7th Workshop

Emotion & Computing

Current Research and Future Impact

Editor / Organizer

Prof. Dr. Dirk Reichardt, DHBW Stuttgart

Scientific Committee

Contact

www.emotion-and-computing.de

Introduction

The workshop series “emotion and computing – current research and future impact” has been providing a platform for discussion of emotion related topics of computer science and AI since 2006. The main focus of the workshop shifts within the wide range of topics and fields of research related to emotions interpreted and generated by a computer. Motivations for emotional computing are manifold. The scientific papers we discuss this year come from many different fields of research which also go beyond the scope of artificial intelligence. We are expecting a fruitful exchange between researches in these fields of research, especially in the demo session and the moderated discussion session which has become a key component of the workshop.

Dirk Reichardt

Overview

On the Relevance of Sequence Information for Decoding Facial Expressions of Pain and Disgust - An Avatar Study
Michael Siebers, Tamara Engelbrecht, Ute Schmid

A Method for Extracting Colors from Text to Visualize Readers’ Impressions
Tomoko Kajiyama, Isao Echizen

General Purpose Textual Sentiment Analysis and Emotion Detection Tools
Alexandre Denis, Samuel Cruz-Lara, and Nadia Bellalem

Automatic Speech for Poetry - The Voice behind the Experience
Diana Arellano, Cristina Manresa-Yee, Volker Helzle

Analyzing for emotional arousal in HMD-based head movements during a virtual emergency
C. Becker-Asano, D. Sun, C. N. Scheel, B. Tuschen-Caffier, B. Nebel

Using Text Classifcation to Detect Alcohol Intoxication in Speech
Andreas Jauch, Paul Jaehne, David Suendermann

No matter how real: Out-group faces convey less humanness
Aleksandra Swiderska, Eva G. Krumhuber, Arvid Kappas

TARDIS - a job interview simulation platform (demo)
Hazaël Jones, Nicolas Sabouret

On the Relevance of Sequence Information for Decoding Facial Expressions of Pain and Disgust

An Avatar Study

Michael Siebers, Tamara Engelbrecht, and Ute Schmid

Otto-Friedrich-Universität Bamberg, Germany

Abstract. Since Ekman and Friesen published their Facial Action Coding System in 1978 the typical facial expressions of many mental states have been studied. However, most research concentrates on the set of action units present during some facial expression—the sequence of the action units is ignored.

In this paper we investigate human facial expression decoding capabilities on videos of pain and disgust. Stimuli are videos taken of the emotional agent Amber of the EMBR system. Base condition is the simultaneous onset of all involved action units. Treatment conditions are all 6 sequential permutations of mouth, eye, and brow related action units. Additionally we compare the results with the decoding capabilities on still images taken at maximal intensity.

The study with 87 subject shows that facial expressions of disgust are decoded with significantly higher accuracy (82.9%) than expressions of pain (72.0%, $p < .000$). For expressions of pain an ANOVA indicates that the simultaneous onset of action units improves rating accuracy. However, no difference between base and target conditions can be found for facial expressions of disgust. For all conditions the raters' gender and age influences the accuracy significantly ($p < .01$). Across both expression types the rating accuracy is higher for videos (79.5%) than for stills (48.9%).

1 Motivation

Communication is an important part of every-day life. The main part of conversation are verbal utterances. However, facial expressions and other nonverbal cues also have a significant value in face-to-face communication. They unconsciously convey the mental state of the utterer. Facial expressions are also a part of human-computer-interaction. On the one hand it is desirable to guess the mental state or sentiment the user is in. On the other hand avatars gain credibility displaying facial expressions.

The facial action coding system (Ekman and Friesen, 1978) is the most used facial expression description system in the behavioural sciences. Ekman and Friesen described 43 independent facial movements—called *action units*. In most research the relation between mental states and facial expressions is reported by the occurrence and intensity of action units during some time period. However, there might be relevant information in the sequence in which those action units appear.

We have selected two sentiments—pain and disgust—and investigate whether the sequence of action units influences the sentiment decoding performance of human observers.

In the next section we will present related work trying to model facial expressions of pain or disgust. In Section 3 we will describe the used stimuli and the experiment set-up. The results of these experiments will be presented in Section 4. Finally, we will conclude in Section 5.

2 Related Work

Fabri et al. developed an avatar system able to display 11 action units directly. They claim that these action units are sufficient to encode the six basic emotions (Ekman and Friesen, 1971). They validated this using a study with 29 subjects (Fabri et al., 2004). They compared decoding correctness between videos of their avatar and images of actors taken from the *Pictures of Facial Affect* database (Ekman and Friesen, 1975). The identification ratio was significantly higher (78.6%) in the images than in the virtual head (62.2%). *Disgust* in the virtual head was identified in the fewest cases (around 20%). They used different action unit combinations for each emotion. No details are available on the sequence of action unit motion onsets.

Paleari and Lisetti (2006) modelled facial expressions according to Scherer's (1982) multi-level theory of emotions. Scherer postulates that emotions are the result of the sequential evaluation of events. The evaluation steps are always conducted in the same sequence. Scherer associated action units with some of the steps performed. Paleari and Lisetti used the action units in the sequence of the checks postulated by Scherer. They designed a single time pattern for each of the six basic emotions. In an evaluation study with an unknown number of participants *disgust* was identified in 64% of the cases.

To our knowledge no work was conducted regarding the animation of facial expressions of pain. However, Schmid et al. (2012) tried to identify the temporal action unit pattern of facial expressions of genuine pain. Pain was mechanically induced to 124 voluntary subjects. The shown facial expressions were coded into action unit sequences. An artificial grammar for the resulting sequences was induced using alignment based learning. The coverage of the resulting grammar was estimated at 65% using cross validation.

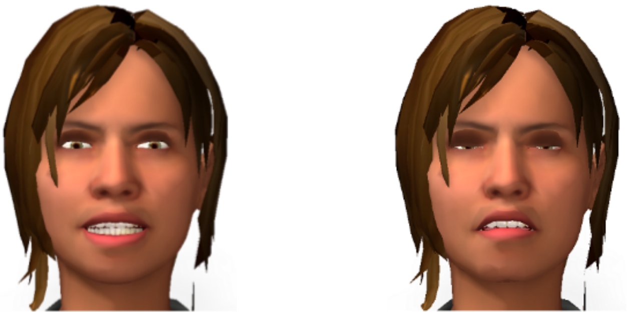
3 Experiments

We conducted an online experiment to investigate the influence of facial expression animation type on human decoding possibilities. Stimuli showing facial expressions of pain, disgust, or a neutral expression were presented to participants. Then subjects had to categorize the stimuli as *pain*, *disgust*, or *neutral.*

In the next section we will present the different types of stimuli and their generation. Afterwards we will detail the structure and realization of the online experiment.

3.1 Stimuli Generation

Using the EMBR System (Heloir and Kipp, 2010) developed by the Embodied Agents Research Group at DFKI we animated facial expressions of pain and disgust. Pain is expressed using the action units 4 (Brow Lowerer), 7 (Lid Tightener), and 9 (Nose Wrinkler) according to findings from Prkachin (1992). Action units 9 (Nose Wrinkler), 15 (Lip Corner Depressor), 16 (Lower Lip Depressor), and 7 (Lid Tightener) were used for disgust. Though the facial expressions in the EMBR system are not based on action units, the movement of brows and eyes can be approximated well. However, nose wrinkling (action unit 9) can not be modelled directly. We tried to mimic this action unit using mouth trajectories intended for phonemes. Static images of the modelled sentiments can be seen in Figure 1.



(a) Pain

(b) Disgust

Fig. 1. Facial expressions for sentiments at their highest intensity.

The facial expressions were animated using seven different schemes. In the *simultaneous* scheme all facial regions were animated simultaneously. In the sequential schemas the three regions eyes, brows, and mouth were animated sequentially. All six permutations of the region sequence were used. Additionally we extracted still images from the simultaneous schemes showing all regions at full effect (like in Figure 1).

3.2 Procedure

The experiment was conducted as online experiment using the soSci-System (`www.sosci.de`). Subjects were randomly assigned to three experiment conditions: (a) *images*, (b) *simultaneous*, or (c) *sequential*. The three conditions differ

Table 1. Distribution of subjects' age and gender over the experiment conditions. Age is given as mean ± standard deviation.

condition	gender		age
	female	male	
image	12	19	28.3 ±10.1
simultaneous	12	13	28.0 ± 10.6
sequential	11	19	29.8 ± 12.0
overall	35	51	28.7 ± 10.8

Table 2. Answer probabilities for pain and disgust stimuli in the experiment conditions.

(a) Simultaneous

	sentiment	
answer	pain	disgust
pain	85.9%	19.2%
neutral	3.8%	0.00%
disgust	10.3%	80.8%

(b) Sequential

	sentiment	
answer	pain	disgust
pain	70.7%	7.3%
neutral	20.5%	8.4%
disgust	8.8%	84.3%

(c) Image

	sentiment	
answer	pain	disgust
pain	52.7%	32.6%
neutral	26.9%	10.9%
disgust	20.4%	56.5%

only in the stimuli used. The images conditions uses images, the simultaneous condition uses videos with simultaneous animation onset, and the sequential condition uses videos with all six sequential schemas.

The setting was identical for each stimulus: The stimulus is presented and the subject is asked, which sentiment the subject would attribute to this stimulus[1]. Possible answers were *pain*, *neutral*, and *disgust*. Additionally each subject was asked for his age and gender.

4 Results

The experiment was completed by 89 subjects. Two of the subjects showed rather many unanswered questions (33% and 23%, respectively). Both subjects were excluded from the evaluation. Of the remaining 87 subjects 35 were female, 51 male, and 1 subject did not give his gender. Subjects were between 14 and 58 years old (mean=28.7).

A total of 31 subjects were assigned to the *image* condition, 26 to the *simultaneous*, and 30 to the *sequential* condition. There is no bias in age or gender in the condition assignment (see Table 1).

At first we analyse the sequential condition. Subject in this group saw animated videos where the sequential movement onset of eyes, brows, and mouth was permuted. We compared the response behaviours for the six permutations using a chi-squared test. The null hypothesis that all six permutations trigger

[1] Welche Empfindung würden Sie dem abgebildeten Gesichtsausdruck zuweisen?

Table 3. Answer correctness for simultaneous and sequential condition. Age is grouped by quartiles.

(a) Gender

	simultaneous		sequential	
sentiment	female	male	female	male
pain	86.1%	84.6%	73.5%	69.1%
disgust	88.9%	74.4%	84.6%	84.2%

(b) Age

	simultaneous				sequential			
sentiment	14–22	23	24–27	27–58	14–22	23	24–27	27–58
pain	91.7%	92.6%	83.3%	66.7%	75.1%	79.4%	73.6%	56.2%
disgust	91.7%	85.2%	83.3%	53.3%	92.8%	93.5%	85.9%	66.9%

the same response behaviour could not be dismissed ($p > .969$). So the sequential onset of facial movements must be considered as one condition disregarding the exact sequence. The answer probabilities for this group are presented in Table 2(b).

The data shows a clear difference in the response behaviour between the sequential and the simultaneous group.[2] On the one hand pain is identified more often in the simultaneous condition, on the other hand disgust is identified more often in the sequential condition. To analyse this difference in more detail we evaluated the answers for correctness. The percent of correct answers for both video conditions and their subgroups can be seen in Table 3.

The tables show that the age and the gender of the subject influence the rating. To investigate on this we fitted a logit-model for each sentiment individually. We included the subjects age, gender, the experiment condition, and all two-variable interactions of those in the model. After fitting the model we conducted an analysis of variants.

For the *disgust* sentiment gender and age have a significant influence. The ANOVA showed a test statistic of $\chi^2 = 6.176$ ($p = .01295$) for gender. The influence of age was highly significant $p < .001$ ($\chi^2 = 67.908$). Other variables or interactions have no significant influence. All reported influences are negative. Odds-ratios for the logit-model are shown in Table 4(a).

The results for the *pain* sentiment were similar. Gender has a significant influence ($\chi^2 = 5.440$, $p = .01968$), age has a highly significant influence ($\chi^2 = 42.640$, $p < .001$). However, animating the facial expression sequentially has a significant influence on the pain sentiment ($\chi^2 = 5.826$, $p = .01579$). Other variables or interactions have no significant influence. All reported influences are negative. Odds-ratios for the logit-model are shown in Table 4(b).

[2] A chi-squared goodness-of-fit test shows that this difference is highly significant ($p < .001$).

Table 4. Coefficients, odds-ratios, and p's for the logit-models for pain and disgust identification. Only factors with a significant influence according to the ANOVA for at least one model are shown.

(a) Disgust

factor	coefficient	odds-ratio	p
(Intercept)	5.68525	294.491	5.3e-05
animation=sequential	-0.41769	0.659	.770047
gender=male	-2.42177	0.089	.031650
age	-0.10315	0.902	.000636

(b) Pain

factor	coefficient	odds-ratio	p
(Intercept)	4.268784	71.435	.000136
animation=sequential	-1.605593	0.201	.162424
gender=male	-0.221821	0.801	.805518
age	-0.075276	0.927	.005415

The response behaviour for the *image* condition is shown in Table 2(c). Pain and disgust are only identified in around 50% of the cases. Both sentiments are rather confused with each other than denied. This is different from both video conditions. Chi-squared test confirm this for the simultaneous condition ($\chi^2 = 56.4744$, $p < .001$) and the sequential condition ($\chi^2 = 259.2122$, $p < .001$). For the *disgust* sentiment an analysis of variances over the logit-models for answer correctness shows that only the subjects' gender has a significant influence ($\chi^2 = 6.6294$, $p = .010$). The coefficient in the model cannot be estimated significantly. For the *pain* sentiment the same analysis shows no significant factor.

5 Conclusion

We investigated the identification rate of pain and disgust sentiments. We showed that sentiments on video are easier identified than still images of the sentiments. We showed that there is a difference between animating facial regions simultaneously or sequentially. It is better to use sequential animation for videos of disgust. No influence was detected for pain videos. However, using sequential animation reduced the confusion rate between both sentiments.

Future work will consider a finer granularity of animation onset. It is not necessary that all facial regions are animated simultaneously or sequentially. The most natural expression might be gained animating some regions simultaneously after other regions. Additionally the notion of *sequentially* might be refined. Not only the movement sequence but also the exact time between onsets might be of interest.

Additionally, future stimuli should include the nose wrinkling action unit. Since this action unit is prototypical for both sentiments the overall identification rate should increase.

Bibliography

Ekman, P. and Friesen, W. (1971). Constants across cultures in the face and emotion. *Journal of Personality and Social Psychology*, 17:124–129.

Ekman, P. and Friesen, W. (1975). Pictures of facial affect CD-rom. University of California, San Francisco.

Ekman, P. and Friesen, W. (1978). *Facial Action Coding System: A Technique for the Measurement of Facial Movement.* Consulting Psychologists Press, Palo Alto.

Fabri, M., Moore, D., and Hobbs, D. (2004). Mediating the expression of emotion in educational collaborative virtual environments: an experimental study. *Virtual Reality*, 7(2):66–81.

Heloir, A. and Kipp, M. (2010). Real-time animation of interactive agents: Specification and realization. *Applied Artificial Intelligence*, 24(6):510–529.

Paleari, M. and Lisetti, C. (2006). Psychologically grounded avatars expressions. In Reichardt, D., Levi, P., and Meyer, J.-J. C., editors, *Proceedings of 1st Workshop on Emotion and Computing at KI 2006.*

Prkachin, K. M. (1992). The consistency of facial expressions of pain: a comparison across modalities. *Pain*, 51:297–306.

Scherer, K. (1982). Emotion as a process: Function, origin and regulation. *Social Science Information*, 21:555–570.

Schmid, U., Siebers, M., Seuß, D., Kunz, M., and Lautenbacher, S. (2012). Applying grammar inference to identify generalized patterns of facial expressions of pain. In Heinz, J., de la Higuera, C., and Oates, T., editors, *Proceedings of the 11th International Conference on Grammatical Inference*, Heidelberg. Springer.

A Method for Extracting Colors from Text to Visualize Readers' Impressions

Tomoko Kajiyama[1] and Isao Echizen[2]

[1]Aoyama Gakuin University
5-10-1 Fuchinobe, Chuo-ku, Sagamihara-shi, Kanagawa, 252-5258 JAPAN
tomo@ise.aoyama.ac.jp

[2]National Institute of Informatics
2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo, 101-8430 JAPAN
iechizen@nii.ac.jp

Abstract. The image on a book cover gives potential buyers not only an impression of the book's contents but also a clue for search and browsing before or after buying the book. We have developed a color extraction method for use as the first step in automatically creating book cover images that reflect readers' impressions. We constructed a database expressing the relationships between adjectives and colors and extracted colors from text such as sentences in a book and reader reviews. We performed an experiment with 15 participants in which the task was to read one of the books, write a report about it, select the color pattern that best expressed the reader's impression, and write the reason for selecting it. We found that the colors extracted using this method were more consistent with the colors selected by participants than the colors in the actual cover and that our method is effective for readers who are satisfied with the book.

Keywords: color extraction, adjective, book cover image, user review, color psychology

1 Introduction

The use of electronic devices for reading digital books is quickly spreading, and the market for digital books is growing rapidly [1]. There are two basic types of digital books: printed books that have been digitalized and books prepared only in digital form. The use of electronic devices for reading has changed not only how people read books but also how they select and buy books. Moreover, some digital books come without a cover, unlike printed books.

A book cover serves an important function—it gives potential buyers an impression of the book. Digital books sold online [2,3] that are in the public domain or are original digital books and are provided without book cover images are often given images containing only the book title with a standard design. The colors in the image are selected on the basis of the genre or are simply randomly chosen. As a result, the

adfa, p. 1, 2011.

cover image is of little use in searching for a book and browsing the book's contents both before purchasing the book and when it is on the purchaser's virtual bookshelf. Several services and applications have been developed for browsing a book's contents. A service has been developed for designing an image for the cover of a digital book [4], and an application has been developed that enables users to design such an image by themselves [5]. However, using this service is costly, and using this application is time-consuming.

To support a user's search for a digital book, it is important to create book cover images automatically. In general, the higher the person's expectations before reading, the lower the level of satisfaction after reading, and the lower the expectations before reading, the greater the chance of opportunity loss [6]. It is thus important to reduce the gap between the impression obtained from a book's cover image and the impression gained by reading the book. To realize a function that can do this, we have to consider not only the contents of the book but also the impressions of its readers. Reader impressions can be found in the reader reviews commonly found on sites selling digital books. Potential buyers can read the reviews and use them to make a purchase decision. However, a reader's emotions and latent comments about a book are most likely to be represented in images, not text [7]. In addition, the meaning of something can generally be understood more quickly from an image than from characters [8].

We are developing a method for automatically creating book cover images that reflect reader impressions. As the first step, we focused on defining colors for the image because colors in images are generally considered to be the most significant aspect [9]. This is because colors create various kinds of associations and affect faculties such as imagination and fantasy. To develop a method for extracting color from texts, we focused on the relationships between colors and adjectives since adjectives typically represent the inner emotional state. We constructed a database expressing the correspondence between adjectives and colors and created a method for extracting color from texts such as the text in a digital book and the text in reader reviews. To evaluate this method, we performed a usability test in which 15 participants were tasked with reading a book, writing a report about it, selecting the color pattern that best expressed the reader's impression, and writing the reason for selecting it

2 Color Extraction Method

2.1 Overview

Colors reflecting the reader impressions of a digital book are extracted by using the text in the book and reader reviews of the book. To visualize the emotions described in the book and the feelings of the reader, it is necessary to extract the words describing them. There are two types of emotion: emotion that can be observed externally by others and described objectively, such as "surprised" and "fired up," and emotion that represents an inner state based on emotion, such as "terrible (story)" and "moving (event)." In general, the former type is described using verbs, and the latter type is described using adjectives [10].

In the method we developed, colors are extracted using adjectives because our objective is to reflect reader impressions, i.e., to reflect the inner emotions of readers. Figure 1 shows the flow of this method. Morphological analysis is performed on the input text (sentences in the book and reader reviews) to extract the adjectives. Scores are then calculated for the adjectives. A score is calculated for each adjective on the basis of the number of occurrences of adjectives and other words. A score is calculated for each color by using a score for each adjective and a color database expressing the relationship between adjectives and colors. Colors with higher scores are then extracted for use in the cover image.

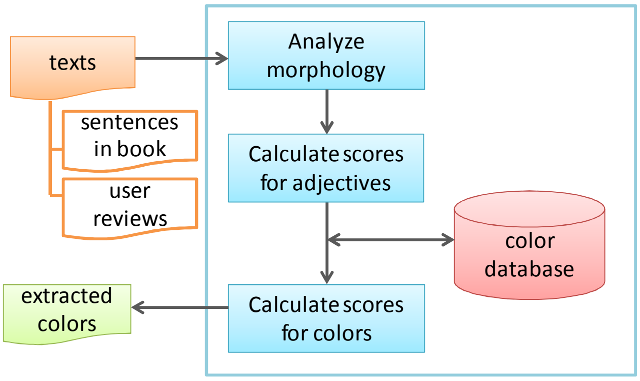


Fig. 1. Flow of proposed method

2.2 Color Database

We constructed the color database using a color image scale [11] expressing the relationships between adjectives and colors. This scale defines 180 adjectives representing basic emotions for 130 colors that capture experiences psychologically. This database has three attributes: color (RGB), adjective, and frequency of use. The frequency of use is defined on a five-star scale—the greater the number of stars, the more strongly the color represents the image of the adjective [11]. Each color corresponds to more than 1 but less than 25 adjectives. The 180 adjectives were extended to 3184 by using a thesaurus [12].

2.3 Scores for Adjectives

The morphological analysis extracts the adjectives from the input text, and a score is calculated for each one. The total number of occurrences of adjectives in the book $(a_{11}, a_{12}, a_{13}, \cdots, a_{1n})$ is defined as $(x_{11}, x_{12}, x_{13}, \cdots, x_{1n})$, and the total number of occurrences of adjectives in the reader reviews $(a_{21}, a_{22}, a_{23}, \cdots, a_{2m})$ is defined as $(x_{21}, x_{22}, x_{23}, \cdots, x_{2m})$. The total number of words in the book is n_1, and the total number of words in the reader reviews is n_2. Weight w_{1i} for an adjective in book a_{1i} is calculated using $w_{1i} = a_{1i}/n_1$, and weight w_{2i} for an adjective in user review a_{2i} is calculated using $w_{2i} = a_{2i}/n_2$.

The set of adjectives $(b_1, b_2, b_3, \cdots, b_\mathrm{r})$ is created by eliminating duplication between $(a_{11}, a_{12}, a_{13}, \cdots, a_{1n})$ and $(a_{21}, a_{22}, a_{23}, \cdots, a_{2m})$. The score $(y_1, y_2, y_3, \cdots, y_r)$ for $(b_1, b_2, b_3, \cdots, b_\mathrm{r})$ is calculated using

$$\mathrm{y}_i = \begin{cases} w_{1j} + w_{2k} & (\mathit{if} \;\; \mathrm{b}_i = a_{1j} = a_{2k}) \\ w_{1j} & (\mathit{if} \;\; \mathrm{b}_i = a_{1j}\;) \\ w_{2k} & (\mathit{if} \;\; \mathrm{b}_i = a_{2k}\;) \end{cases} .$$

2.4 Scores for Colors

The colors in the color database correspond to various numbers of adjectives, so we normalized the frequency of use because the sums of the frequency of use are different. If a given color C_j has defined adjectives $(d_{j1}, d_{j2}, d_{j3}, \cdots, d_{jp})$ and frequency of use $(t_{j1}, t_{j2}, t_{j3}, \cdots, t_{jp})$, the weights for the adjectives $(z_{j1}, z_{j2}, z_{j3}, \cdots, z_{jp})$ are calculated using

$$z_{jk} = t_{jk} / \textstyle\sum_{m=1}^{p} t_{jm} .$$

The score for C_j $S_j$is calculated using the adjective scores and the weight for each adjective:

$$S_j = \sum y_i \times z_{jk} \; (\mathit{if} \; b_i = d_{jk}\,) .$$

The calculated S_j are arranged in ascending order, and the colors with higher scores are extracted on the basis of a threshold.

3 Evaluation

3.1 Overview

An experiment was performed to determine how much the colors extracted using our method were consistent with the colors expressing readers' impressions. The participants were 15 employed people in their 20s and 30s. We randomly selected 3 novels (paperback version) from Amazon's Japanese site[1] for which there were more than 20 reader reviews. Table 1 summarizes the details.

Table 1. Features of each book

ID	Book		Reviews		
	No. of words	No. of adjectives	No.	Average no. of words	Average no. of adjectives
(a)	61,728	6,062	28	409	51
(b)	46,773	4,316	35	284	36
(c)	42,569	3,937	22	215	27

[1] http://www.amazon.co.jp

The participants were tasked with reading one of the books, writing a report about it, selecting the color pattern that best expressed the reader's impression of the book, and writing the reason for selecting it. The books were provided to the participants without a cover to avoid influencing their impressions.

3.2 Details

There are two ways of using our color extraction method; one way is to extract colors from all reviews en masse, and the other is to extract the colors from each review one-by-one and then select the high-frequency colors. We defined the threshold mentioned in section 2.4 as an upper value of 15%, i.e., the average ratio for the top three colors, because the average number of extracted colors for the three books was 3.3. If the first way is used, adjectives that occur with low frequency from the individual review viewpoint could be become high-frequency ones from the total viewpoint. We thus used the second way to better capture the readers' impressions.

We prepared 14 color patterns on the basis of colors extracted using our method and colors extracted from the actual book cover images. Figure 2 shows the ones prepared for book (a). We limited the number of patterns to 14 because a large number of color combinations can give people many different kinds of impressions [11]. The pattern extracted using our method is shown at the top left in Figure 2.

The pattern extracted from the actual book cover is shown at the bottom right in Figure 2. Extraction from the cover was done in three steps: 1) change each pixel in the printed color image into 1 of the 130 colors in the color database by selecting the closest color in RGB color space; 2) calculate number of occurrence of each color; 3) select high-occurrence colors. The number of selected colors was same number of colors extracted using our method with threshold described above.

Except for the color pattern at the bottom right in Figure 2, the first (leftmost) color in each pattern was created on the basis of the extracted color using our method; 12 color patterns and 1 monotone pattern. We divided the hue values into 12 values on the basis of the extracted color to create color patterns, and changed the extracted color to monotone to create a monotone pattern.

For the second-to-last color in each pattern, the extracted colors and colors from the printed covers were alternatively used, e.g., in case of book(a) shown in Figure 2, it means that the second (rightmost) color in each pattern because the number of extracted colors using our method was only two. We printed the 14 color patterns randomly on the questionnaires given to the participants.

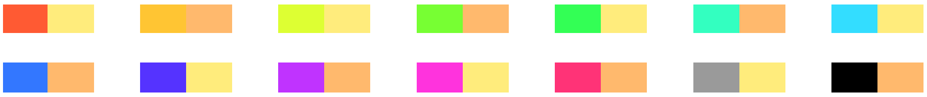


Fig. 2. Color patterns

3.3 Results

Figure 3 shows the percentage of participants who selected each pattern for each book. Pattern (3) is the pattern extracted using our method. The pattern numbers between 1 and 12 shows the patterns based on hue, so pattern (1) is actually next to pattern (12) because hue value is a circular continuous quantity. Only one participant selected the pattern of the printed book cover (book (c)).

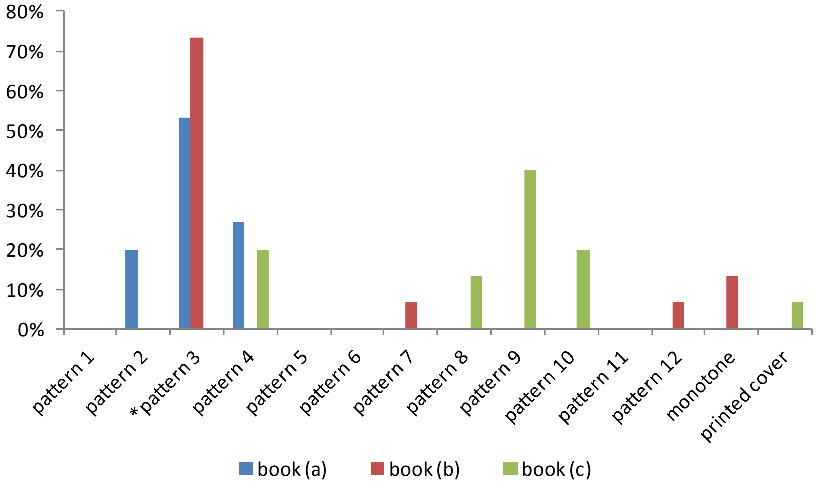


Fig. 3. Results

For book (a), all of the participants selected a pattern near the pattern extracted with our method, so there was not a big difference in terms of hue. More specifically, eight participants selected reddish orange, three selected orange-yellow, and four selected reddish pink. We attribute this finding that book (a) gave a similar impressions to all of the participants to the fact that the colors extracted from the reader reviews were highly consistent.

For book (b), many participants selected the pattern extracted with our method. Two participants selected the monotone pattern, and we determined that they were not satisfied with reading the book. Their reports were simply summaries of the story and contained only four of the adjectives registered in the color database. Eleven of the participants who selected the pattern extracted with our method were satisfied with reading the book, and they described their impression in their reports with words such as “exciting,” “interesting,” and “attractive.” This demonstrates that our method is effective for readers who are satisfied with the book.

For book (c), none of the participants selected the pattern extracted with our method. Three participants selected purple (pattern 4), which is next to the pattern extracted with our method, two selected bluish green, six selected green, and three selected yellowish green. Many participants selected colors related to green because they put a higher priority on the colors of particular scenes rather than on their impression after reading the book. The participants who selected green and yellowish green wrote words such as “woods” and “tree” in their reports. This means that we should consider not only adjectives, which express reader impressions, but also nouns, which represent scenes in the story.

3.4 Discussion

There are various approaches to browsing books in the virtual world, including using a graphical search interface that replicates the image of a bookshelf [13] and automatically capturing a book cover on a library counter [14]. These approaches focus on recreating the real work browsing environment in a virtual world. Using our method to create book cover images should make book browsing and searching in the virtual world more effective because it enables users to intuitively catch the general impressions of readers.

Possible applications of our method include a real-time book recommendation system The book recommendation systems on e-commerce sites are generally based on text such as reader reviews and information about the books purchased or browsed by other users. Since it is difficult to assess a user’s present emotions and preferences because they shift over time, recommendations can sometimes be misleading [15]. The display of a cover image based on general reader impressions during a book search should enable users to assess the impressions generally created by reading the book and to determine whether they are line with the user’s present feelings.

4 Conclusion

We have developed a method for automatically extracting colors from text for use in creating a book cover image that reflects reader impressions. We constructed a color database expressing the relationships between adjectives and colors and developed a method for extracting colors from texts such as digital books and reader reviews. We evaluated this method experimentally by tasking 15 participants to read a book, write a report about it, and select the color pattern that best expressed the reader’s impression. We found that the colors extracted using this method were more consistent with the colors in the images drawn by the participants than the colors in the actual cover and that our method is effective for readers who are satisfied with the book.

To create an effective cover image, i.e., one reflecting reader impressions, it is necessary not only to improve the accuracy of the color extraction by enhancing the word analysis but also to optimize the arrangement of the extracted colors, to extract nouns which represent scenes in the story or symbolic objects from body text, and to represent bibliographic information (title, authors, etc.).

References

1. ICT Research & Consulting, http://www.ictr.co.jp/report/20120710000020.html

2. Aozora Bunko, http://www.aozora.gr.jp/
3. Calibre, http://calibre-ebook.com/
4. libura, http://libura.com/
5. Interwired, TimelyResearch, http://www.dims.ne.jp/timelyresearch/2009/090202/
6. Dawes, R.M., Singer, D., Lemons, F. An experimental analysis of the contrast effect and its implications for intergroup communication and the indirect assessment of attitude, Journal of Personality and Social Psychology, Vol. 21, No. 3, pp. 281–295, 1972.
7. N. Akamatsu. The Effects of First Language Orthographic Features on Second Language Reading in Text, Language Learning, Vol. 53, No. 2, pp. 207–231, 2003.
8. K. Shimazaki, Psychological Color, Bunka Shobo Hakubunsha, 1990.
9. T. Oyama, Introduction of Color Psychology. Chuokoron-Shinsha, 1994.
10. M. Ohso, Verbs and Adjectives of Emotion in Japanese, Studies in Language and Culture, Vol. 22, No. 2, pp. 21–30, 2001.
11. S. Kobayashi. Color Image Scale, Kodansha, 2001.
12. Weblio Thesaurus, http://thesaurus.weblio.jp/
13. Shinsho-map, http://bookshelf.shinshomap.info/
14. T. Miyagawa et al., Automation of back cover image generation in virtual bookshelf, Vol. 30, pp. 25–38, 2006.
15. Marketing Charts, Online Product Recommendations Miss Mark, http://www.marketingcharts.com/interactive/online-product-recommendations-miss-mark-11848/

Acknowledgments. This research was supported in part by a Japan Society for the Promotion of Science Grant-in-Aid for Young Scientists (B) (No. 20454085) and the Telecommunications Advancement Foundation. We greatly appreciate the help of the participants in the experiment.

General Purpose Textual Sentiment Analysis and Emotion Detection Tools

Alexandre Denis, Samuel Cruz-Lara, and Nadia Bellalem

LORIA UMR 7503, SYNALP Team
University of Lorraine, Nancy, France
{alexandre.denis, samuel.cruz-lara, nadia.bellalem}@loria.fr

Abstract. Textual sentiment analysis and emotion detection consists in retrieving the sentiment or emotion carried by a text or document. This task can be useful in many domains: opinion mining, prediction, feedbacks, etc. However, building a *general purpose* tool for doing sentiment analysis and emotion detection raises a number of issues, theoretical issues like the dependence to the domain or to the language but also pratical issues like the emotion representation for interoperability. In this paper we present our sentiment/emotion analysis tools, the way we propose to circumvent the difficulties and the applications they are used for.

Keywords: sentiment analysis, emotion detection, applications

1 Sentiment Analysis and Emotion Detection from text

1.1 Definition

One of the most complete definition of the sentiment analysis task is proposed by Liu [13] in which a sentiment is defined as a quintuple $\langle e, a, s, h, t \rangle$ where e is the name of an entity, a an aspect of this entity, s is a sentiment value about a, h is the opinion holder and t is the time when the opinion is expressed by h. The sentiment analysis task consists in outputing for a given text or sentence the set of sentiments that this text conveys. Whereas sentiment analysis is limited in general to binary sentiment value (positive, negative) or ternary (positive, negative, neutral), emotion detection consists in determining the sentiment value among a larger set of emotions, typically Ekman's emotions [7]: joy, fear, sadness, anger, disgust, or surprise.

1.2 Difficulties and existing approaches

A difficult aspect of sentiment analysis is the fact that a given word can have a different polarity in a different context. In [25], authors oppose the *prior polarity* of a word, that is the polarity that a word can have out of context, and the *contextual polarity*, that is the polarity of a word in a particular context. There are many complex phenomena that influence the contextual valence of a word

[17, 25, 13] as the table 1 shows, and most of the approaches thus reduce the scope of the problem to $\langle s \rangle$ the sole sentiment value or sometimes to $\langle e, s, h \rangle$, the entity, the sentiment value and the holder like in [11].

Phenomenon	*Example*	*Polarity*
Negation	it's not good ; no one thinks it is good	negative
Irrealis	it would be good if... ; if it is good then ...	neutral
Presupposition	how to fix this terrible printer? can you give me a good advice?	negative neutral
Word sense	this camera sucks ; this vaccum cleaner sucks	negative vs positive
Point of view	Israel failed to defeat Hezbollah	negative or positive
Common sense	this washer uses a lot of water	negative
Multiple entities	Ann hates cheese but loves cheesecake	negative wrt cheese positive wrt cheesecake
Multiple aspects	this camera is awesome but too expensive	positive wrt camera negative wrt price
Multiple holders	Ann hates cheese but Bob loves it	negative wrt Ann positive wrt Bob
Multiple time	Ann used to hate cheese and now she loves it	negative wrt past positive wrt present

Fig. 1. Examples of linguistic phenomena that influence the final valence of a text

Given the complexity of the task, it is not a surprise that the first approaches to the problem use machine learning. In [24] an unsupervised approach is proposed. It uses the Pointwise Mutual Information distance between online reviews and the words "excellent" and "poor". Its accuracy ranges from 84% for the automobile reviews to 66% for the movie reviews. In [16], a corpus of movie reviews and their annotation in terms of number of stars is used to train several classifiers (Naive Bayes, Support Vector Machine and Maximum Entropy) and obtain a good score for the best (around 83%). More recent work in machine learning approaches to sentiment analysis explore successfully different kinds of learning algorithms such as Conditional Random Fields [15] or autoencoders which are a kind of Neural Networks and which offer good performance on the movie reviews domain [21]. A detailed and exhaustive survey of the field can be found in [13].

2 Cross Domains problems

A general purpose sentiment analysis or emotion detection tool is meant to work in different domains with different applications and thus faces at least three problems: the dependence of the algorithms to the domain they were developed on, the representation of emotions/sentiments for interoperability, and the fact that different applications may require other languages than English, and as such the multilinguality issue must be considered. We detail these three problems in this section.

2.1 Domain dependence

Machine learning dependence An important issue of supervised machine learning is the dependence to the training domain. Classical supervised algorithms require a new training corpus each time a new domain is tackled. In [2] several methods are tried to overcome the domain-dependence of machine learning and they show that the best results can be obtained by combining small amounts of labeled data from the training domain and large amounts of unlabeled datas in the target domain. Actually, unsupervised or semi-supervised machine learning seems more adequate than purely supervised machine learning to reach domain-independence [18]. Another approach [1] is to use hybrid methods, classifiers trained on corpora and polarity lexicons. Indeed, polarity lexicons, such as Sentiwordnet [8] or the Liu Lexicon [9], being in general domain-independent seem to be an interesting track to follow.

Types of emotions dependence Moreover the set of relevant emotions depends on the domain. In a generic emotion analysis tool, there is not much choice apart providing a set of the least domain specific emotions, hence the frequent choice of Ekman's emotions. These may not describe accurately the affective states in all domains, for instance their use is criticized in the learning domain [12, 3], but their independence to the domain and the existence of Ekman's based emotional lexicons such as WordNet-Affect [22] makes them a common practical choice.

2.2 Interoperability

The representation of emotions for interoperability is an important issue for a sentiment/emotion analysis tool that is meant to work in several domains and with several applications. We advocate the use of EmotionML [20] a W3C proposed recommendation for the representation of emotions. An interesting aspect of EmotionML is the acknowledgement that there exists no consensus on how to represent an emotion, for instance is an emotion better represented as a cognitive-affective state like [12], as a combination between pleasure and arousal like [3], or as an Ekman emotion [7]? Thus, EmotionML proposes instead an emotion skeleton whose features are defined by the target application. An emotion is defined by a set of descriptors, either dimensional (a value between 0 and 1) or categorical (a discreet value), and each descriptor refers to an emotional vocabulary document. An emotion and its vocabulary can be embedded in one single document or the emotion can refer to an online vocabulary document.

2.3 Multilinguality

A general purpose sentiment/emotion analysis tool is also required to be working in other languages than English. Most of the work related to multilinguality is tied to subjectivity analysis, a simpler sentiment-like analysis which consists in determining whether a text conveys an objective or subjective assessement. Several solutions are possible, for instance training classifiers on translated corpora, using translated lexicons, building lexicons or corpora for targeted language [5]. Recent experiments with automatic translation for sentiment analysis show that the performance of machine translation does not degrade the results too much [4]. We refer the interested reader to [5, 13] for a good overview of the topic.

3 Tools and applications

We present here the sentiment/emotion analysis tools and the applications that use them in the context of the Empathic Products ITEA2 project (11005)[1], a european project dedicated to the creation of applications that adapt to the intentional and emotional state of the users.

3.1 Sentiment/emotion analysis tools

We implemented several sentiment analysis and emotion detection engines, and will briefly present them here. All of them are integrated in one Web API that takes text in input and returns a single emotion formatted in EmotionML format, dimensional valence for sentiment analysis engines and categorical emotion for the emotion detection engines[2]. A support also exists for non-English languages following [4] using Google machine translation but this service has not yet been evaluated.

For emotion detection, the approach uses an emotion lexicon, namely WordNet-Affect [22] by detecting emotional keywords in text complemented with a naive treatment of negation which inverts the found emotion, ad hoc filters (smileys, keyphrases) and simple semantic rules. Despite its simplicity this tool manages to reach performance similar to other approaches when evaluated on the Semeval-07 affective task dataset [23], it obtains 54.9% of accuracy given an emotion to valence mapping (joy is positive and the others are negative).

For sentiment analysis, we are currently exploring two opposed approaches, one symbolic and one with machine learning. The symbolic approach follows [17] as an attempt to both tackle the linguistic difficulties we mentioned thanks to valence shifting rules and domain dependence by using general purpose lexicons. It works by first retrieving the prior polarity of words as found in the Liu lexicon [9] after a part-of-speech tagging phase with the Stanford CoreNLP library. Then, a parsing phase enables to construct the dependencies (also with Stanford CoreNLP), the resulting dependencies are filtered such that prior word valence is propagated along the dependencies following manually crafted rules for valence shifting or inversion. This approach enables to be more precise, for instance the sentence "I don't think it's a missed opportunity" would be tagged as positive

[1] http://www.empathic.eu/

[2] The Web API is currently accessible on http://talc2.loria.fr/empathic

by the application of two valence flipping rules, a modifier rule for "missed opportunity", and a verb negation rule for "don't think". This approach obtains 56.3% accuracy on the Semeval-07 dataset and 65.86% accuracy on the Semeval-13 data set. The impact of rules has also been evaluated and show that the rules enable to gain 5% accuracy on the Semeval-13 dataset as opposed to a simple lexical approach that only takes the average valence of all words contained in a sentence.

The second approach relies on machine learning by training a classifier, namely a Random Forest classifier evaluated on the Semeval-13 dataset. It uses simple features such as the stemmed words and the part-of-speech but manages to obtain 64.30% accuracy on Semeval-07 and 60.72% accuracy on Semeval-13 both evaluated with 10-fold cross validation. We also performed preliminary evaluation of the cross-domain abilities of the statistical approach and observed that when trained on the Semeval-07 dataset and evaluated on Semeval-13 it obtains 47.08% accuracy. While training it on Semeval-13 and evaluating it on Semeval-07, it obtains 55.5% accuracy. The significant difference is likely caused by the dataset dissimilarities: first the Semeval-07 dataset is much smaller than Semeval-13 (1000 utterances vs 7500 utterances) and then, training on Semeval-07 is less efficient, and second the Semeval-07 dataset only consists in short news headline while the Semeval-13 dataset is composed of tweets which are much longer and then a better source for training. We assume that the difference in text length could also explain the difference in the results for the symbolic approach (56% vs 65.86%) since it is known that text length can influence the performance of sentiment analysis engines. A less difficult cross-domain evaluation would then rely on datasets that share the same properties in terms of text length and available data size.

3.2 Applications

Video conference feedback One problem of video conference is the lack of feedback that the presenter can have about its remote audience. The first application of our sentiment/emotion analysis tool consists in providing the presenter an aggregated feedback of the emotional state of its audience. We assume that the audience is both attending to the videoconference remotely and expressing its feelings over a textual channel, using Twitter, Facebook or by chat. Moreover in the context of the Empathic Products project, the audience video feedback is also analyzed with regards to visual emotions. The interoperability solution based on EmotionML proves to be an efficient option for combining the two kinds of feedbacks. The most generic emotional output is the binary valence which offers a basic yet more reliable characterization of the affective state of the audience. Emotions are also possible, but depending on the video domain (e-learning, news, etc.), Ekman's emotions may not be fully relevant. The sentiment/emotion of all messages sent by audience participants is averaged; when using valence it is possible to animate a gauge, when using emotions, it is possible to animate iconic emotion representations (fig. 2)

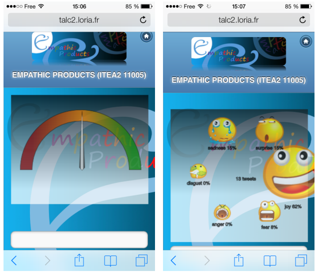


Fig. 2. Display of valence and emotions on mobile device

E-learning virtual world emotion tagging E-learning in a virtual world requires some level of investment by the participants and is eased by their collaboration. It has been shown that emotional information can enhance the collaboration. For instance [10] show that participants interact more if provided with emotional clues about their partner's current state. We propose to integrate our sentiment analysis tool to the Umniverse collaborative virtual world [19]: the virtual world works as a situated forum in which participants can move around and submit posts. When participants submit posts they can annotate by hand the emotion that their post carries (with Ekman's emotions). Our tool can then be used to pre-annotate each post by proposing automatically an emotion. After posts have been annotated and published to the forum, it is possible to filter the existing posts by their annotated emotion and as such find all the posts that carry sadness for example.

Global opinion of TV viewers An early application for sentiment analysis has been the annotation of movie reviews in order to automatically infer the sentiment of viewers towards a movie [16]. We propose to apply the same idea to the TV shows. It is known that regular TV shows have Twitter fan-base who discusses the show. The idea is thus to conduct sentiment/emotion analysis on Twitter streams that are related to a particular TV show. The ongoing work related to that application is thus inline with recent work in sentiment analysis and emotion detection in Twitter, see for instance [14].

4 Conclusion

When developing a sentiment/emotion analysis service that is meant to be generic enough to work with several different applications, it is important to consider whether the algorithms are tied to a particular domain, whether the representation of output emotions is homogenous for all applications and whether the algorithms may be adapted to other languages than English. We detailed these three problems while mentioning the existing solutions to them. We introduced our own sentiment/emotion analysis service developed in the context of the Empathic Products ITEA project which partially adresses these problems.

While interoperability seems satisfactory enough and multilinguality support has already been shown to be robust when using machine translation, the domain dependence aspects could be improved. In particular we evaluated the algorithms on two quite different domains, the news headlines provided by the Semeval-07 affective task evaluation and the tweets provided by the Semeval-13 sentiment analysis in Twitter evaluation. The results show significant difference, probably caused by the difference of text length between the two types of dataset. Nevertheless, for future work we are considering approaches that are more hybrid such as [1] in order to tackle domain dependence.

This work was conducted in the context of the ITEA2 "Empathic Products" project, ITEA2 1105, and is supported by funding from the French Services, Industry and Competitivity General Direction.

References

1. Andreevskaia, A., Bergler, S. 2008. When Specialists and Generalists Work Together: Overcoming Domain Dependence in Sentiment Tagging. In Proceedings of ACL-2008: HLT, pp 290-298. Columbus, USA. 2008.
2. Aue, A., Gamon, M. 2005. Customizing Sentiment Classifiers to New Domains: a Case Study. In Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP-05). 2005.
3. Baker, R., D'Mello, S., Rodrigo, M.M., Graesser, A.C. 2010. Better to be frustrated than bored: the incidence, persistence, and impact of learner's cognitive-affective states during interactions with three different computer-based learning environments. In International Journal of Human-Computer Studies, Volume 68, Issue 4, April, 2010, pp 223-241.
4. Balahur, A., Turchi, M. 2012. Multilingual Sentiment Analysis using Machine Translation? In Proceedings of the 3rd Workshop in Computational Approaches to Subjectivity and Sentiment Analysis. Korea. 2012.
5. Banea, C., Mihalcea, R., Wiebe, J. 2011. Multilingual Sentiment and Subjectivity Analysis. In Multilingual Natural Language Processing, editors Imed Zitouni and Dan Bikel, Prentice Hall, 2011.
6. Chaumartin, F-R. 2007. UPAR7: A Knowledge-Based System for Headline Sentiment Tagging. In Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval 2007), pp. 422-425, 2007.
7. Ekman, P. 1972. Universals And Cultural Differences In Facial Expressions Of Emotions.. In J. Cole (Ed.), Nebraska Symposium on Motivation, 1971 (Vol. 19, pp. 207-282). Lincoln: University of Nebraska Press.
8. Esuli, A., Sebastiani, F. 2006. SentiWordNet: A Publicly Available Lexical Resource for Opinion Mining. In Proceedings of the 5th Conference on Language Resources and Evaluation (LREC 2006), Genova, IT, 2006, pp. 417-422.
9. Hu, M., Liu, B. 2004. Mining and summarizing customer reviews. In KDD-2004.
10. Kamada M., Ambe M., Hata K.,Yamada E.,Fujimura Y. 2005. The Effect of the Emotion-related Channel in 3D Virtual Communication Environments. PsychNology Journal, 3(3), 312-327.
11. Kim, S., Hovy, E. 2006. Identifying and Analyzing Judgment Opinions. In Proceedings of the main conference on Human Language Technology Conference of the

North American Chapter of the Association of Computational Linguistics (HLT-NAACL '06), pp. 200-207.

12. Kort, B., Reilly, R. Picard, R.W. 2001. An Affective Model of Interplay Between Emotions and Learning: Reengineering Pedagogy Building a Learning Companion. In Proceedings of IEEE International Conference on Advanced Learning Technologies. Madison.
13. Liu, B. 2012. Sentiment Analysis and Opinion Mining. Morgan & Claypool Publishers, May 2012.
14. Mohammad, S. 2012. #Emotional Tweets. In Proceedings of the First Joint Conference on Lexical and Computational Semantics (*Sem), June 2012, Montreal, Canada.
15. Nakagawa, T., Inui, K., Kurohashi, S. 2010. Dependency Tree-based Sentiment Classification using CRFs with Hidden Variables. In Proceedings of Human Language Technologies: the 2010 Annual Conference of the North American Chapter of the ACL (HLT 2010), pp. 786-794. Los Angeles. 2010.
16. Pang, B., Lee, L., Vaithyanathan, S. 2002. Thumbs up? Sentiment classication using machine learning techniques. In Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP 2002), pp. 7986, Philadelphia, PA.
17. Polanyi, L., Zaenen, A. 2004. Contextual valence shifters. In J. Shanahan, Y. Qu, and J. Wiebe (eds.), Computing Attitude and Affect in Text: Theory and Applications, pp. 19. The Information Retrieval Series, Vol. 20, Springer, Dordrecht, The Netherlands.
18. Read, J., Caroll, J. 2009. Weakly Supervised Techniques for Domain-Independent Sentiment Classification. In Proceeding of the 1st international CIKM workshop on Topic-sentiment analysis for mass opinion (2009), pp. 45-52. 2009.
19. Reis, F., Malheiro, R.. 2011. Umniversity virtual world platform for massive open online courses University Platform. IE Challenges 2011 VII International Conference on ICT in Education, Braga, Portugal, May 2011.
20. Schröder, M., Baggia, P., Burkhardt, F., Pelachaud, C., Peter, C., Zovato, E. 2011. EmotionML – an upcoming standard for representing emotions and related states. In Proceedings of the 4th International Conference on Affective Computing and Intelligent Interaction - Volume Part I (pp. 316-325).
21. Socher, R., Pennington, J., Huang, E.H., Ng, A.Y., Manning, C.D. 2011. Semi-Supervised Recursive Autoencoders for Predicting Sentiment Distributions. In Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing (EMNLP 2011), pp 151-161. Edinburgh, 2011.
22. Strapparava, C., Valitutti, A. 2004. WordNet-Affect: an affective extension of WordNet. In Proceedings of the Language Resources and Evaluation Conference (LREC 2004), Lisbon, May 2004, pp. 1083-1086.
23. Strapparava, C., Mihalcea, R. 2007. SemEval-2007 Task 14: Affective Text. In Proceedings of the 4th International Workshop on the Semantic Evaluations (SemEval 2007), Prague, Czech Republic, June 2007.
24. Turney, P. 2002. Thumbs up or thumbs down? Semantic orientation applied to unsupervised classication of reviews. In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL 2002), pp. 417424, Philadelphia, PA.
25. Wilson, T., Wiebe, J., Hoffmann, P. 2009. Recognizing Contextual Polarity: An Exploration of Features for Phrase-Level Sentiment Analysis. Computational Linguistics, Vol. 35, No. 3, pp 399-433. 2009.

Automatic Speech for Poetry - The Voice behind the Experience

Diana Arellano[1], Cristina Manresa-Yee[2], and Volker Helzle[1]

[1] Filmakademie Baden-Wuerttemberg, Institute of Animation, Germany
[2] Mathematics and Computer Science Dept., University of Balearic Islands, Spain
{diana.arellano,volker.helzle}@filmakademie.de
cristina.manresa@uib.es
http://research.animationsinstitut.de/

Abstract. This paper presents the advances in expressive speech for the interactive installation The Muses of Poetry, as well as the results of an evaluation to assess the interaction and emotional experience of the participants. The objective of the installation is to bring poetry closer to a wider audience through the use of animated characters who not only read poetry, but also manifest the emotional content of the poems. The latter is done using facial expressions and affective speech, which is the focus of this paper. The results of the evaluation show that in general the installation was pleasant and appealing. Also, the three characters managed to convey emotions and awake emotions in the users.

Keywords: Affective Speech, Computational Creativity, User Experience

1 Introduction

Poetry is a form of literary art that can be characterized by its capacity to transport the reader into another, more ethereal world. However, as Kwiatek and Woolner[1] expressed it, *poetry is not always easy to understand, especially for young people.*

In our attempt to bring poetry closer to a wider audience, we developed an interactive installation named The Muses of Poetry, where virtual animated characters recite poetry in an emotional way. A semantic affective analysis of the text of existing poems allows the system to extract their intrinsic affective content, which is manifested by the characters through facial expressions and expressive speech, both automatically generated in real-time.

Unlike previous works that have dealt with poetry and virtual characters [2], automatic generation of poetry with emotional content [3], or without it [4], [5], [6], the work we present combines different fields like real-time computer animation, semantic analysis, human-computer interaction and affective computing, in order to create a believable and engaging expression of the emotions in the poems. We think that The Muses of Poetry might help users not familiar with poetry to understand better the context of the poem, thus awakening an interest in this art. Moreover, given the real-time feature of our installation, any poem can be analysed, integrated and recited on-site, without the need to spend a large amount of time and resources.

The objective of this paper is twofold. On the one hand, we explain the procedure to enrich the synthetic speech of the characters, or muses with the emotionality of the poems. On the other, we present the results of a user-experience evaluation performed on The Muses of Poetry during a public exhibition.

2 Emotional Speech

One of the characteristics of poetry is its freedom of interpretation, which can lead to different ways of reading it aloud. Pauses, intonation, melody, and emotions are some of the elements that need to be taken into account when reciting poetry. Their correct use can enhance the poetic experience to a level that is capable to engage a wider audience.

In The Muses of Poetry one of our objectives is to transmit the emotionality of the poems not only with visual manifestations, but also with changes in the speech. To that end, a semantic analysis of the text of the poems is performed in order to extract their affective content. From this analysis, both general and "line-per-line", emotional states are obtained, which in the case of The Muses are: pleasant, nice, fun, unpleasant, nasty and sad. As for the poems, these have been provided by real poets from the Cordite Poetry Review Magazine [7], and by poets who have their poems under the Creative Commons licence.

Once the analysis of the text is carried on, the results are integrated into the system, so they can be interpreted by the TTS tool. In our installation we are using the third-party voice synthesizer provided by SVOX[3].

As it is, the TTS produces voices of relative good quality, but with no changes in the prosody. For that reason, we developed a real-time algorithm that indicates, without bias from the human perception, where exactly the changes in the pitch and speed of the voice have to be generated. However, due to the static nature of the poems, this prosodical analysis is performed only once and stored as tags inside the text of the poem.

2.1 Poem into Lines

The structure of the text of the poems follows an XML structure, which facilitates the extraction of the different parts of the poem: author, title and body; as well as the tagging of the emotions.

Before getting into details, it is worth noting that a "line" in the poem is not necessarily the same as the written line (i.e. separated by new lines in the text). We defined a "poetic line" as the set of words, or lines, that enclose one idea of the poem. To divide the text into lines we follow the logic presented in this pseudo-algorithm:

[3] http://www.nuance.de/products/SVOX/index.htm

```
begin
   repeat
     Check if the LINE ends with a period ‘.’
        if TRUE then
           add it to the list of LINES
           go back to repeat
        else
           if NUMBER_LINES_READ > 3 then
              check for conjunctions in LINE, add a comma before them
              add it to the list of LINES
              go back to repeat
           if LINE ends with a comma ‘,’ then
              go back to repeat
           else if First Letter in NEXT_LINE is capitalized then
              random:
              add it to the list of LINES, OR
              add a long pause ‘..’ at the end of LINE
              go back to repeat
   until endOfFile
end
```

As a result, each time the system encounters a new line or punctuation mark in this new set of lines (`LINES`), it will be interpreted as a pause. The length of the pause depends on the mark, if it is a new line or two points ‘..’, then it is a long pause; on the contrary, if it is a comma ‘,’ or a period ‘.’, then it is a short pause (comma pause is in general shorter than the period pause).

2.2 Tagging the Poem

Once the poem is divided into lines, the semantic affective analysis to extract the emotional states is carried on. The details of the analysis can be found in [8].

As previously mentioned, the result of the global affective analysis of the poem gives one of the following emotional states: pleasant, nice, fun, sad, unpleasant, or nasty. In order to simplify the tagging of the poems, we re-grouped these states in: *happy* (i.e. pleasant, nice, or fun), *unpleasant* (i.e. unpleasant or nasty), and *sad.*

The tagging of a poem is performed line by line, and in the particular case of pitch, also per word. The values we use for speed and pitch are obtained from the global and line-per-line analysis of the poem, according to the following rules:

1. Compute the increment, or decrement, of the neutral pitch and speed, according to the poem emotional state (i.e. if it is happy, sad, or unpleasant). In the case of a human-like character, we use the following values:
   - If the state is *happy*, the neutral speed (i.e. neutral speed=90) is incremented by 5%, and the neutral pitch (i.e. neutral pitch=90) by 10%.
   - For an *unpleasant* state, the decrement in neutral pitch and speed is 8%.
   - For the *sad* state, the decrement in neutral pitch and speed is 5%.

2. Compute for each "poetic line" its resultant emotional value by multiplying (a) the poem emotional state value obtained from the affective analysis using the Whissell Dictionary of Affect in Language [9] (e.g. pleasant = 11.45) by (b) the line emotional state value, which is basically the number of words with the same emotional state of the line (e.g. if the line is pleasant and has 4 words rated as pleasant, then the line state value will be 4)
3. To the speed increased in step 1, add (or subtract, depending on the emotional state) the emotional value obtained in step 2. This value will be used to tag the speed of the whole line.
4. The pitch value to tag the whole line is obtained from step 1. Moreover, for each line, the value obtained in step 2 is added (or subtracted, depending on the emotional state) to the neutral pitch. This value is used to tag the pitch of each word with emotional state similar to the one of the line. If no words are rated with the same state as the line, then no pitch changes are applied.

The subtle variations in step 1 are due to the fact that abrupt changes in a realistic character are seen as very uncanny, diminishing the whole experience. However, this changes in the case of more cartoon or abstract characters, where the common guidelines followed in animation include the idea that cartoon characters should be exaggerated to better convey emotion and intent [10].

The tags we used are provided by the TTS tool, SVOX to change the pitch and speed: `[synthesis:pitch level=PVALUE]`, where `PVALUE` $\in (0, 200)$ and `[synthesis:speed level=SVALUE]`, where `SVALUE` $\in (0, 500)$.

The following excerpt are the "poetic lines" of the poem *For the Road*, by Carol Jenkins. The poem was assessed as *happy* and tagged accordingly:

Line 1

`[synthesis:emotion id='JOY_TRIGGER'][synthesis:pitch level='99'][synthesis:speed level='100']` *First as a dare and then for the* `[synthesis:pitch level= '124']` *warm* `[/synthesis:pitch]` *languor of the tar, at midnight* `[synthesis:pitch level= '124']` *walking* `[/synthesis:pitch]` *to my house, we lay down our bodies on the middle of Moana Road and* `[synthesis:pitch level='124']` *kissed,* `[/synthesis:pitch]` *.. Those long dreamy kisses of abandonment, to each other, to the road, to the dark pines looking on, to the locked light of houses with blinds drawn tight on quarter acre blocks,* `[/synthesis:speed] [/synthesis:pitch ]`

Line 2

`[synthesis:emotion id='PLEASANT_TRIGGER'][synthesis:pitch level='99'] [synthesis:speed level='100']` *The stars' bright and dizzy mass arcing over us, and we'd get to our feet,* `[synthesis:pitch level='101']` *like* `[/synthesis:pitch]` *angels coming to in a strange world,* `[/synthesis:speed] [/synthesis:pitch]`

Line 3

`[synthesis:emotion id='JOY_TRIGGER'][synthesis:pitch level='99'][synthesis:speed level='100']` *To walk down the road, arms and hands tangling,* `[synthesis:pitch level='124']` *laughing, like* `[/synthesis:pitch]` *we'd swallowed a* `[synthesis:pitch level='124']` *universe* `[/synthesis:pitch]` *and it was exploding out of our fingertips.* `[/synthesis:speed] [/synthesis:pitch]`

It can be seen that at the beginning of each line, the pitch and speed have the same value (pitch = 99, speed = 100). Nonetheless, the pitch-tagged words (e.g. *warm*, *walking*, and *kissed*) have different values (*Line 1 y 3*, pitch = 124; *Line 2*, pitch = 101), obtained from the rules previously explained.

3 User Experience

To assess the interaction and the emotional engagement produced by The Muses of Poetry, we carried on an user evaluation during its exhibition at FMX 2013, the Conference on Animation, Effects, Games and Transmedia. This conference reunites professionals, students and amateurs focused in the development, production and distribution of digital media, as well as computer graphics and animation, all of whom constituted the main audience who interacted with The Muses of Poetry.

Regarding the installation, it was a stand resembling an open book, where the slides that simulated the pages formed a kind of cave, where the user could enter and interact with the characters. The interaction was thought to be free-of-devices, for which we installed a clip microphone on the top of the second slice, almost invisible to the human eye, and approximately 1 meter away from the projection screen.

As for the virtual characters, we designed and implemented three types that could cover the wide spectrum of animated characters. The first one was a realistic human-like female, Nikita, which we have used in previous research applications. In this case, we added a veil to make her look more ancient-Greek like. The second was designed by one of the students at the Institute of Animation, following the premise of a more abstract character. This was made of particles that are shaped in the form of a human head, and were constantly moving when the character spoke. The third one, Myself, was a 2D cartoon character designed by the German animator Andreas Hykade, which came to complement this first repertoire of muses. Figure 1 shows the installation and the evaluated characters.

Participants experienced the installation alone with the support of a member of the development team, who explained how the system worked. Each participant interacted with only one character: Nikita, Particles, or Myself. At the beginning, the character asked the user to select two words from a word cloud displayed on the screen. Once the two words were recognized, a poem containing those two words was selected and recited. At the end of the experience, the participant completed a questionnaire regarding the installation, the character and his or her feelings towards the interactive system and the poem reading. The written questionnaire consisted in eight 5-point Likert scale questions ranging from "Strongly Disagree" on one end to "Strongly Agree" on the other:

Q1: The character is attractive as a poetry reader
Q2: The character conveys emotion
Q3: The character read the poem in a way I understood the topic of the poem
Q4: I think I would like to visit this installation frequently or I would recommend it to my friends

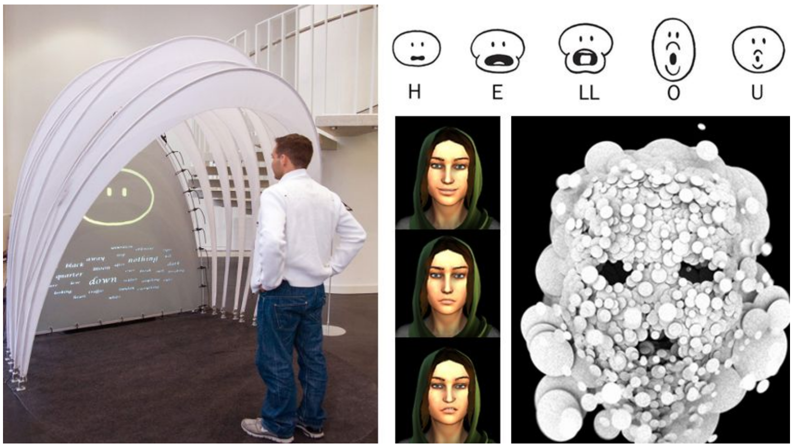


Fig. 1. Left: Installation and location of the mic. Right: (top) Myself, 2D character; (down-left) Nikita, 3D realistic character; (down-right) Particles, 3D abstract character

Q5: I felt a variety of emotions while listening to the poems
Q6: The system is pleasant
Q7: The system is inviting
Q8: The system is appealing

In the end 51 questionnaires were gathered: 35 from male participants and 16 from female participants, with ages ranging between 19 and 45 years (average age was 27). Data was analysed for each specific question using two approaches: separately for each character to find differences among the three poetry readers and together to conclude global insights from the interactive system.

Figure 2 shows the boxplots for each of the eight questions, considering the results of the three characters together. Based on the interquartile ranges (IQR) of questions Q1, Q2 and Q5, all related with the emotional aspect of the installation, we could not conclude if the user felt emotions when listening to the poems, or if he or she felt the characters emotional enough, because the bars are widely distributed. However, Q3, which evaluated the degree of empathy with poetry, throws *median*=4, which indicates that one of the objectives of the installation (make a wider audience understand poetry) might have been achieved. As for questions Q6, Q7 and Q8, despite the outliers, we can conclude that the installation was pleasant, inviting and appealing.

Figure 3 shows the boxplots for each of the eight questions, for each character. The boxplot on the right of figure 3 shows the IRQs when evaluating Nikita. For Q1, although the median value was above the mid point, it cannot be concluded that Nikita was attractive as a poet reader. The analysis of Q2 and Q3 do not throw decisive results either. However, people felt they understood the topic of the poem (Q4) and felt the installation more inviting (Q7) and appealing (Q8) than pleasant (Q9). From the boxplot corresponding to Particles we can conclude

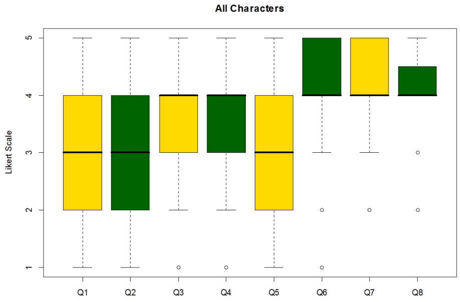


Fig. 2. Inter-quartile analysis for each question and all characters

that people indeed understood the content of the poems (*median*=4), although 50% of the sample is under this value. As for the appealing, pleasantness and inviting nature of the installation, results are not so satisfactory as with Nikita and Myself. Nonetheless, these elements were rated as positive. Finally, from the boxplot with the evaluation of Myself, we cannot conclude that the character was attractive as a poet reader (Q1), but it did conveyed emotions (Q2) and 50% of the sample felt emotions while listening to the poems. Participants also felt that with this character the installation was pleasant, inviting and appealing. A last thing to note is that with Myself users would recommend the installation to friends, in a higher scale than with the other characters.

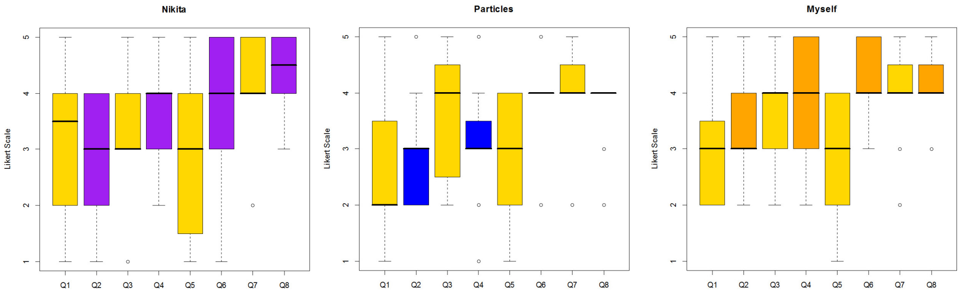


Fig. 3. Inter-quartile analysis for each question and each character: Nikita (left), Particles (middle), Myself (right)

A last analysis focused on the mode and median differences based on genders showed that women found the interactive installation more inviting and pleasant than men, and in general they rated higher values than men in all question, as seen in Table 1. In a more detailed analysis, women also felt more likeliness for the Myself character, while men for Nikita. Regarding the mode, it was observed that women felt more emotions while listening to the poem.

Table 1. Data of participants and evaluated characters

	Q1	**Q2**	**Q3**	**Q4**	**Q5**	**Q6**	**Q7**	**Q8**
median (women)	3	3	4	4	3	4.5	5	4
median (men)	3	3	3	4	3	4	4	4
mode (women)	3	3	4	5	4	5	5	4
moda (men)	2	3	3	4	2	4	4	4

4 Discussion and Future Work

We presented the advances in affective speech generation and interaction results of the on-going project The Muses of Poetry, an interactive installation where animated characters recite poetry in an emotional way. Regarding speech, the generated affective voices were in general satisfactory, enhancing the conveyance of emotions. The results of the user experience evaluation showed that the majority of the participants found the installation pleasant, inviting and appealing. However, these results did not allow us to conclude which character performed better as a poetry reader. In the future we will continue working on the speech generation to produce a more natural intonation, we will add more emotional expressions in the current characters, as well as new animated characters.

Acknowledgments. The project is funded by the *Innovationsfonds Kunst* of the Ministry for Science, Research and Art in Baden-Württemberg (54-7902.513/65)

References

1. Kwiatek, K., Woolner, M:. Let me understand the poetry. Embedding interactive storytelling within panoramic virtual environments. In: EVA 2010, pp. 199–205 (2010)
2. Tosa, N.: Interactive poem. In: ACM SIGGRAPH 98 Conference abstracts and applications, SIGGRAPH 98, pp. 300 (1998)
3. Kirke, A., Miranda, E.: Emotional and Multi-agent Systems in Computer-aided Writing and Poetry. In: Symposium on Artificial Intelligence and Poetry, pp. 17–22 (2013)
4. Colton, S., Goodwin, J., Veale, T.: Full-FACE Poetry Generation. In: Proceedings of the 3rd International Conference on Computational Creativity, pp.95–102 (2012)
5. Gervás, P., Hervás, R., Robinson, J. R.: Difficulties and Challenges in Automatic Poem Generation: Five Years of Research at UCM. In: E-Poetry 2007 (2007)
6. Cope, D.: Comes the Fiery Night. NY: Amazon Books (2011)
7. Cordite Poetry Review Magazine, `http://cordite.org.au/`
8. Arellano, D., Spielmann, S., Helzle, V.: The Muses of Poetry - In search of the poetic experience. In: Symposium on Artificial Intelligence and Poetry, pp. 6–10 (2013)
9. Duhamel, P., Whissell, C.: The dictionary of affect in language [software] (1998)
10. Hyde, J., Carter, E. J., Kiesler, S., Hodgins, J. K.: Perceptual effects of damped and exaggerated facial motion in animated characters. IEEE International Conference on Automatic Face and Gesture Recognition - FG 13 (2013)

Analyzing for emotional arousal in HMD-based head movements during a virtual emergency

C. Becker-Asano[1], D. Sun[1], C. N. Scheel[2], B. Tuschen-Caffier[2], and B. Nebel[1]

[1] Institut für Informatik, Georges-Köhler-Allee 52, 79110 Freiburg, Germany
{basano,sun,nebel}@informatik.uni-freiburg.de
[2] Institut für Psychologie, Engelbergerstraße 41, 79106 Freiburg, Germany
{corinna.scheel,brunna.tuschen-caffier}@psychologie.uni-freiburg.de

Abstract. This paper reports on results of a statistical analysis of human players' head-movements. Forty-one participants were asked to cope with an unexpected emergency in a virtual parking lot. Before the virtual reality exposure began, half of the participants watched an emotion-inducing movie clip and the other half an emotionally neutral one. The analysis of the acquired questionnaire data reveals, however, that this emotion induction method seems to have been rather ineffective. Thus, it is not surprising that only very weak between group effects are found when analyzing for differences in head movements around the emergency event. In general, horizontal head movement speed is found to be on average significantly faster during the first fifteen seconds directly after the emergency event as compared to just before and another fifteen seconds later. These findings are in line with previous results of an analysis of the acquired physiological data, further substantiating the conclusions drawn.

1 Introduction and motivation

Recently, the visual quality of virtual characters in computer games reached such a high level that they are able to convey a wide range of emotions very convincingly by body posture and facial expressions. In addition, the visual effects of such interactive games are now comparable to those of cinematic productions. The Affective Computing community can benefit from this high realism in that new means to acquire data on emotions during interaction are realizable. The recent availability of affordable head-mounted displays with in-built inertial measurement units (IMU) (e.g. Oculus Rift) not only enables novel gaming experiences for the consumer market, but the acquired head movement data might also be useful to recognize emotions during gameplay.

Accordingly, we aim to develop and test means to detect emotional arousal online based on the available head movement data. In contrast to the rather slowly changing physiological attributes, such as heart rate or skin conductance level, a participant's head movement can be expected to respond rather quickly to emotion arousing or stressful events. In addition, even a prevailing background emotion or general increase in arousal could lead to a significant change in head movements. This paper reports on our first attempt to extract and analyze such features from empirical data that were collected during an interdisciplinary collaboration between computer scientists and psychologists.

The remainder of this paper is structured as follows: Section 2 presents and discusses related work, before in Section 3 the research goal and the technological background are explained. Section 4 details the procedures taken in the study and, in Section 5 its results are given. At last, general conclusions are drawn.

2 Related work

Virtual Reality (VR) technology has been used, for example, to train surgeons [1], to treat posttraumatic stress disorder (PTSD) in veterans of the Iraq war [2, 3], or as a means to evaluate emotion simulation architectures driving virtual humans [4, 5]. Furthermore, a VR setup has evoked similar responses to violent incidents in human observers as can be expected in real world situations [6] given that a high degree of plausibility could be achieved and maintained. The software used to realize these applications range from game engines such as Epic's Unreal Tournament [2, 3, 5] to a number of custom made installations [7, 4] with proprietary software components. They are combined with different display technologies such as panoramic, auto-stereoscopic, or head-mounted displays (HMDs), or even CAVEs (CAVE Automated Virtual Environment, [8]). The VR-related aspects of a project for PTSD treatment are meant to teach the patient "coping skills" [9] through virtual exposure. For an empirical study on the link between presence and emotions Riva and colleagues [10] used an HMD with head-tracking and a joystick for navigation. They successfully induced an anxious mood in participants only by systematically changing visual and auditory components of a virtual park scenery.

Already more than ten years ago Cowie et al. [11] expected entertainment to be one of the applications for computational emotion recognition. Their overview, however, does not include any work on emotion recognition based on body or head movements. Later, head tilt frequency was used as a parameter to detect head nods in the context of a fatigue detection system [12]. In the human-computer dialog context the importance of head movements is generally acknowledged [13, 14] and a system for automatic detection of the mental states "agreeing, concentrating, disagreeing, interested, thinking and unsure" [15] from video streams consequently includes a number of head orientations. To the best of our knowledge, however, mechanisms to derive emotion-related parameters from head movements during computer games have not been investigated.

3 Experiment outline

3.1 Research goal

In general, we aim to develop novel technological means to detect emotional arousal of humans while they are interactively exposed to potentially dangerous

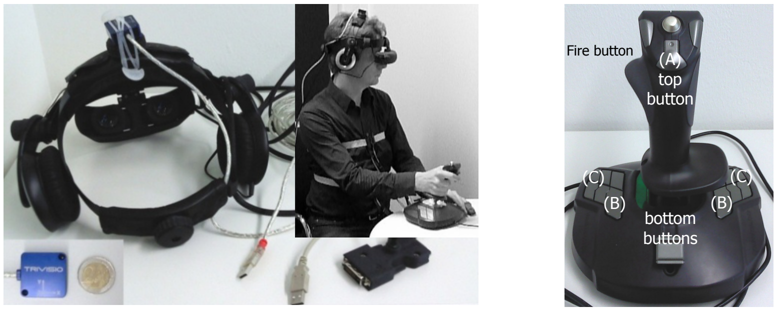


Fig. 1: Technical setup with the head-mounted display and the Colibri IMU (left) and the joystick's button allocation (right)

events. The technological hard- and software setup (see Section 3.2) has already been shown to be similarly emotion arousing as watching a short clip of a horror movie [16].

Here we want to explore, if increased stress or fear levels affect a players head movements. Thus, we set out to analyze the acquired head movement data and relate the result to previously analyzed physiological data. Two research questions summarizes our concerns:

1. RQ_1: Did our emotion induction method have the desired effect of inducing fear and stress and, if so, to what extent?
2. RQ_2: Does a sudden, emotion eliciting event during the VR exposure significantly change a player's head movement speed and do previously induced emotions affect these movements as well?

The first research questions is addressed by performing a between-groups, repeated measures analysis of the questionnaire data. In order to address the second research question within-subject, repeated measures analyses of variance (ANOVAs) of four segments of head movement data around a decisive moment during the experimental session is conducted. The complete study design is described in Section 4, after the hard- and software setup has been explained next.

3.2 Technology

Hardware setup To achieve an immersive setup we opted for Trivisio's "VRvision HMD" [17], which features two SVGA AMLCD 800x600 color displays with 24 bit color depth, 60 Hz video frame rate, and a field of view of 42° diagonally and 25° vertically; cp. Fig. 1, left. The USB-powered HMD features a pair of Sennheiser HD 205 headphones, which are connected to the same PC. An ATI Sapphire Radeon 5870 together with an Intel Core-i5-760 CPU drives the HMD under Windows 7 (64bit). A USB-powered "Colibri" tracker—mounted on top of the HMD (cp. Fig. 1, left)—provides us with the participant's head movements.

The participants used a Thrustmaster T-16000M joystick to navigate inside the virtual environment; cp. Fig. 1, right. They can move forward and backward

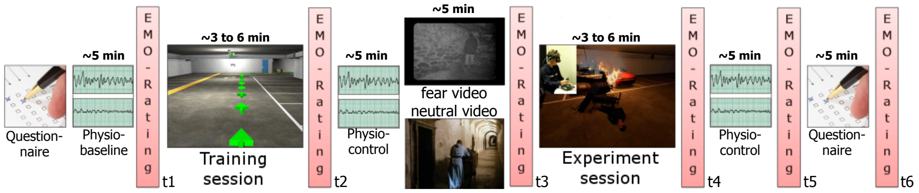


Fig. 2: Overall design of the empirical study

by pushing and pulling the joystick, respectively. Leaning the joystick left or right results in sidesteps, whereas turning it slightly to the left or right makes the participant turn accordingly. Finally, as long as "button B" is being pressed the participant's character is "crouching". "Button A" enables participants to take an object or a tool, the latter being either a spray can during training or a fire extinguisher during the experimental session. While holding an object, pressing "button A" again results in dropping it. With just a tool (or nothing at all) in his or her virtual hands the participant opens and closes doors or pushes buttons by pressing "button A." By pressing the joystick's "fire button," the participants throw an object or use a tool. During the training sessions, for example, they are instructed to practice using a tool by coloring a wall with a spray can. Subsequently, they have to throw it away pressing "base button C."

The setup of the physiological sensors for measuring skin conductivity, heart rate variability, and breath rate are detailed elsewhere [18, 16]. Both sessions of the experiment took place in a darkened room with only the joystick emitting some light for reference. A desktop monitor was used for online questionnaire assessment before, between, and after the experimental sessions.

Software setup Valve's Source Engine as was chosen as a software framework following similar work by Smith & Trenholme [19]. In addition to their simulation system, we also modified the source code of the Source 2007 engine to include tools such as a spray can and a fire extinguisher and to implement mechanisms for synchronizing the in-game events with the external sensor recordings for later analysis. In addition, we designed an underground parking lot from scratch that features signs, doors, stairways, elevators, cars, and additional models such as a coke vendor machine to make it look most convincingly[3].

4 Study design

The overall design of our study can be split up into five parts (cp. Fig. 2). First, socio-demographical and psychometric data as well as previous experience with computer games and VR technology are acquired through questionnaires.

[3] Videos of the final setup can be found here: `https://www.becker-asano.de/index.php/research/videos/49-videos1#COVE`

Physiological baseline data is recorded for five minutes, followed by a first rating of felt emotions. Then, participants are guided through a training session. A second rating of felt emotions is acquired afterwards; cp. Fig. 2, t_2. After a control of the physiological measurement the participant either watches a neutral video clip (control) or a fear inducing video clip (experimental manipulation). Both are around five minutes long and the latter is a clip taken from the movie “Blair Witch Project.” Then, at t_3, the participants rate their feelings again. The experimental session starts with the participant standing in front of the elevator on the ground, see Section 4.1. After the VR experiment, at t_4, the participant has to rate his or her felt emotions again. Finally, the physiological measurement is being controlled again, after which the participants have to rate their emotions once more (t_5). After a final questionnaire they are asked to report one last time on their felt emotions (t_6).

The training sessions start on underground level five of the parking lot and are acoustically guided both to get used to the control interfaces as well as to the situation they are supposed to deal with.

4.1 Experimental session

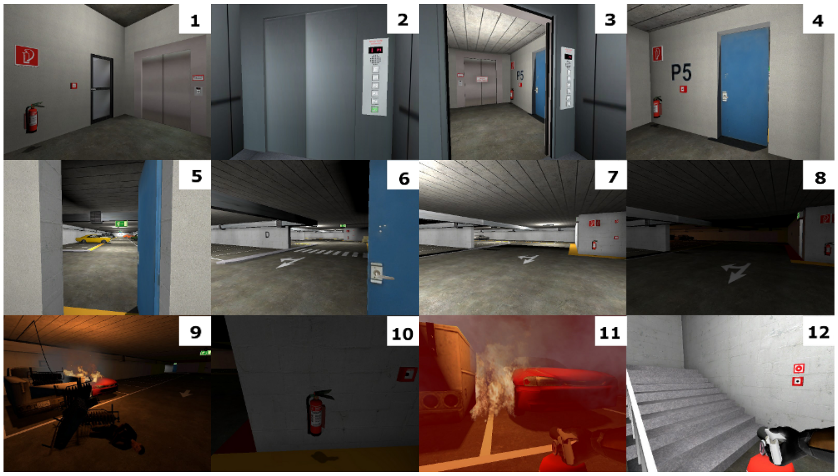


Fig. 3: The experiment session between t_3 and t_4 (see main text for explanations)

The experimental session starts with the participant on the ground floor inside the same virtual parking lot as the one used for training. The participants are instructed to go down by the elevator back to their red sports car and drive it out of the parking lot. The individual way of reacting to challenging situations might show differences in the participant’s emotional skills. Therefore the participants had no further instructions but to react adequately in any situation they might get into. The most appropriate way to deal with the sudden explosion (cp. Fig. 3, #7–8) is to approach the sports car (#9) with the injured person in front of the fire, then, to get back to press the alarm button and to take a fire extinguisher (#10) to extinguish the fire (#11). Finally, it is best to exit the parking lot taking the stairs (#12).

5 Experiment results

The outlined experiment was approved by the University's ethics committee. A total of 48 university students participated in the study after they had provided informed consent. Seven of them had to be excluded due to technical errors and/or missing data. The remaining 41 participants (age: $M = 23.4$ years, $SD = 3.1$ years, 18 male, 23 female) were randomly assigned to the experimental conditions, with 20 watching the neutral movie clip ("control condition", 5 male), and 21 the fear inducing movie clip ("fear condition", 13 male).

5.1 Procedure and previous results

We concentrate our analysis on two different data sets, first, the emotion ratings, which were acquired through a visual analogue scale ranging from zero to ten, and, second, the head movement data. Ratings of felt intensity for the emotions fear, anger, shame, sadness, happiness, boredom, guilt, and stress were gathered a total of six times during the course of the experiment; cp. Fig.2. Only the ratings for fear and stress that followed the movie-based emotion induction, i.e. t_3 through t_6, are included in the analysis, because before the emotion induction at t_3 no between groups difference can be expected.

A previous analysis of the physiological data of 20 participants, all of whom belonging to the control condition, showed that heart rate (HR) and skin conductance level (SCL) varied significantly [18]. The mean values of both physiological parameters were higher during the training session (SCL: $M = 8.74$, $SD = 1.83$; HR: $M = 76.21$, $SD = 11.12$), than during the neutral movie (SCL: $M = 8.09$, $SD = 1.19$; HR: $M = 74.54$, $SD = 11.88$), and highest during the minute following the sudden explosion in the experimental session (SCL: $M = 9.16$, $SD = 1.82$; HR: $M = 88.47$, $SD = 13.44$). Thus, the general emotional arousal significantly increased during the virtual emergency as compared to both the training session and the neutral movie.

5.2 Analysis of emotion ratings (RQ_1)

Two repeated-measures ANOVAs with time (from t_3 until t_6, four levels) as within-groups factor and condition (fear versus control, two levels) as between-groups factor were performed for fear and stress.

For fear the main effect of both condition, $F(1, 152) = 1.21$, n.s., and time, $F(3, 152) = 1.84$, n.s., remained below the desired five percent level of significance. Also, no significant interaction effect was found, $F(3, 152) = 0.69$, n.s.

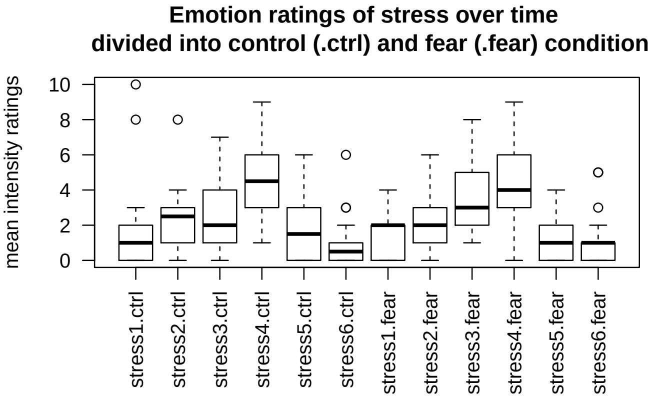


Fig. 4: The questionnaire results for "stress" compared between conditions

The repeated-measures ANOVA of the stress ratings, in contrast, showed a significant main effect for time, $F(3, 152) = 4.255$, $p < 0.01$, but, again, not for condition, $F(1, 152) = 0.37$, n.s.; cp. Fig. 4. No significant interaction effect was found, $F(3, 152) = 1.27$, n.s. A post-hoc paired t-test (bonferroni corrected) revealed a significant increase ($p < 0.01$) of stress levels from before the experimental session (stress3, $M = 3.02$, $SD = 2.04$) to just after this session (stress4, $M = 4.51$, $SD = 2.33$), and a significant decrease ($p < 0.01$) from just after the experimental session to after the third baseline (stress5, $M = 1.54$, $SD = 1.45$); cp. Fig.4.

5.3 Analysis of head movements (RQ_2)

Pitch values along the sagittal and yaw values along the horizontal plane of every participant were recorded with a sampling frequency of approx. 60 Hz during both the training and the experimental session. The yaw values are a combination of turning the joystick and looking around with the HMD, whereas the pitch values are only changing in relation to HMD (i.e. head-) movements. Accordingly, these two data streams are preprocessed and analyzed independently.

To investigate RQ_2 two consecutive 15 seconds intervals from just before (B_1 and B_2) and another two consecutive 15 seconds intervals from immediately after (A_1 and A_2) the sudden explosion during the experimental session are analyzed (cp. Fig. 3, #8). Nearly all participants were still waiting for the elevator doors to open (cp. Fig. 3, #3) 30 seconds before the explosion(cp. Fig. 3, #8). After the explosion none of the participants reached the injured person depicted in frame #9 of Fig. 3 within 30 seconds. The features extracted from the corresponding pitch and yaw data streams are subjected to two separate within-subject, repeated-measures ANOVAs.

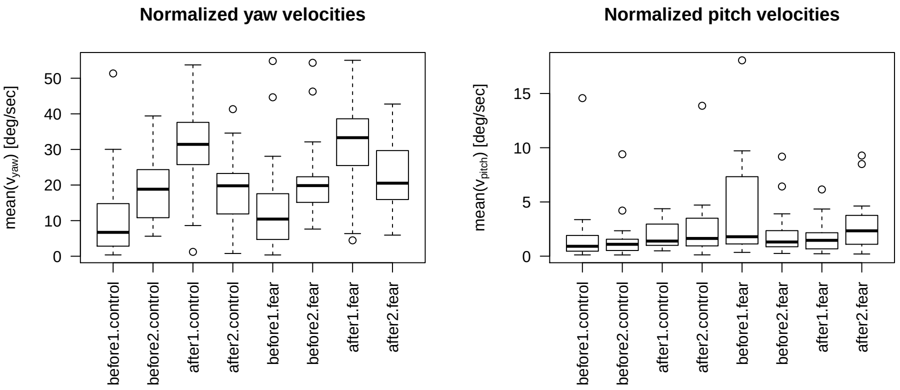


Fig. 5: Average, absolute pitch (left) and yaw (right) velocities of the 15 seconds intervals immediately before and after the explosion compared between conditions

Preprocessing and feature extraction We preprocessed the raw head-movement data as follows to remove noise:

1. A small number of consecutive values with the same timestamp due to measurement delays are deleted.
2. Turns above 180 degrees or below -180 degrees are corrected by adding or subtracting 360 degrees resp. to/from all subsequent data.
3. A low-pass butterworth filter with cut-off frequency 0.9 Hz is applied to eliminate noise.
4. Applying the primary difference quotient resulted in a sequence of velocities $v_{pitch/yaw}$ in degrees / sec.

The mean of the absolute values of pitch and yaw, respectively, are calculated as features $f_{pitch/yaw}$ per participant and data set.

Results Mean pitch and yaw velocities are plotted in Fig. 5. Although the previous analysis indicates only a weak effect of the experimental variation, for completeness the two conditions are included as between-groups factor in the following two repeated-measures ANOVAs in addition to time (four levels) as within-groups factor.

The repeated-measures ANOVA of the average pitch velocities showed no significant main effect for time, $F(3, 150) = 0.266$, n.s., but, a main effect for condition, $F(1, 150) = 4.564$, $p < 0.04$.; cp. Fig. 5, right. A post-hoc pairwise t-test, however, reveals that the difference between fear-group ($M = 2.73$, $SD = 3$) and control-group ($M = 1.96$, $SD = 2.45$) is not significant ($p > 0.07$). The interaction effect was not significant either, $F(3, 150) = 1.76$, n.s.

The repeated-measures ANOVA of the average yaw velocities, in contrast, showed a significant main effect for time, $F(3, 150) = 3.135$, $p < 0.03$. For condition, however, the main effect is not significant $F(1, 150) = 2.1$, n.s.; cp. Fig. 5, left. Again, no significant interaction effect was found, $F(3, 150) = 0.155$, n.s. A post-hoc paired t-test (bonferroni corrected) reveals a significant increase of average yaw velocity from B_1 ($M = 12.57$, $SD = 13.18$) to B_2 ($M = 19.98$, $SD = 10.75$; $p < 0.4$) and from B_2 to A_1 ($M = 31.52$, $SD = 13.26$; $p < 0.01$). Subsequently, from A_1 to A_2 ($M = 20.87$, $SD = 9.68$; $p < 0.01$) the average yaw velocity decreased significantly to a level similar to that just before the explosion occurred.

6 Conclusions

We set out to search for correlations between a human player's emotional arousal and his or her head movements while having to cope with a virtual emergency (RQ_2). In addition, we checked whether our video-based method of emotion induction was effective (RQ_1), which seemed only to be the case for stress, but not for fear.

A significantly higher average horizontal head movement speed, however, was found that might be interpreted as an immediate response to a sudden, stressful event. In the light of results derived previously from physiological data analysis, these findings suggest that increased physiological arousal might, in general, be correlated with faster horizontal head movements.

A number of challenging questions remain for future research, such as (1) do further features extracted from the acquired physiological data support the conclusions drawn with regard to RQ_2, (2) how can we better detect and account for inter-individual differences in head movement profiles, and (3) which other scenarios might be implemented to address these questions?

In summary, if emotional arousal indeed results in a change of head movement speed, then our results seem to indicate that this kind of arousal is of rather short duration. Already fifteen seconds after the unexpected events the average movement speed returned to the same level as just before the event.

Acknowledgment

This research was supported by the Freiburg Institute for Advanced Studies.

References

1. N. Seymour, "VR to OR: A Review of the Evidence that Virtual Reality Simulation Improves Operating Room Performance," *World Journal of Surgery*, vol. 32, pp. 182–188, 2008.
2. G. M. Reger and G. A. Gahm, "Virtual reality exposure therapy for active duty soldiers," *Journal of Clinical Psychology*, vol. 64, no. 8, pp. 940–946, 2008.
3. P. Kenny, T. D. Parsons, J. Gratch, and A. A. Rizzo, "Evaluation of justina: A virtual patient with PTSD," in *Intl. Conf. on Intelligent Virtual Agents*, 2008, pp. 394–408.

4. C. Becker-Asano and I. Wachsmuth, "Affective computing with primary and secondary emotions in a virtual human," *Autonomous Agents and Multi-Agent Systems*, vol. 20, no. 1, pp. 32–49, January 2010.
5. J. Gratch and S. Marsella, "Evaluating a computational model of emotion," *Autonomous Agents and Multi-Agent Systems*, vol. 11, pp. 23–43, 2005.
6. A. Rovira, D. Swapp, B. Spanlang, and M. Slater, "The use of virtual reality in the study of people's responses to violent incidents," *Front Behav Neurosci*, vol. 5, no. 0, pp. 1–10, 12 2009.
7. B. Dunkin, G. Adrales, K. Apelgren, and J. Mellinger, "Surgical simulation: a current review," *Surgical Endoscopy*, vol. 21, pp. 357–366, 2007.
8. J. Brooks, F.P., "What's real about virtual reality?" *Computer Graphics and Applications, IEEE*, vol. 19, no. 6, pp. 16 – 27, 1999.
9. G. Riva, S. Raspelli, D. Algeri, F. Pallavicini, A. Gorini, B. K. Wiederhold, and A. Gaggioli, "Interreality in practice: Bridging virtual and real worlds in the treatment of posttraumatic stress disorders," *Cyberpsychology, Behavior, and Social Networking*, vol. 13, pp. 55–65, 2010.
10. G. Riva, F. Mantovani, C. S. Capideville, A. Preziosa, F. Morganti, D. Villani, A. Gaggioli, C. Botella, and M. Alcañiz, "Affective interactions using virtual reality: The link between presence and emotions," *CyberPsychology & Behavior*, vol. 10, no. 1, pp. 45–56, 2007.
11. R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, and J. Taylor, "Emotion recognition in human-computer interaction," *Signal Processing Magazine, IEEE*, vol. 18, no. 1, pp. 32–80, 2001.
12. Q. Ji, P. Lan, and C. Looney, "A probabilistic framework for modeling and real-time monitoring human fatigue," *Systems, Man and Cybernetics, Part A: Systems and Humans, IEEE Transactions on*, vol. 36, no. 5, pp. 862–875, 2006.
13. D. Heylen, "Head gestures, gaze and the principles of conversational structure," *Intl. Journal of Humanoid Robotics*, vol. 3, no. 03, pp. 241–267, 2006.
14. G. Castellano, I. Leite, A. Pereira, C. Martinho, A. Paiva, and P. McOwan, "Affect recognition for interactive companions: challenges and design in real world scenarios," *Journal on Multimodal User Interfaces*, vol. 3, no. 1-2, pp. 89–98, 2010.
15. R. Kaliouby and P. Robinson, "Real-time inference of complex mental states from facial expressions and head gestures," in *Real-Time Vision for Human-Computer Interaction*. Springer US, 2005, pp. 181–200.
16. C. Becker-Asano, D. Sun, B. Kleim, C. N. Scheel, B. Tuschen-Caffier, and B. Nebel, "Outline of an empirical study on the effects of emotions on strategic behavior in virtual emergencies," in *Affective Computing and Intelligent Interaction*, ser. LNCS, vol. 6975. Springer, 2011, pp. 508–517.
17. Trivisio Prototyping GmbH, "VRvision HMD," www.trivisio.com/trivisio-products/vrvision-hmd-5/, June 2013.
18. C. N. Scheel, B. Kleim, J. Schmitz, C. Becker-Asano, D. Sun, B. Nebel, and B. Tuschen-Caffier, "Psychophysiological stress reactions to a simulated fire in an underground parking lot," *Zeitschrift für Klinische Psychologie und Psychotherapie*, vol. 41, no. 3, pp. 180–189, 2012.
19. S. P. Smith and D. Trenholme, "Rapid prototyping a virtual fire drill environment using computer game technology," *Fire Safety Journal*, vol. 44, no. 4, pp. 559 – 569, 2009.

Using Text Classification to Detect Alcohol Intoxication in Speech

Andreas Jauch[1,2], Paul Jaehne[1,2], and David Suendermann[2]

[1] IBM, Böblingen, Germany
[2] DHBW, Stuttgart, Germany

andreas.jauch@de.ibm.com paul.jaehne@de.ibm.com david@suendermann.com

Abstract. This paper focuses on text-based classification of the Munich Alcohol Language Corpus (ALC) which contains speech from persons in intoxicated as well as in sober state. In order to classify whether a person is intoxicated or not, several combinations of classifiers and feature extraction approaches have been examined. One major finding was that the expressiveness of a test was tightly coupled to its type of speech and topic. The best result was achieved by classifying picture description tests using logistic regression which resulted in an unweighted average recall of 89.4%.

1 Introduction

One of the most severe problems in traffic is the abuse of alcohol. In the United States, every day about 30 people die in crashes involving alcohol-impaired drivers totaling more than 50 billion US$ annual cost [1]. Therefore, measuring intoxication by interviewing drivers and analyzing their answers is an encouraging topic of current research activity. The Ludwig Maximilians University of Munich put together a foundation for work related to the detection of alcohol intoxication by producing a publicly available speech corpus. The *Alcohol Language Corpus* (ALC) [7, 8] contains speech recordings and their transcriptions in intoxicated as well as in sober state. To inspire research teams to work on classifying intoxication state on this corpus, the *Interspeech Speaker State Challenge 2011* has been brought up to serve as a stage for competitions [9]. Hence, there already exist a number of publications covering this topic, though to the best of our knowledge all of them—including the intoxication subchallenge winners from Interspeech [3]—used audio-based features either on its own or in combination with others to perform classification. Solely the team from the University of Erlangen [2] measured accuracy[3] of a text-only based system, but later on, they combined it with different kinds of features to improve their results. Furthermore, they excluded textual features from their major final result

[3] text-only based results are provided on their development set only with an unweighted average recall of 59,1% [2]

system because they decreased accuracy. Encouraged by the fact that no text-only focused publications on this challenge exist, we hereby present our results on classification using textual features only.

2 System Description

2.1 Basic Setup

Although the *ALC* was originally created as a speech corpus, it is exhaustively transcribed allowing for easy feature extraction from these transcripts[4]. Using the WEKA toolkit [6], bag-of-word feature vectors in form of word presence or word count vectors were generated on the transcribed speech. Additionally, information on speech irregularities like stutters, repetitions or noticeable pauses was added to the feature set. Apart from this, no other information provided by the corpus—like audio data or phonetic information—was used for feature generation.

In the beginning, the tests were executed using multilayer perception neural networks, decision tables, J48, JRip, naïve Bayes, logistic regression and SMO as classifiers. However, since the latter three produced considerably better results than the others, the final experiments—and such all of those presented in this paper—were only performed on those three. All experiments have been executed using 10-fold cross-validation. To be able to compare results with other publications on the same matter, we used unweighted average recall (UAR) as performance metric as calculated by

$$\text{UAR} = \frac{\text{recall(alc)} + \text{recall(nonalc)}}{2} = \frac{\frac{tp}{tp+fp} + \frac{tn}{fn+tn}}{2} \tag{1}$$

We are making the code written to conduct these experiments available to the public in the form of an open-source GIT repository on

http://suendermann.com/corpus/alc.html

and researchers are encouraged to live up to these results.

2.2 Enhancements

Tests were done in several iterations, where the goal of each iteration was to top the resulting accuracies of the previous tests.

In the first iteration, feature selection using information gain was applied to discard features whose contribution of information is lower than a certain threshold. Information gain is computed by evaluating the differences in entropy [5] with or without the knowledge of that feature. In order to optimize the final

[4] The transcripts, which build the basis for this paper, where generated by manual annotation.

set of features, the ideal threshold needed to be found, which was achieved by running a series of test runs using different thresholds.

To profit from the different approaches of the used classifiers, SMO, logistic regression and naïve Bayes were combined into a majority voting system that predicts the resulting class by way of majority vote. As the corpus contains speech from different topics[5] ranging from simple tasks as reading a telephone number, over tongue twisters to picture description, the corpus was further divided into its 11 document classes to asses differences in their expressiveness. Since tests were executed on both the entire corpus and all individual document class sub-corpora in the same manner, the tests can be easily compared.

3 Experiments

3.1 Corpus Description

The ALC provides German speech of 77 female and 85 male speakers, recorded both sober and intoxicated. Its vocabulary size is 15776 words, where all the different dialectical forms of one word are counted separately. The ALC focuses not only on read speech, but it also contains a variety of different spontaneous speech samples. In addition to that, the corpus also contains tests on command and control speech for its applicability in an automotive environment. Altogether, there are 11 different document classes that can be divided into spontaneous and non-spontaneous speech as shown in Table 1.

When running experiments using feature selection on the corpus, some words resulted in a surprisingly high information gain in contrast to the rest of the corpus. This turned out to be due to the fact that there are some tests not available in both states—e.g., one tongue twister only appears in intoxicated state. Of course, this would give a text classifier considerable advantage. Thus, we removed all those tests not available in both states from the corpus corresponding to a decrease in size by about one third. Still containing 4698 intoxicated samples and 4978 sober ones, this now modified corpus is almost equally balanced with a distribution of 48.6% to 51.4%. Hence, WEKA's default classifier ZeroR always picking the most frequent class produced a baseline UAR of 51.4%. The modified ALC corpus has a total of 11386 words vocabulary.

[5] in the following referred to as *document classes*

Table 1. description of sub-corpora

Doc Class	Description	Speech Type	#Types	#Samples	%Samples ALC	%Samples NonALC
LN	list numbers	read	264	1660	48.80%	51.20%
LT	list tongue twister	read	174	344	47.09%	52.91%
LS	list spelling	read	107	344	47.09%	52.91%
RT	read tongue twister	read	527	1316	49.24%	50.76%
RR	read command	read	175	1356	47.79%	42.21%
RA	read address	read	491	1356	47.79%	42.21%
DQ	dialogue question	spontaneous	4651	324	50.00%	50.00%
DP	dialogue picture description	spontaneous	2974	344	47.09%	52.91%
MQ	monologue question	spontaneous	2961	344	47.09%	52.91%
MP	monologue picture	spontaneous	4177	648	50.00%	50.00%
EC	elicited command	spontaneous	979	1640	49.39%	50.61%
all	complete corpus	various	11386	9676	48.55%	51.45%

3.2 Experiment I - Entire Corpus

Our first approach was to feed the entire corpus into all three classifiers, which produced results hardly outperforming the ZeroR baseline. Furthermore, this approach comes along with immense computational requirements due to more than 11000 features to be processed. While naïve Bayes and SMO were able to produce results in a reasonable time frame, logistic regression took more than two weeks to complete. Concluding from this test, it can be said that this set of features is too large to be applicable for a text-only-based classification.

After application of feature selection, logistic regression achieved the best accuracy with 58.80% UAR[6] which relates to a relative improvement of more than 14% over the ZeroR baseline. Nevertheless, it was still lower than the Interspeech Speaker State Challenge 2011 baseline[7] of 65.9% UAR [9].

3.3 Experiment II - Individual Document Classes

Since the combination of all features did not lead to the desired results, the next experiment was targeted on checking whether it is useful to further divide the corpus into its different document classes. This approach was also motivated by the question how the performance of a simple task like reading a telephone number compares to that of a rather difficult test such as describing a picture. The

[6] This result was achieved with an information gain threshold of 0.0002 which included 971 featues.

[7] That number was provided in the call for participation which was, however, not restricted to text-only-based features. As some of the samples have been removed to avoid discrepancies in the corpus—as described in section 3.1—these results are not directly comparable.

expectation was that a classifier specifically trained on one document class could be more effective than a classifier trained on the whole corpus. Consequently, the corpus was divided into 11 sub-corpora each of which containing only transcriptions from one document class. Table 1 shows details about each sub-corpus. A similar idea was published in [11], where a comparison between the different prompt types *spontaneous speech*, *tongue twister* and *command-and-control* was performed.

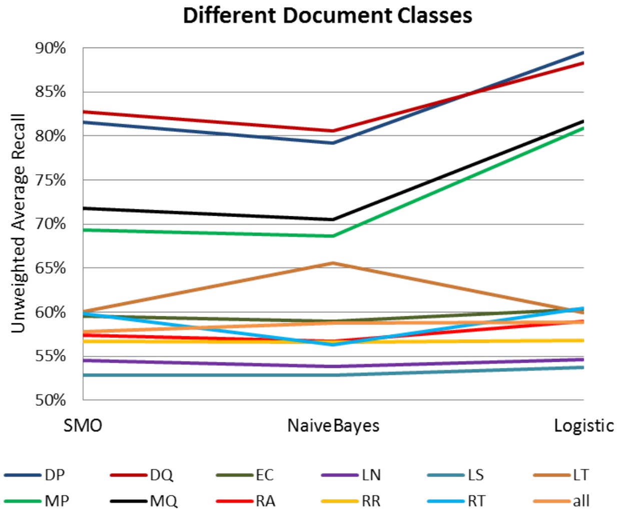


Fig. 1. accuracy of document classes

As shown in Figure 1, there are considerable differences on achieved accuracies supporting our conjecture that document classes vary in terms of expressiveness. The diagram compares the unweighted average recall achieved on each document class using SMO, naïve Bayes and logistic regression classifiers. Just as before, feature selection was applied on the basis of an information gain threshold optimized for each document class. Figure 2 shows the influence of the information gain threshold for the class DP (picture description) in detail. As expected, classes containing spontaneous speech performed best. Among them, dialogue-speech-based document classes (DP, DQ) achieved an unweighted average recall between 79.17% and 89.43% performing considerably better than monologue classes. The latter form the next group in the result set, performing between 68.67% UAR and 80.86% UAR. Table 1 shows that these classes have a considerably higher number of word types than the non-spontaneous classes, being a likely reason for the superior performance.

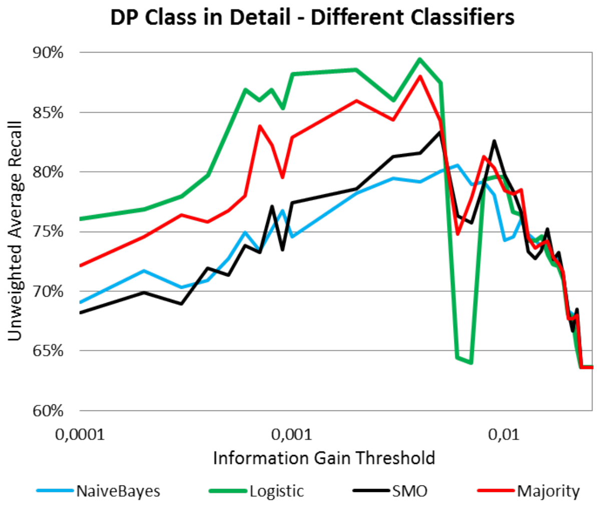


Fig. 2. DP (picture description) optimizing IG threshold

This diagram shows the positive impact of feature selection using information gain on the results. Furthermore, a strange finding from this experiment is that SMO and logistic regression show a steep decline right after their peaks, whereas naïve Bayes does not suffer from such a drop. We were able to recreate this effect in multiple test iterations, but were not able to positively identify its cause as of yet.

3.4 Experiment III - Word Counts

Since the previous tests were all performed using word presence bag-of-word features, the next test examined the potential of adding word count information. Figure 3 reveals that, contrary to our hope, the change from word presence to word count deteriorates accuracy. When using logistic regression the difference in accuracy between word count and word presence is rather small, whereas much higher differences can be observed when using naïve Bayes.

3.5 Experiment IV - Combining Document Classes

The preceding results made us wonder whether a combination of the best performing document classes, e.g. DP, DQ, and MQ, could further improve results due to synergetic effects. At first, we merged sub-corpora containing these three classes into one corpus. Although this combination performed considerably better than the run on the undivided corpus, synergetic effects were not as strong as anticipated and did not result in an improvement compared to individual document classes.

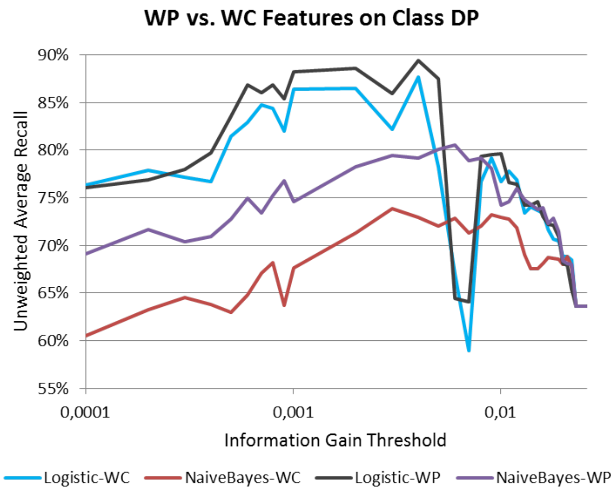


Fig. 3. word presence vs. word count features

3.6 Experiment V - Combining Classifiers

As a last experiment, the three classifiers under consideration were combined into a majority voting system, whose results are shown in Figure 2, too. It turned out that even classifier combination was not able to beat logistic regression, which seems to be the best on this specific domain.

4 Conclusion and Outlook

This study showed the power of pure text-based features to determine whether somebody is intoxicated or sober by analyzing speech transcriptions. There are two major findings. First, it is enough to limit analysis to a single spontaneous speech task as it is much more expressive than read speech. Second, the use of text-based features turned out to be very effective especially in conjunction with logistic regression.

Although the final result of 89.4% UAR on the most expressive document class (DP) is an excellent achievement, it needs to be said that the accuracy still needs to be improved before considering operational scenarios. Yet it is an important step forward to understand that it is not necessary to combine all document classes of the ALC, but that it is more worthwhile to concentrate on one or maybe two classes only. This also allows for further improvement of the test procedure itself since tests can now be developed with a special focus on spontaneous speech.

To improve classification accuracy even further, we will be considering n-gram features to model word order dependencies with a proven record of performance gain in text classification [4, 10]. Furthermore, we plan to consider a weighting system for classifier combination, such that better classifiers are less likely to be outvoted by worse ones. Also, majority voting could be applied across multiple document classes as suggested in [11] as well.

Another area to look into is the applicability of this research to real-world scenarios. As manual transcription of speech is not available in real-time, software for analysis of intoxication based on text will have to rely on automatic speech recognition. It will therefore be necessary to analyze the influence of speech recognition performance on the accuracy of classification. This is particularly interesting considering that voice and speech characteristics of users may be subject to substantial change under the influence of alcohol.

References

1. L. Blincoe, A. Seay, E. Zaloshnja, T. Miller, E. Romano, S. Luchter, and R. Spicer. The Economic Impact of Motor Vehicle Crashes 2000. Technical report, U.S. Department of Transportation, 2002.
2. T. Bocklet, K. Riedhammer, and E. Nöth. Drink and Speak: On the Automatic Classification of Alcohol Intoxication by 46 Acoustic, Prosodic and Text-Based Features. In *Proc. of the Interspeech*, Florence, Italy, 2011.
3. D. Bone, M. Black, M. Li, A. Metallinou, S. Lee, and S. Narayanan. Intoxicated Speech Detection by Fusion of Speaker Normalized Hierarchical Features and GMM Supervectors. In *Proc. of the Interspeech*, Florence, Italy, 2011.
4. C. Boulis and M. Ostendorf. Text Classification by Augmenting the Bag-of-Words Representation with Redundancy-Compensated Bigrams. In *Proc. of the FSDM*, Newport Beach, USA, 2005.
5. I. Guyon, S. Gunn, M. Nikravesh, and L. Zadeh. *Feature Extraction: Foundations and Applications (Studies in Fuzziness and Soft Computing)*. Springer, New York, USA, 2006.
6. M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. Witten. The WEKA Data Mining Software: An Update. *SIGKDD Explorations*, 11(1), 2009.
7. F. Schiel and C. Heinrich. Laying the Foundation for In-Car Alcohol Detection by Speech. In *Proc. of the Interspeech*, Brighton, UK, 2009.
8. F. Schiel, C. Heinrich, S. Barfüsser, and T. Gilg. ALC - Alcohol Language Corpus. In *Proc. of the LREC*, Marrakesch, Marokko, 2008.
9. B. Schuller, S. Steidl, A. Batliner, F. Schiel, and J. Krajewski. The INTERSPEECH 2011 Speaker State Challenge. In *Proc. of the Interspeech*, Florence, Italy, 2011.
10. C. Tan, Y. Wang, and C. Lee. The use of bigrams to enhance text categorization. *Information Processing and Management*, 38(4), 2002.
11. F. Weninger and B. Schuller. Fusing Utterance-Level Classifiers for Robust Intoxication Recognition from Speech. In *Proc. of the ICMI 2011*, Alicante, Spain, 2011.

No matter how real: Out-group faces convey less humanness

Aleksandra Swiderska, Eva G. Krumhuber, Arvid Kappas

School of Humanities and Social Sciences,
Jacobs University Bremen,
Campus Ring 1, 28759 Bremen, Germany
{a.swiderska, e.krumhuber, a.kappas}@jacobs-university.de

Abstract. Past research on real human faces has shown that out-group members are commonly perceived as lacking human qualities, which links them to machines or objects. In this study, we aimed to test whether similar out-group effects generalize to artificial faces. Caucasian participants were presented with images of male Caucasian and Indian faces and had to decide whether human traits (naturally and uniquely human) as well as emotions (primary and secondary) could or could not be attributed to them. In line with previous research, we found that naturally human traits and secondary emotions were attributed less often to the out-group (Indian) than to the in-group (Caucasian), and this applied to both real and artificial faces. The findings extend prior research and show that artificial stimuli readily evoke intergroup processes. This has implications for the design of animated characters, suggesting that out-group faces convey less humanness regardless of how life-like their representation is.

Keywords: objectification; realism; face perception; emotion; out-group

1 Introduction

A long-standing question within the field of computer science and artificial intelligence concerns the degree of human likeness required by animated characters, computer agents, and robots. In general, human appearance is viewed as advantageous as it provides a more intuitive and effective interface [1], [2], thereby facilitating various aspects of human-computer interaction. This relates particularly to the face being the most immediate source of communication [3]. Consequently, attempts to increase its humanness have aimed at the development of photorealistic faces that strongly resemble those of living humans [4], [5]. On the other hand, the strive for realism has been countered by arguments about possible experiences of alienation due to the ‘uncanny valley’ [6]. That is, if computer-generated faces become too close to humans, without making people fully believe that they are real, feelings of discomfort and repulsion may arise [7].

Research exploring perception of artificial facial stimuli, and whether this process is different from that of real faces, has been so far inconclusive. For example, studies involving functional magnetic resonance imaging (fMRI) and event-related brain potentials (ERP) found that artificial faces may be processed distinctly from real faces

[8], [9]. Also, the recognition of emotions seems to vary depending on the type of the face [10], [11]. On the contrary, there is evidence suggesting that people do respond to various kinds of artificial faces, as well as to face-like objects, similarly as to real faces [12], [13]. Given that external features and expressions can be easily manipulated and controlled in synthetic faces, they are commonly employed in social categorization studies, the results of which map onto those employing real faces [14], [15].

Independently of controversies linked to the level of realism, faces convey diverse qualities and are processed quickly and possibly preferentially by our brains [16], [17]. They attract attention faster than non-face objects [18] and as a visual cue are favored over other types of input [19]. Furthermore, faces are a social stimulus of major functional significance, prompting rapid evaluations of people on a number of dimensions. These include gender, age, and ethnicity on the most basic level [20], as well as other traits, for instance attractiveness, likeability, and trustworthiness [21]. Importantly, faces are often the first source of information pertaining to group membership. Categorization of people as belonging to one's in-group or out-group is common and in fact unavoidable in real-life social encounters [22]. Numerous studies have shown that in-groups are generally favored over out-groups [23], [24], and that out-group members are frequently the targets of prejudice [25], [26]. One facet of prejudice is the failure to perceive out-group members as complete human beings [27]. In this case, out-groups are not granted the full range of human qualities, including personality traits [28], emotions [29], and mental states [30]. Denial of traits that have been identified as natural or essential to all human beings (naturally human traits, e.g., warmth, depth) reduces people to objects, such as machines or automata [31]. The equation of humans with objects is referred to as objectification [32]. It is associated with the perceived lack of mind and what follows, compromised capacities which are distinctive for humans [33], for example the ability to experience refined emotions (secondary emotions, e.g., elevation, envy) [29].

Although prior studies using real faces have demonstrated that out-group members appear overall less human compared to in-group members [34], [35], [36], this intergroup phenomenon has not been investigated yet in the context of artificial faces. Respective issue seems of importance since the (differential) attribution of human characteristics to group members should apply only to faces of real human beings, that is, entities that actually are alive. Alternatively, if synthetic faces of out-group members are seen as representing diminished humanness, just as real faces are in daily interactions, the most realistic animation may not be adequate for them to be perceived as human-like. Therefore, regardless of how authentic the representation is, out-group faces may still be viewed as machines/ objects.

In the present study, we wanted to explore the process of objectification as it applies to real as well as artificial faces. For this, artificial facial stimuli were used that were highly human-like, but clearly distinguishable from their real counterparts in terms of aliveness (see [37]). To maximize differences in perception, Caucasian participants were presented with both real and artificial faces of Caucasian (in-group) and Indian (out-group) individuals. Participants' task was to decide whether certain human traits (naturally and uniquely human) and emotions (primary and secondary) could or could not be attributed to each face. As out-groups are generally associated more with objects, we would predict that less human qualities and secondary emotions would be attributed to them. Furthermore, such effects should be similar for real and artificial faces.

2 Experiment

Thirty-one students (11 men), ranging in age from 19 to 27 years (M = 21.84, SD = 2.12), at Warsaw University, Poland, participated on a voluntary basis and were paid 20 PLN (~5€). All of them identified themselves as Polish, with three people holding a double Polish-American citizenship. Information about the experiment was distributed in English and directed primarily at the students of an English-language psychology program to ensure fluency in this language. However, the mother tongue of all participants was Polish.

Facial stimuli (see Fig. 1) consisted of photographs of eight neutral faces of Caucasian and Indian (four of each) adult males, obtained from the Center for Vital Longevity Face Database [38]. These photographs were selected from a larger set of realistic facial stimuli which did not differ in a pretest (N = 30) with regard to their perceived attractiveness (scale 1-5; $M_{Caucasian}$ = 1.97 vs. M_{Indian} = 1.83, p = 0.062), intelligence ($M_{Caucasian}$ = 2.83 vs. M_{Indian} = 2.72, p = 0.348), trustworthiness ($M_{Caucasian}$ = 2.62 vs. M_{Indian} = 2.49, p = 0.152), and likeability ($M_{Caucasian}$ = 2.64 vs. M_{Indian} = 2.71, p = 0.480). The eight photographs served as a basis for creating the faces' artificial analogues by applying a variety of modifications in Photoshop (CS3-ME, Adobe Systems Inc., 2007) while preserving their identity (i.e., no changes in facial morphology). The modified faces (four Caucasian, four Indian) composed the set of artificial versions of the stimuli and differed significantly in perceived aliveness, as determined in an independent study (N = 60, M_{real} = 6.15 vs. $M_{artificial}$ = 1.52, $p <$ 0.001, scale 1-7). Additionally, eight images of cars were included as filler items with the purpose to distract from the target manipulation. This resulted in a set of 24 pictures which measured 627 x 479 pixels and were displayed on a white background.

Participants took part in the study individually. Their task was to decide "whether a certain characteristic could or could not be ascribed to a stimulus". This decision had to be made as quickly as possible for every stimulus and characteristic which added up to 384 trials (24 images x 16 characteristics, described below), presented randomly in four blocks. To signal the beginning of each trial, a fixation cross appeared on the top of the screen for 500 ms; it was then replaced by a word (label of a trait), which was displayed for 1000 ms; finally, underneath the word, a picture of the stimulus appeared. The word and the picture remained on the screen until participants gave a response by pressing a key on the keyboard, corresponding to either a '*yes*' (the characteristic can be attributed to the stimulus) or a '*no*' (the characteristic cannot be attributed to the stimulus) judgment. The experimental task was delivered using DirectRT software (Version 2010, Empirisoft Co., NYC, USA).

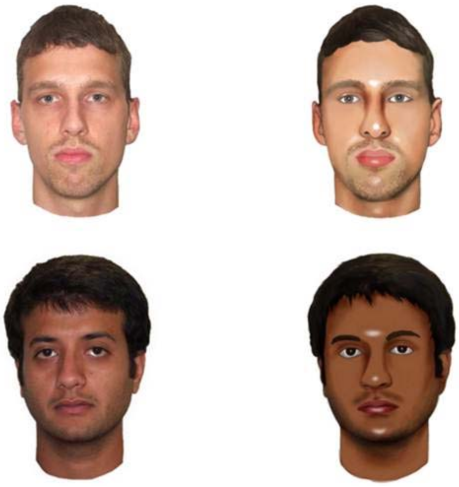


Fig. 1. Examples of real (left) and artificial (right) Caucasian and Indian faces.

Measures targeted the frequency with which a given characteristic (human traits and emotions) was ascribed to a stimulus. Human traits were selected based on dehumanization research by Haslam and colleagues [31], [39], and comprised four uniquely human traits (positive: *organized, broadminded*; negative: *rude, shallow*) and four naturally human traits (positive: *friendly, trusting*; negative: *shy, impatient*). Emotion terms were drawn from research by Demoulin et al. [40], and comprised four primary emotions (positive: *pleased, calm*; negative: *fearful, angry*) and four secondary emotions (positive: *sympathetic, hopeful*; negative: *ashamed, guilty*). As both positive and negative characteristics were included, valence was treated as an additional factor in the analysis of results.

3 Results

A multivariate analysis of variance (MANOVA) with Ethnicity (Caucasian, Indian), Realism (real, artificial), and Valence of the traits (positive, negative) as within-subjects factors was conducted on the four dependent variables (uniquely and naturally human traits, primary and secondary emotions). For all univariate analyses, the Greenhouse-Geisser adjustment to degrees of freedom was applied. No significant main effect emerged for Realism $F(4, 27) = 1.57$, $p = 0.212$, $\eta_p^2 = 0.19$, suggesting similar responses to real and artificial faces. The multivariate main effects were significant for Ethnicity, $F(4, 27) = 5.32$, $p = 0.003$, $\eta_p^2 = 0.44$, and Valence, $F(4, 27) = 8.07$, $p < 0.001$, $\eta_p^2 = 0.55$. These two main effects were qualified by a significant multivariate interaction between Ethnicity and Valence, $F(4, 27) = 4.07$, $p = 0.01$, $\eta_p^2 = 0.38$. In terms of univariate tests, this interaction was significant for the naturally human traits, $F(1, 30) = 8.99$, $p = 0.005$, $\eta_p^2 = 0.23$, as well as for secondary emotions,

$F(1, 30) = 4.44$, $p = 0.044$, $\eta_p^2 = 0.13$. Analyses of simple effects showed that participants attributed more naturally human traits and secondary emotions to Caucasians ($M = 0.56$ and $M = 0.52$) in comparison to Indians ($M = 0.40$ and $M = 0.42$). However, this was the case only in the context of positive characteristics. No such differences occurred for negative naturally human traits and negative secondary emotions ($p > 0.05$). The proportions of characteristics attributed to Caucasian and Indian faces can be seen in Fig 2.

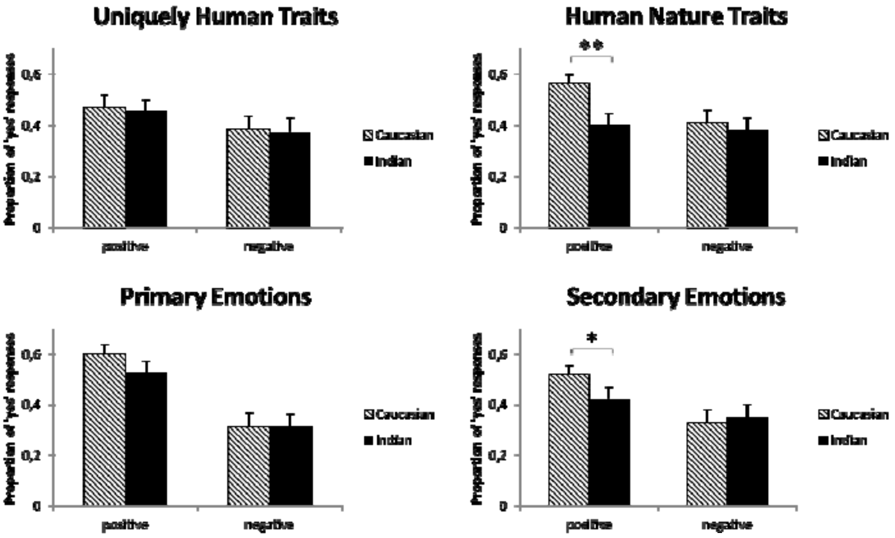


Fig. 2. Mean proportions of human traits and emotions attributed to Caucasian and Indian faces; $^{**}p < 0.01$, $^{*}p < 0.05$.

4 Discussion

Out-group members are commonly denied an array of human qualities which reduces them to objects, such as machines or automata. These specifically include traits perceived as naturally human [31] and refined emotions that entail high cognition and morality (secondary emotions), but not emotions shared with other species (primary emotions) [29]. In the current study, our objective was to extend previous findings that demonstrated the process of objectification for real faces and to investigate whether this generalizes to artificial faces. In line with previous research, the results showed that participants attributed less naturally human traits to faces of out-group members in comparison to in-group members. Moreover, they attributed less secondary emotions to faces of out-group members than to in-group members, while there were no differences in how primary emotions and uniquely human traits were attributed. The findings applied to both real and artificial stimuli, suggesting that the latter readily evoke intergroup processes, bringing about out-group effects comparable in their nature to real faces.

Although the pattern of differential attribution of human traits and emotions to in-group and out-group members was consistent with predictions implying objectification [31], it concerned largely positive characteristics. Besides similar findings in the literature [35], there is evidence suggesting that the valence of human characteristics may not play a crucial role in how these are associated with groups [41]. In fact, sometimes negative naturally human traits tend to be connected even more with the in-group, justifying the “only human”, inborn and therefore uncontrollable nature of their flaws [42]. Differential attribution of positive characteristics can thus be seen as one facet of objectification, thereby indicating a positivity bias towards the in-group.

The findings have important implications for the design and animation of computer-generated characters. Up to now, the developments in computer graphics have focused on increasing the realism of synthetic characters to the point that they are indistinguishable from living humans. In this context, major efforts have been devoted to the generation of photorealistic faces. To circumvent deficiencies in appearance that lead to unsettling impressions on the part of viewers, ascribed to the ‘uncanny valley’ [6] research in turn has scrutinized the potentially problematic elements of faces. This was typically done as if the faces were a collection of separate features, colors, and textures [7], [37] rather than a whole that functions as a social stimulus in interaction with the human perceiver.

In the current paper, we have shown that group membership plays a major role in how human-like a face appears to be. One possible extension of this research would be to conduct it in a different country. For instance, would Indian participants attribute greater humanness to Indian (in-group) faces and perceive Caucasian (out-group) faces as objects? Further, how would categories other than race or ethnicity influence perception of human qualities in another? People go beyond what is directly observable and constantly make inferences about the underlying states, intentions, and qualities of other interactants. A crucial function of faces is that they represent human qualities and are associated with minds that powerfully suggest a potential for mental connection, constantly sought after by humans [43]. Nonetheless, people will not connect with everybody in the same manner. Group membership proved to be a basic criterion that determines to what extent such mental connection is achieved. Independently of whether the face is real or artificial, we demonstrated that out-group members were generally viewed as more machine/ object-like than in-group members, hence embodying lesser humanness and what follows, reduced promise of bonding. Accentuated realism of computer-generated faces alone may consequently not be sufficient to capture the complexities and subtleties of human perception. Rather, the design of human-like agents necessitates consideration of purpose-related, social psychological processes.

Acknowledgments. This work has been conducted within the European Commission project eCUTE – Education in Cultural Understanding, Technologically-Enhanced (FP7-ICT-2009.4.2).

References

1. Appel, J., von der Pütten, A., Krämer, N. C., Gratch, J.: Does humanity matter? Analyzing the importance of social cues and perceived agency of a computer system for the emergence of social reactions during human-computer interaction. Adv. Hum.-Comput. Int. 2012, 1--10 (2012)
2. Breazeal, C. L.: Designing Social Robots. MIT Press, Cambridge (2002)
3. Sproull, L., Subramani, M., Kiesler, S., Walker, J.H., Waters, K.: When the Interface Is a Face. Hum-Comput. Interact. 11, 97--124 (1996)
4. Alexander, O., Rogers, M., Lambeth, W., Jen-Yuan Chiang, Wan-Chun Ma, Chuan-Chang Wang, Debevec, P.: The Digital Emily Project: Achieving a Photorealistic Digital Actor. IEEE Comput. Graph. 30, 20--31 (2010)
5. Takács, B., Kiss, B.: The Virtual Human Interface: A Photorealistic Digital Human. IEEE Comput. Graph. 23, 38--45 (2003)
6. Mori, M.: The Uncanny Valley. (K. F. Macdorman & N. Kageki, Trans.). IEEE Robot. Autom. Mag. 19, 98--100 (2012)
7. MacDorman, K. F., Green, R. D., Ho, C. C., Koch, C. T.: Too real for comfort? Uncanny responses to computer generated faces. Comput. Hum. Behav. 25, 695--710 (2009)
8. Mühlberger, A., Wieser, M. J., Herrmann, M. J., Weyers, P., Tröger, C., Pauli, P.: Early cortical processing of natural and artificial emotional faces differs between lower and higher socially anxious persons. J. Neural Transm. 11, 735--746 (2009)
9. Wheatley, T., Weinberg, A., Looser, C., Moran, T., Hajcak, G.: Mind perception: Real but not artificial faces sustain neural activity beyond the N170/VPP. PLoS One 6, 1--7 (2011)
10. Dyck, M., Winbeck, M., Leiberg, S., Chen, Y., Gur, R. C., Mathiak, K.: Recognition profile of emotions in natural and virtual faces. PLoS ONE 3, e3628, 1--8 (2008)
11. Kätsyri, J., Sams, M.: The effect of dynamics on identifying basic emotions from synthetic and natural faces. Int. J. Hum.-Comput. St. 66, 233--242 (2008)
12. Blascovich, J., Loomis, J., Beall, A. C., Swinth, K. R., Hoyt, C. L., Bailenson, J. N.: Immersive virtual environment technology as a methodological tool for social psychology. Psychol. Enq. 13, 103--124 (2002)
13. Krumhuber, E., Manstead, A. S. R., Cosker, D., Marshall, D., Rosin, P. L.: Effects of dynamic attributes of smiles in human and synthetic faces: A simulated job interview setting. J. Nonverbal Beh. 33, 1--15 (2009)
14. Corneille, O., Hugenberg, K., Potter, T.: Applying the attractor field model to social cognition: Perceptual discrimination is facilitated, but memory is impaired for faces displaying evaluatively congruent expressions. J. Pers. Soc. Psychol. 93, 335--352 (2007)
15. Hugenberg, K., Bodenhausen, G. V.:Ambiguity in social categorization: The role of prejudice and facial affect in race categorization. Psychol. Sci. 15, 342--345 (2004)
16. Kappas, A., Olk, B.:The concept of visual competence as seen from the psychological and brain sciences. Visual Stud. 23, 162--173 (2008)
17. Liu, J., Harris, A., Kanwisher, N.: Stages of processing in face perception: an MEG study. Nat. Neurosci. 5, 910--916 (2002)
18. Langton, S. R. H., Law, A. S., Burton, A. M., Schweinberger, S. R.: Attention capture by faces. Cognition 107, 330--342 (2008)
19. Beckett, N. E., Park, B.: Use of category versus individuating information making base rates salient. Pers. Soc. Psychol. B. 21, 21--31 (1995)
20. Fiske, S. T.: Stereotyping, prejudice, and discrimination. In: Gilbert, D. T., Fiske, S. T., Lindzey, G. (eds.) The Handbook of Social Psychology, 4th edition, pp 357--411. McGraw-Hill, New York (1998)
21. Willis, J., Todorov, A.: First impressions: Making up your mind after a 100-ms exposure to a face. Psychol. Sci. 17, 592--598 (2006)
22. Allport, G. W.: The Nature of Prejudice. Addison-Wesley, Reading (1954)

23. Brewer, M. B.: The psychology of prejudice: Ingroup love or outgroup hate? J. Soc. Issues 55, 429--444 (1999)
24. Hewstone, M., Rubin, M., Willis, H.: Intergroup bias. Ann. Rev. Psychol. 53, 575--604 (2002)
25. Brewer, M. B., Brown, R. J.: Intergroup relations. In: Gilbert, D. T., Fiske, S. T., Lindzey, G. (eds.) The Handbook of Social Psychology, 4th edition, pp 554--594. McGraw-Hill, New York (1998)
26. Fiske, S. T., Cuddy, A. J. C., Glick, P., Xu, J.: A model of (often mixed) stereotype content: Competence and warmth respectively follow from perceived status and competition. J. Pers. Soc. Psychol. 82, 878--902 (2002)
27. Harris, L. T., Fiske, S. T.: Social neuroscience evidence for dehumanized perception. Eur. Rev. Soc. Psychol. 20, 192--231 (2009)
28. Haslam, N., Bain, P., Douge, L., Lee, M., Bastian, B.: More human than you: Attributing humanness to self and others. J. Pers. Soc. Psychol. 89, 973--950 (2005)
29. Leyens, J. P., Paladino, P. M., Rodriguez-Torres, R., Vaes, J., Demoulin, S., Gaunt, R.: The emotional side of prejudice: The attribution of secondary emotions to ingroups and outgroups. Pers. Soc. Psychol. Rev. 4, 186--197 (2000)
30. Harris, L. T., Fiske, S. T.: Dehumanizing the lowest of the low: Neuroimaging responses to extreme out-groups. Psychol. Sci. 17, 847--853 (2006)
31. Haslam, N.: Dehumanization: An integrative review. Pers. Soc. Psychol. Rev. 10, 252--264 (2006)
32. Fiske, S. T.: From dehumanization and objectification to rehumanization: Neuroimaging studies on the building blocks of empathy. Ann. New York Academy of Sciences 1167, 31--34 (2009)
33. Epley, N., Waytz, A.: Mind perception. In: Fiske, S. T., Gilbert, D. T., Lindzey, G. (eds.) The Handbook of Social Psychology, 5th edition, pp 498--541. Wiley, Hoboken (2010)
34. Bain, P., Park, J., Kwok, C., Haslam, N.: Attributing human uniqueness and human nature to cultural groups: Distinct forms of subtle dehumanization. Group Process. Interg. 12, 789--805 (2009)
35. Boccato, G., Cortes. B. P., Demoulin, S., Leyens, J. P.:, 2007 “The automaticity of infra-humanization” Eur. J. Soc. Psychol. 37, 987--999 (2007)
36. Saminaden, A., Loughnan, S., Haslam, N.: Afterimages of savages: Implicit associations between ‘primitives’, animals, and children. Brit. J. Soc. Psychol. 49, 91--105 (2010)
37. Looser, C. E., Wheatley, T.: The tipping point of animacy: How, when, and where we perceive life in a face. Psychol. Sci. 21, 1854--1862 (2010)
38. Minear, M., Park, D. C.: A lifespan database of adult facial stimuli. Behav. Res. Meth. Instr. 36, 630--633 (2004)
39. Loughnan, S., Haslam, N.: Animals and androids: Implicit associations between social categories and nonhumans. Psychol. Sci. 18, 116--121 (2007)
40. Demoulin, S., Leyens, J. P., Paladino M. P., Rodriguez-Torres, R., Rodriguez-Perez, A., Dovidio, J. F.: Dimensions of “uniquely” and “non-uniquely” human emotions. Cognition Emotion 18, 71--96 (2004)
41. Paladino, M. P., Leyens, J. P., Rodriguez, R., Rodriguez, A., Gaunt, R., Demoulin, S.: Differential association of uniquely and non uniquely human emotions with the ingroup and the outgroup. Group Process. Interg. 5, 105--117 (2002)
42. Koval, P., Laham, S. M., Haslam, N., Bastian, B., Whelan, J. A.: Our flaws are more human than yours: Ingroup bias in humanizing negative characteristics. Pers. Soc. Psychol. B. 38, 1--13 (2011)
43. Wheatley, T., Kang, O., Parkinson, C., Looser, C. E.: From mind perception to mental connection: Synchrony as a mechanism for social understanding. Soc. Pers. Psychol. Compass 6, 589--606 (2012)

TARDIS - a job interview simulation platform

Hazaël Jones[1] and Nicolas Sabouret[2]

[1] LIP6 - UPMC, Paris, FRANCE, hazael.jones@lip6.fr
[2] LIMSI, Orsay, FRANCE, nicolas.sabouret@limsi.fr

The TARDIS[3] project, funded by FP7, aims at building a serious game for NEETs[4] and employment/inclusion organisations which supports social training and coaching in the context of job interviews [5].

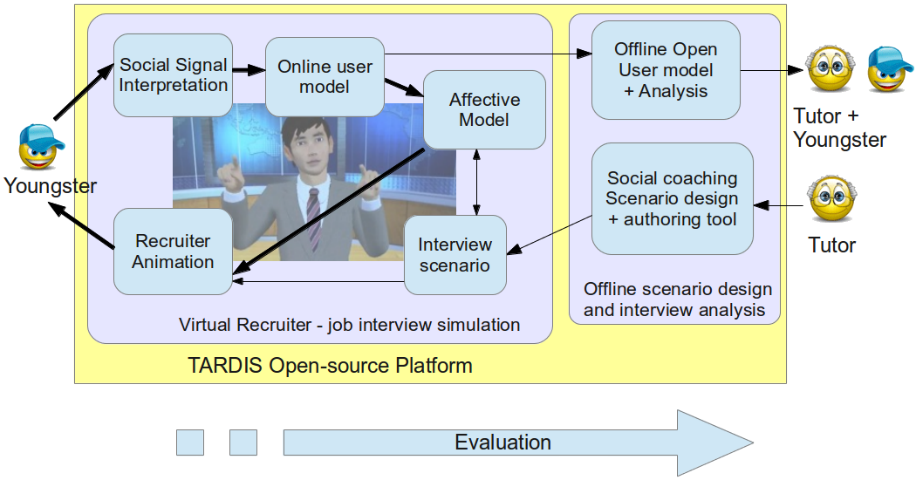

Social Signal Interpretation
Online user model
Affective Model
Offline Open User model + Analysis
Tutor + Youngster
Youngster
Social coaching Scenario design + authoring tool
Tutor
Recruiter Animation
Interview scenario
Offline scenario design and interview analysis
Virtual Recruiter - job interview simulation
TARDIS Open-source Platform
Evaluation

Fig. 1. Architecture of the TARDIS platform

This project has several objectives: 1) to define socio-emotionally credible interactions between a virtual agent and a human [4], by integrating the 3 dimensions of this process (real-time signal processing of the human affects, cognitive evaluation and adaptation of the virtual recruiter, and emotion expression), 2) to allow NEETs to train their social skills thanks to a simulation platform, 3) to provide empowerment organisations a new tool in their work with youngsters.

In this demonstration, we focus on the first part: the interaction loop. The TARDIS architecture (Fig. 1) is composed by 4 modules:

– *Social Signal Interpretation.* This module allows the detection of youngster affects thanks to a Kinect (a motion sensing input device) and a microphone.

[3] TARDIS stands for Training young Adult's Regulation of emotions and Development of social Interaction Skills. url: www.tardis-project.eu

[4] NEET is a government acronym for young people not in employment, education or training.

- *Interview scenario.* It defines the interview progress and the expectation of the recruiter after a question.
- *Affective model of the virtual recruiter.* This module updates the internal state of the virtual agent thanks to detected affects from the system and expected ones from the scenario. Our affective model [1], specially conceived for job interview, is composed of emotions [2], moods and social attitudes.
- *Virtual recruiter animation.* It allows the real-time display of the recruiter affective states thanks to the GRETA conversational agent [3].

The interaction loop and the behaviour of our virtual recruiter (Fig. 2) will be illustrated on a 10 minutes scripted scenario.

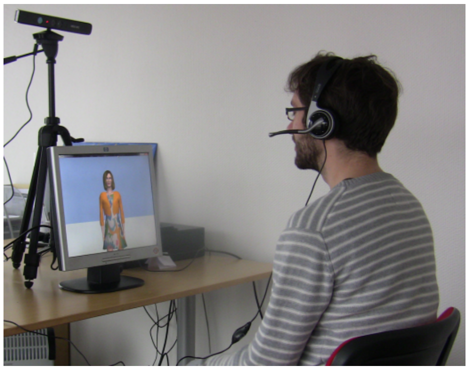


Fig. 2. Setting of the simulation of job interview

Acknowledgement: This research has received funding from the European Programme FP7-ICT-2011-7 under grant agreement 288578.

References

1. H. Jones and N. Sabouret. TARDIS - A simulation platform with an affective virtual recruiter for job interviews. In *IDGEI (Intelligent Digital Games for Empowerment and Inclusion)*, 2013.
2. A. Ortony, G. L. Clore, and A. Collins. *The Cognitive Structure of Emotions.* Cambridge University Press, July 1988.
3. I. Poggi, C. Pelachaud, F. de Rosis, V. Carofiglio, and B. De Carolis. Greta. a believable embodied conversational agent. In *Multimodal intelligent information presentation*, pages 3–25. Springer, 2005.
4. H. Prendinger and M. Ishizuka. the Empathic Companion: a Character-Based Interface That Addresses Users' Affective States. *Applied Artificial Intelligence*, 19(3-4):267–285, Mar. 2005.
5. M. Sieverding. 'Be Cool!': Emotional costs of hiding feelings in a job interview. *International Journal of Selection and Assessment*, 17(4), 2009.