intelligence, vol. 19, no. 3, pp. 215-233, 2005.

[4] F. Rosis, C. Pelachaud, I. Poggi, V et al., "From greta's mind to her face: modelling the dynamics of affective states in a conversational embodied agent," International Journal of Human-Computer Studies, vol. 59, no. 1-2, pp. 81-118, 2003.

[5] Wik P, Hjalmarsson A. Embodied conversational agents in computer assisted language learning. Speech communication, 51(10), 1024-1037, 2009.

[6] Brustoloni, Jose C. (1991), "Autonomous Agents: Characterization and Requirements," Carnegie Mellon Technical Report CMU-CS-91-204, Pittsburgh: Carnegie Mellon University

[7] Engwall, O., Balter, O. Pronunciation feedback from real and virtual language teachers. Journal of Computer Assisted Language Learning, 20(3),(2007), 235-262.

[8] S. Planalp, "Varieties of cues to emotion in naturally occurring situations," Cognition & Emotion, vol. 10, no. 2, pp. 137-154, 1996.

[9] V. Vinayagamoorthy, M. Gillies, A. Steed, E. Tanguy, X. Pan, C. Loscos, and M. Slater, "Building expression into virtual characters," 2006.

[10] L.S. Chen, T.S. Huang, et al., "Multimodal human emotion/expression recognition," in Proc. the Third IEEE International Conference on Automatic Face and Gesture Recognition, 1998, pp.366-371.

[11] R. Cowie, and E. Douglas-Cowie, et al., "Emotion recognition in human-computer interaction," IEEE Signal Processing Magazine, Jan 2001, pp.33-80.

[12] C.Y. Chen, Y.K. Huang, and P. Cook, "Visual/Acoustic emotion recognition," in Proc. International Conference on Multimedia and Expo, 2005, pp.1468-1471.

[13] B. Schuller, G. Rigoll, and M. Lang, "Hidden Markov Model based speech emotion recognition," in Proc. ICASSP, 2003, pp.1-4.

[14] Y. Tian, T. Kanade, and J.F. Cohn, "Recognizing action units for facial expression analysis," IEEE Tans. Pattern Analysis and Machine Intelligence, VOL.23, No.2, Feb. 2001, pp.97-115.

[15] D. Silva, T. Miyasato, and R. Nakatsu, "Facial emotion recognition using multi-modal information," in Proc. International Conference on Information and Communications Security, 1997, pp.397-401.

[16] C. Busso, Z. Deng, and S. Yildirim, et al., "Analysis of emotion recognition using facial Expressions, Speech and Multimodal Information," in Proc. the 6th Intl. Conf. on Multimedia Interfaces, 2004, pp.205-211.

[17] S. Kshirsagar, N. Magnenat-Thalmann, A. Guye-Vuillme, D. Thalmann, K. Kamyab, and E. Mamdani, Avatar markup language,?pp. 169?77, 2002.

[18] N. DeCarolis, V. Carofiglio, and C. Pelachaud, From discourse plans to believable behavior generation.

[19] J. Cassell, H.H. Vilhjlmsson, and T. Bickmore, Beat: the behavior expression animation toolkit,?pp. 477-486, 2001.

[20] P. Ekman, "Facial expressions of emotion: New findings, new questions," Psychological science, vol. 3, no. 1, pp. 34, 1992.

[21] R. Plutchik, "A general psychoevolutionary theory of emotion," Emotion: Theory, research, and experience, vol. 1, no. 3, pp. 3-33, 1980.

[22] S. Kshirsagar, "A multilayer personality model," pp. 107-115, 2002.

[23] T. F. Cootes, C. Taylor, D. Cooper, and J. Graham, "Active shape models - their training and their applications," Computer Vision and Image Understanding, January 1995, pp.38-59.

[24] J.H. Tao, and Y.G. Kang, "Features importance analysis for emotion speech classification," in Proc. the 1st International Conference on Affective Computing and Intelligence Interaction, 2005, pp.449-457.

[25] R.E. Schapire, and Y. Singer, "Improved boosting algorithms using confidence-rated prediction," Machine Learning, Vol 37, 1999, pp.297-336.

[26] Lewis J (1991) Automated lip-sync: Background and techniques. The Journal of Visualization and Computer Animation, 2:118-122

[27] Wen Z, Hong P and Huang T,S (2002) Real-time speech driven face animation with expressions using neural networks. IEEE Transactions on neural networks, 13:916-927

[28] Yamamoto, E., Nakamura, S. and Shikano, K. 1998. Lip movement synthesis from speech based on Hidden Markov Models. Speech Communication. 26, 1-2 (Oct. 1998). 105-115.

[29] Xin L, Tao J and Yin P (2009) Realistic visual speech synthesis based on hybrid concatenation method. IEEE Transactions on Audio, Speech, and Language Processing, 17:469-477

[30] Badler N, Steedman M, Achorn B, et al.(1994) Animated conversation: Rule-based generation of facial expression gesture and spoken intonation for multiple conversation agents, Proceedings of SIGGRAPH, pp 73-80.

[31] Graf H.P, Strom Cosatto, Huang F (2002) Visual prosody: facial movements accompanying speech. In Fifth IEEE International Conference on Automatic Face and Gesture Recognition.

[32] Oki B.M, Goldberg D, Nichols D and Terry D (1992) Using collaborative filtering to weave an information tapestry. Communications of the ACM, 35:61-70

[33] Kaihui Mu, Jianhua Tao, Jianfeng che, Minghao Yang, "Real-Time Speech-Driven Lip Synchronization" 4th International Universal Communication Symposium (IUCS2010), Oct. 2010, pp 377-381