CCAI Vol. 23, N°1-4 (2006), pp. 89-98.

AN AUTOMATED TEXT-BASED SYNTHETIC FACE WITH EMOTIONS FOR WEB LECTURES

Siska Fitrianie and Leon J.M. Rothkrantz

Abstract

Web lectures have many positive aspects, e.g. they enable learners to easily control the learning experiences). To develop high-quality online learning materials takes a lot of time and human efforts [2]. An alternative is to develop a digital teacher. We have developed a prototype of a synthetic 3D face that shows emotion associated to text-based speech in an automated way. As a first step, we studied how humans express emotions in face-to-face communication. Based on this study, we have developed a 2D affective lexicon and a set of rules that describes dependencies between linguistic contents and emotions.

Categories and Subject Descriptors

K.3.1. Computer Uses in Education, I.2.7 Natural Language Processing, I.2.1 Applications and Expert Systems, I.3.7 3D Graphics and Realism.

General Terms

Experimentation, Human Factors, Languages

Keywords

Emotion, multimodal communication, knowledge acquisition, natural language processing

Introduction

Among the many changes brought by the Internet is the emergence of electronic learning over the Web, e.g. [1][8][22]. Contrast to the traditional education flow, web-based environment students can choose their own paces for learning. Learners can easily connect to the Internet anytime and anywhere. At MMI-Group TUDELFT, there is a project running on developing web-lectures [10] (see Fig. 1). The lectures are facilitated by a real-time navigable table of content, Microsoft PowerPoint slides and recorded audios/videos of human teachers. This project aimed at developing web lectures that present lecture material in advance of class, therefore, more in-class time can be spent with authentic learning activities.

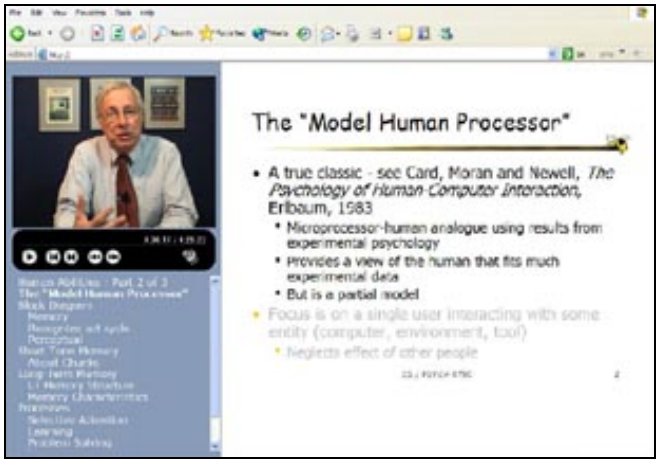


Fig. 1. Web lecture playback in a browser [10]

Developing a high quality online learning is typically an intense process [2]. A basic requirement for web lectures is the availability of excellent lectures. Teachers, who develop web lectures, need to spend a great investment of time and energy and the ongoing evaluation and modification processes require additional and ongoing investments of time. During recording they have to be present at specific time and location. The aim of our research is to create a

digital teacher that is available any time and includes the performance of excellent human teacher.

Recent studies of the effect of visual appearance of virtual agents on learners have been carried out [3][4][16]. The results indicated that visual design of agents could still affect students' self-efficacy as well as their motivation to pursue studies in a field. Students perceive such agents as being helpful, credible and entertaining. By the employment of virtual human actor(s) many researchers showed that imitating human face-to-face conversation could facilitate a robust and natural human-machine interaction [24][30]. Motivated by this, we have developed a system that allow average users to generate facial animations using a synthetic 3D face based on Facial Coding System (FACS – [13]) in a simple manner [31]. We have also built a dictionary of facial expression (FED – [19]) that stores the facial expressions that naturally occur in face to face communication.

In the area of virtual agents, research has focused on support for conversation interaction, body language, navigation issues and adding knowledge and intelligence [12]. However, the problem of producing autonomous virtual actors relies not just on physical models of human motion but also on modeling human behavior. Facial expressions change perpetually. They do not occur randomly, but rather are synchronized to one's speech or to the speech of other [9] [26]. Humans are used to convey their thought through their (conscious and unconscious) choice of words. Some words possess emotive meaning together with their descriptive meaning. The meaning of this type of words along with a sentence structure informs the interpretation of a nonverbal behavior and vice versa. Seeing faces, interpreting their expression, understanding the linguistics contents of speech are all part of our development and growth.

The project described in this paper aims at forming knowledge for a system that is able to reason about emotions automatically from natural language text and show appropriate facial expressions as its response to the emotion content of the text. Our final goal is to develop a virtual agent by animating a synthetic 3D face that can act as a human teacher in web lectures. This paper is organized as follows. In the following sections we will give related work and an overview of the system. Section 4 describes our data acquisition experiments. Further, an affective lexicon database is presented and the knowledge base construction based on the experimental results is described, in section 5 and 6 respectively. Finally, we conclude the work.

Related work

Recent years have seen a growing amount of research on the use of virtual agents in e-learning. Particularly important are animated agents, e.g. talking face and avatar, due to their ability to communicate nonverbally using gestures, gaze, and facial expressions. One such example is Steve [27], an animated agent that helps students to learn how to perform physical and procedural tasks. Steve supports nonverbal signals to regulate the flow and turn taking of conversation, but does not support other nonverbal signals that are closely tied to spoken dialogue. Another example is Rea [7] that is able to track the user's nonverbal communication. The research field of ECA [5] attempts to create agents that include emotion, personality and convention properties as humans do in face-to-face conversation.

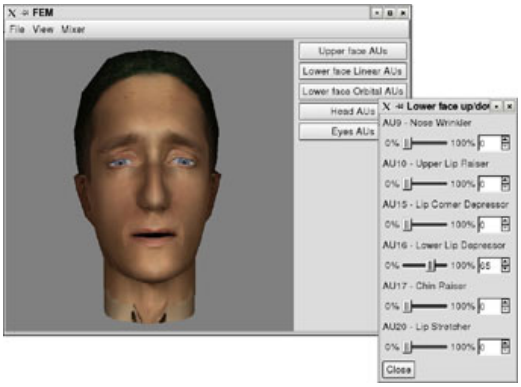


Fig. 2. A snapshot facial expression modulator [31]

Directly animating human faces with speech is challenging because there are so many parameters to be controlled for realistic facial expressions. To alleviate such difficulties for animators, the BEAT system takes annotated text to be spoken by an animated figure as input, and outputs appropriate synchronized nonverbal behaviors and synthesized speech in a form that can be sent to a number of different animation systems [6]. A rule based approach has been used to govern the animation by Wojdel and Rothkrantz [31] and connect their facial expression modulator to a dictionary of predefined facial expressions (see Fig. 2).

Most existing virtual agents, to the best of our knowledge, have not considered explicitly the emotions existing in the speech content. The

difficulty lies in the fact that emotional linguistic content consists of entities of complexity and ambiguity such as syntax, semantics and emotions. The use of simple templates has proven to be useful for the detection of subjective sentences and of words having affective semantic orientation (e.g. [18][21]). These simplistic models describe how words with an affective meaning are being used within a sentence, but fail to offer a more general approach. Furthermore, current developments of affective lexicon database (e.g. [25][29]) are based on subjective meaning of the emotion words and do not provide information about the (relative) distance between words in regards to their emotion loading context. The lack of a large-scale affective lexicon resource database makes a thorough analysis difficult. As a consequence, although important, an automated emotional expression from natural language is still rarely developed.

System overview

Fig. 3 shows the pipeline processes of the developed system. The Emotion Analysis module has a parser that associates input text to emotions and transforms it into XML format. For this purpose, this module exploits the affective lexicon database. In the next version of the system, the input will be the teacher's talk that has been synchronized with the PowerPoint presentation timetable. The Reaction Module processes the XML results by assigning appropriate facial expressions. The module has a FED in which a lexeme contains an emblem of a facial expression, a description of which AUs are activated, semantic interpretations, and an example using a synthetic face [19].

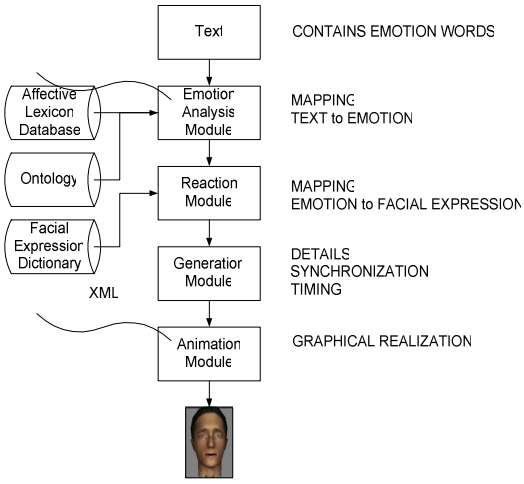


Fig. 3. System architecture

The Generation module plans the display to be synchronized with speech. It estimates speech timings and constructs an animation schedule prior to execution. The Animation module uses the results to generate facial animation based on "facial script language", where basic variables are AUs and their intensity. This module employs a parametric model for facial animation and a method for adapting it to a specific person based on performance measurements of facial movements [31].

Data acquisition experiments

The purpose of our experiments was to find out: (1) what kind of emotion expressions are shown most often in a conversation; (2) what is the typical course of particular expressions; (3) whether the expressions depends on each other; and (4) whether the expressions are related in any way to used certain words. To collect data for the analysis we have prepared ten scenarios of dialogs between two characters with diverse situations which evoke various affective states. We have asked ten participants to perform a role of one character as many expressions as possible from the scenarios. To obtain diversity of emotion expressions, provided scenarios contained a high number of punctuation marks. The facial expressions of participants during the experiment were recorded on a video recorder to be analyzed afterwards. The video sequences were converted and stored as MPEG-2 streams. They were sampled at 25 frames per second and saved with 645 KB/sec bit rate. As first step, three independent observers marked the onset and offset of an expression. In the next step, these expressions were labeled according to the context. In the final step, we also collected emotion words used in each expression. The agreement rates between the observers in both steps were about 73%.

The experimental results indicated that our participants showed most of the time a neutral face. However, we managed to capture in total 40 different facial expressions; about 20-35 different expressions per participant in each dialog, and 119 emotion words. Our experimental results were endorsed by an experiment conducted by Desmet [11], which found 41 displayed emotion expressions actually used to appraise a product (see Table 3).

To analyze distribution of the elapsed time of facial expressions, we plotted appropriate histograms for each emotion label with interval

length of 5 frames and the range from 0 – 104 frames (all expression persisting longer than 4 seconds where put into the last interval). From the histograms (see examples in Fig. 4), we noticed that emotion expressions mostly appear for a rather short period of time, somewhere between half and just more that one second (10-30 frame). However, 40% of "surprise" expression could appear until 4 second. Expressions: "anger" and "disgust" had a shape of distribution similar to the shape of distribution of "surprise", but last usually a little longer than "surprise" (15-19 frames, while most "surprise" 5-9 frames). 90% of "happiness" were distributed in the range between 10-50 frames. Its histogram also contained a tail that comprises of the longest observed expression duration. We attributed this to the dual role of the expression as both short communication signal and a long-lived mood indicator.

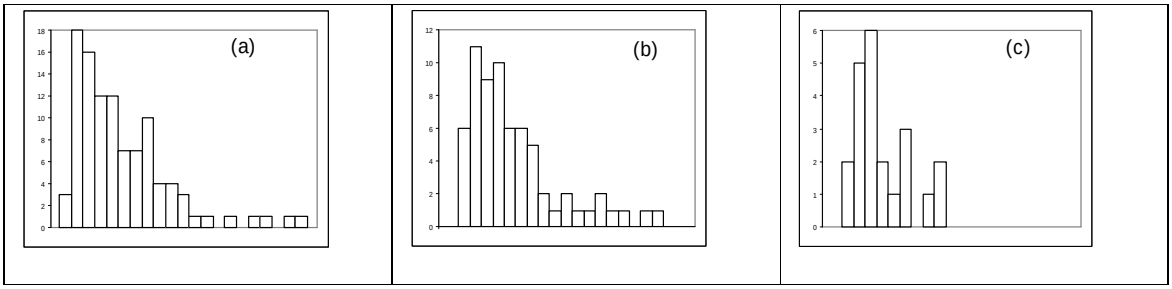


Fig. 4. Histogram for elapsed time: (a) surprised, (b) anger and (c) disgusted

To study the relationship between emotion expressions, we examined which facial expressions could occur at the same time. Generally, almost 13% of all segments covered another segment for at least one frame. We found a high number of occurrences of two kinds of combinations (see Table 1): "surprise" with "sadness" and "anger" with "disgust". Usually "surprise" preceded and overlapped "sadness" and 65% cases of "anger" had segments that contained the whole segment of "disgust". 50% of "disgust" was entirely enclosed in longer segments of "anger".

Expression	Combination #	Frames in total	Shortest segment	Longest segment
Astonishment & sadness	1	4	4	4
Surprise & grief*	15	318	2	60
Surprise & happiness	1	14	14	14
Sadness & anger	1	10	10	10
Regret & grief	1	41	41	41
Grief & anger	1	1	1	1
Anger & disgust*	17	335	7	48

*Table 1. Statistics of combined expressions (*most occurrences)*

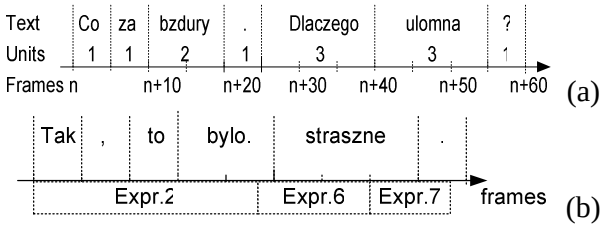


Fig. 5. An example of (a) text synchronization and (b) mapping facial expressions to text

To study the relationships between facial expressions and text, we determined the timing of occurrences of each word and punctuation mark by partitioning the dialog (manually) into basic constituents, which are usually formed by a single sentence and for each constituent into components, i.e. single words and punctuation marks (see an example in Fig. 5(a)). Then, for each component, we determined the time of its occurrence (number of frames in which the given word is pronounced). Next, we determined which components coincide with the shown expressions (see an example in Fig. 5(b)). This means, for each selected facial expression, we had to determine the text that starts with the component synchronized with the first frame of a given facial expression and ends with the component synchronized with the last frame of this expression. The experimental result showed that most of the facial expressions (around 63%) corresponded to the text spoken. For sentences with questions or exclamation marks, distinguishably, "surprise" is the most common facial expression displayed during a question. It appeared almost exclusively in short and single-word question, e.g. "really?", "sure?". We noticed that the sentences ending with exclamation mark were usually accompanied by expression "anger".

Words	Total	Linked to Expression
Non-emotion words	2206	1022 (46.3%)
Emotion words	119	65 (54.6%)

Table 2. Statistics of emotion and non-emotion words in the input text linked to facial expression

In the final experiment, we focused on mapping the shown expressions to emotional words defined by Desmet [11]. The distance of a given facial expression from a particular emotional word was defined as the number of frames with the neutral face, which appear between the facial expressions. The results showed that 54.6% of the emotion words spoken by the participants linked to facial expressions. To check whether the participants really displayed more facial expressions for emotional words than for any other words we compared results from above with the analogous statistics for non-emotional words. The comparison showed that the use of emotion words, indeed, evoked emotions which were expressed by facial expressions (see Table 2). Although, with this experiment, we still could not draw direct link between the emotion words with the facial expressions. It is because some words only occurred once or twice and some other words in different context were related to different facial expressions.

Affective lexicon database

Russell [28] and Desmet [11] analyzed qualitative values of emotion words based on humans' social value and depicted them in a 2D space of pleasureness and activation. The dimension can categorize emotions in a comprehensible way; however, the approach is not yet sufficient to differentiate between emotions, e.g. anger and fear fall close together on the circumplex. Kamps and Marx [20] proposed the quantitative differences between the relatively object notions of lexical meaning and more subjective notions of emotive meaning by exploiting WordNet [15]: (1) the smallest number of synonymy (synset) steps between two words,

e.g. MPL(good, bad) = 4 {good, sound, heavy, big, bad}, and (2) the relative distance of a word to two reference words, e.g. EVA*(proper, good, bad) = 1 and POT*(amazed, active, passive) = 0.75. The EVA* allows us to differentiate between words that are predominantly used for expressing positive emotion (close to 1), for expressing negative emotion (close to -1), or for non-affective words (value = 0).

We used the results of the experiment above and selected the stem of 140 emotion labels and words that were found in the experiment as initial records of our database. We aimed to depict them in a 2D space, by two approaches. Firstly, based on [20], the direction and distance of a vector in the bipolar dimension represent the quality (Pleasant-Unpleasant) and intensity (Active-Passive). Fig. 6 shows an example of the plotting. The degree of correctness by this approach was 78%.

Finally, we applied multidimensional scaling (MDS) to represent emotion words in 2D space. We employed [20]'s MPL to construct an NxN matrix as input. The Euclidian distances among all pairs of points were applied to measure natural distances of those points in the space. Using "similarity" (corresponding meaning) between emotion words, this procedure found the clusters that approximate the observed distance in the best way. By lowering the degree of corresponding between the Euclidian distance among points and the input matrix, the best corresponding MDS map

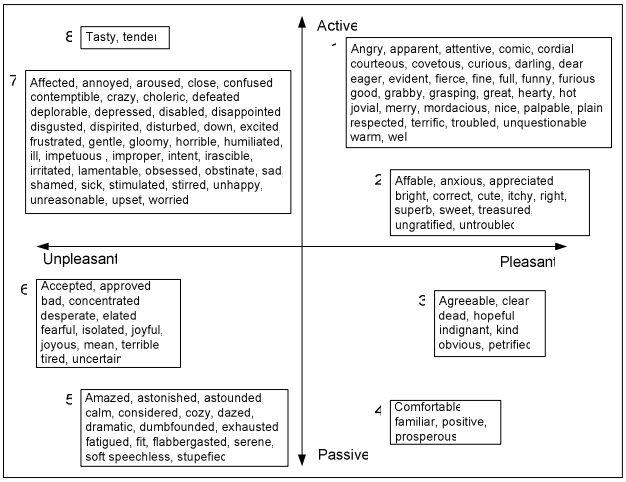


Fig. 6. A 2D space emotion words in octants-related Pleasantness-Activation

can be achieved. Fig. 7 shows an example of emotion words after MDS mapping. The degree of correctness by this approach was 65%. For both approaches, manual checking was still necessary for all mistaken classified emotion words.

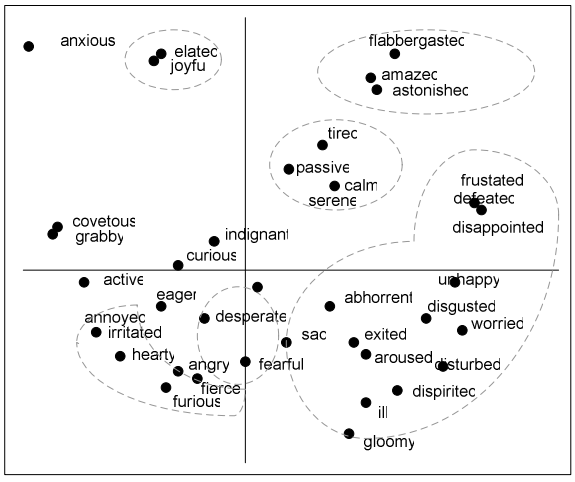
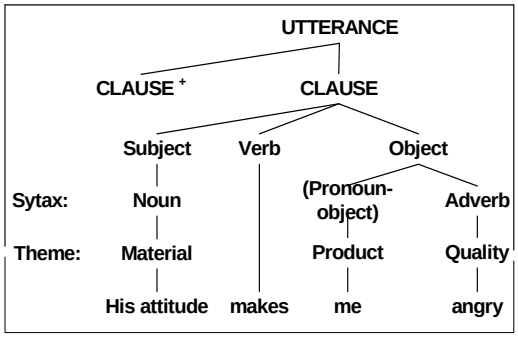


Fig. 7. MDS-Mapping emotion word

Mapping text to emotions knowledge base

Based on the experimental results, we distinguished five types of emotional intentions: (a) emotionally active toward an object, e.g. "I hate vegetable"; (b) emotionally directed by an object, e.g. "she treats me badly"; (c) emotions that provoked by an object, e.g. "his attitude makes me angry"; (d) emotions that experienced towards an object, e.g. "it is beautiful picture"; and (e) appraisal toward an object, e.g. "her mother is ill". These findings were supported by Mulder et. al [23] that stated that emotions in language as having a positive or negative orientation, an intensity and a direction toward an object or event.

We have developed heuristic rules to assess an emotion intention of a sentence and associated it with an emotion type. The rules describe the subjective notion of the sentence using three attributes: the experiencer, the attitude, and the object. The experiencer is the person in a private state of the kind attitude towards the object. By decomposing the syntax structure of the sentence and its thematic roles, we can extract these attributes including direction of the attitude toward the object (passive, active and indirect). Fig. 8 shows an example of the rules. Using the affective lexicon database, we can find the attitude's orientation (positive, negative or neutral), its intensity, and its classification and relative distance with other attitudes (or emotion words).



R

```
If exist(experiencer, product) and
   exist(object, material) and
   exist(experiencer, "main-role")
Then attitude(experiencer="main-
     role", quality,
              product, material,
     direction="passive")
```

Fig. 8. An example of constructing rule base for extracting emotion intentions in text

To control the intensity of emotion expressions from statement to another statement, we have designed six universal "emotion thermometers" based Ekman's universal emotion types [14]. Although we aimed at using 41 emotion expressions, to simplify, we associated the emotions with the universal emotion types (Table 3). If a sentence is analyzed to have an emotion loading *i*, the value of all thermometers *T* will be calculated using equation:

$$T_i(t) = T_i(t-1) + I_i.s$$
$$\forall j \neq i \,|\, T_j(t) = T_j(t-1) - d[j,i]$$

Where, *i* is the active universal emotion type, *s* is a summation factor, *I* is the emotion's activation degree, and *j* is a range of {happiness, sadness, anger, surprise, disgust, and fear}. The distance between two universal emotions *d*[*j*,*i*] follows the work of Hendrix and Ruttkay [17]. The highest value of the thermometers is considered as the sentence's current mood. Using Table 3, the emotion's intensity is the thermometer's value of the corresponding emotion type. A rule-base approach is used to govern the calculation, e.g. to detect negation, the contrast coordinate words (e.g. "but") and contrast subordinate words (e.g. "although").

Universal Emotions	Emotion Expressions
Happy	Inspired, desiring, loving, fascinated, amused, admiring, sociable, yearning, joyful, satisfied, softened.
Sad	Disappointed, contempt, jealous, dissatisfied, disturbed, flabbergasted, cynical, bored, sad, isolated, melancholic, sighing.
Surprise	Pleasantly surprise, amazed, astonished.
Disgust	Disgusted, greedy.
Anger	Irritated, indignant, hostile.
Fear	Unpleasantly surprised, frustrated, alarmed.
Neutral	Curious, avaricious, stimulated, concentrated, eager, awaiting, deferent.

Table 3. Universal Emotion [14] – Emotion Expressions [11]

	Happiness	Surprise	Anger	Disgust	Sadness
Happiness	0	3.195	2.637	1.926	2.554
Surprise		0	3.436	2.298	2.084
Anger			0	1.506	1.645
Disgust				0	1.040
Sadness					0

Table 4 Distance values between emotions [17]

Conclusion

In this paper, we described basic research to be able to design digital teacher for web lectures. A method for analyzing emotion loadings in text has been investigated. A rule-based approach has been selected for emotion reasoning. This gives opportunities for us to extend both our developed human computer interaction system's believability and behavior. To support the reasoning engine, we also have developed an affective lexicon database, which is depicted in a 2D space. The database gives information about the positive/negative orientation of a word, its intensity and its relative distance to other (emotion) words.

Our approach described in this paper assumed that the emotion loadings in linguistic contents are explicit emotions, shown by emotion words. Our experimental results indicated that an emotion expression for a given emotion word depends mostly on the context and a given word used in various situations can have different meaning. The next step will include emotion analysis from discourse information, e.g. moods, personality characteristics, anaphoric information, background contexts of speech.

Future work is necessary to evaluate the knowledge applied by the developed system. An intensive research also needs to be done in evaluating the knowledge to fit with teacher pedagogical requirements and generates a virtual agent according to teacher's specifications. To simulate a realistic virtual teacher, the physical and social environment and interaction phenomena of learners and teachers have to be taken into account.

Acknowledgments

The research reported here is part of the Interactive Collaborative Information Systems (ICIS) project, supported by the Dutch Ministry of Economic Affairs, grant nr: BSIK03024.

References

[1] Abowd G.D., Classroom 2000: An Experiment with the Onstrumentation of a Living Educational Environment, *IBM Systems Journal*, 1999, 38:508-530.

[2] Arsham H., Impact of the Internet on Learning and Teaching, *USDLA Journal*, ISSN 1537-5080, 16(3), 2002.

[3] Baylor A., The Impact of Pedagogical Agent Image on Affective Outcomes, *IUI'05*, CA, 2005.

[4] Baylor A. and Plant A. Pedagogical Agents as Social Models for Engineering: The influence of Agent Appearance on Female Choice, *Proc. of AIED'05*, The Netherlands, 2005.

[5] Cassell J., Sullivan J., Prevost S., and Churchill E., *Embodied Conversational Agents*, MIT Press 2000.

[6] Cassell J., Vilhjálmsson H. and Bickmore T., BEAT: The Behavior Expression Animation Toolkit, *Proc. of SIGGRAPH*, 2001, 477-486.

[7] Cassell J., Bickmore T., Campbell L., and Vilhjálmsson H. Human conversation as a System Framework: Designing Embodied Conversational Agents. *In: Embodied Conversational Agents*, Cassell, J., Sullivan, J., Prevost, S., and Churchill, E. (eds.), MIT Press, 2000, 29-63.

[8] Collard D., Girardot S. and Deutsch H., From the Textbook to the Lecture: Improving Prelecture Preparation in Organic Chemistry, *Journal of Chemical Education*, 2002, 79:520-523.

[9] Condon W. and Osgton W., Speech and Body Motion Synchrony of the Speaker-Hearer. In D. Horton and J. Jenkins (eds.), *The Perception of Language*, Academic Press, 1971, 150–184.

[10] Day J., Foley J., Groeneweg R. and Van der Mast C., Enhancing the Classroom Learning Experience with Web Lectures, *Proc. of ICCE'05*, ISBN 981-05-4005-1, Singapore, 2005, 638-641.

[11] Desmet P., *Designing Emotion*, Doctoral Dissertation, Delft University of Technology, 2002.

[12] Economou D., Mitchell B., Pettifer S., Cook J., and Marsh J., User Centred Virtual Actor Technology. In *Eg Virtual Reality, Archaeology and Cultural Heritage*, 2001.

[13] Ekman P. and Friesen W.F., *Unmasking the Face*. Englewood Cliffs, USA: Prentice-Hall, Inc, 1975.

[14] Ekman P., Basic Emotions, In Dalgleish T. and Power M., (eds.). *Handbook of Cognition and Emotion*, UK: John Wiley and Sons, Ltd. 1999.

[15] Fellbaum C., *WordNet: An Electronic Lexical Database*, The MIT Press, 1998.

[16] Gulz A. and Haake M., Social and Visual Style in Virtual Pedagogical Agents, *Proc. of the Workshop on Adapting the Interaction Style to Affective Factors, UM'05*, Scotland, 2005.

[17] Hendrix J. and Ruttkay Zs. M., Exploring the Space of Emotional Faces of Subjects without Acting Experience, *ACM Computing Classification System*, 1998.

[18] Hatzivassiloglou V. and McKeown K.R., Predicting the Semantic Orientation of Adjectives, *Proc. of ACL'97*, Spain, 1997.

[19] de Jongh E. and Rothkrantz L.J.M, FED: an Online Facial Expression Dictionary, *Euromedia 2004*, Eurosis, Ghent, April 2004, 115-119.

[20] Kamps J. and Marx M., Words with Attitude, *Proc. of Global WordNet CIIL'02*, India, 2002, 332-341.

[21] Liu H., Lieberman H., and Selker T., *A Model of Textual Affect Sensing using Real World Knowledge*, Technical Report, MIT Media Laboratory, Cambridge, 2003.

[22] Moses G., Litzkow M., Foertsch J., and Strikwerda J., *eTeach -- A Proven Learning Technology for Education Reform*, presented at ASEE/IEEE Frontiers in Education Conference, Boston, MA, 2002.

[23] Mulder M., Nijholt A., den Uyl M. and Terpstra P., A Lexical Grammatical Implementation of Affect, *Proc. of TSD'04*, LNAI, Springer, 2004, 171-178

[24] Nass C.I., Steuer J.S. and Tauber E., Computers are Social Actors, *Proc. of CHI'94*, MA, 1994, 72-78.

[25] Ortony A., Clore G. and Foss M., The Referential Structure of the Affective Lexicon, *Cognitive Science*, 1987, 11: 341-364.

[26] Pelachaud C. and Bilvi M., Computational Model of Believable Conversational Agents, in *Communication in MAS: Background, Current Trends and Future*, Marc-Philippe Huget (eds.), Springer-Verlag, 2003.

[27] Rickel J. and Johnson W.L., Task-Oriented Collaboration with Embodied Agents in Virtual Worlds. *In: Embodied Conversational Agents*, Cassell, J., Sullivan, J., Prevost, S., and Churchill, E. (eds.), MIT Press, 2000, 95-154.

[28] Russell J.A., A Circumplex Model of Affect, *Journal of Personality and Social Psychology*, 1980, 39(6): 1161-1178.

[29] Strapparava C. and Valitutti A., WordNet-Affect: An Affective Extension of WordNet, *Proc. of LREC'04*, Portugal, 2004.

[30] Walker J.H., Sproull L., and Subramani R., Using a Human Face in an Interface, in *Companion of CHI'94 on Human Factors in Computing Systems*, MA, 1994.

[31] Wojdel A. and Rothkrantz L.J.M., Parametric Generation of Facial Expressions Based on FACS, *Computer Graphics Forum*, Blackwell Publishing, 2005, 4(24):1-15.

About the authors

Siska Fitrianie, MSc. PDEng., PhD Researcher, Man-Machine-Interaction Group, Delft University of Technology, Mekelweg 4, 2628CD Delft, The Netherlands, phone: +31 15 278 8543, e-mail: s.fitrianie@ewi.tudelft.nl

Dr. Drs. Leon J.M. Rothkrantz, Associate Professor, Man-Machine-Interaction Group, Delft University of Technology, Mekelweg 4, 2628CD Delft, The Netherlands, phone: +31 15 278 7504, e-mail: l.j.m.rothkrantz@ewi.tudelft.nl