GfK Verein

WORKING PAPER SERIES

AUTOMATIC ANALYSIS OF FACIAL EXPRESSIONS IN AN ADVERTISING TEST WITH CHINESE RESPONDENTS

Anja Dieckmann, Matthias Unfried,
Jens Garbas, Marcello Mortillaro

WORKING PAPER /// NO. 5 / 2017

Working papers are in draft form. This working paper is distributed for purposes of comment and discussion only and may not be reproduced or copied without permission of the copyright holder.
The views expressed are those of the authors and do not necessarily reflect the view of the GfK Verein.
An electronic version can be downloaded from the website of the GfK Verein: www.gfk-verein.org.

Automatic analysis of facial expressions in an advertising test with Chinese respondents

Anja Dieckmann*[†] Matthias Unfried[†] Jens Garbas[‡]
Marcello Mortillaro[§]

Abstract— *An automatic system designed to infer valence (positive or negative feeling) from facial expressions, EMO Scan, was employed and evaluated in an advertising test in China. Its validity has already been demonstrated with Western participants (Garbas et al., 2013), and the purpose of the present study was to examine to what extent it generalizes to Asian participants. In a computer-based interview in a studio setting, participants' faces were recorded while watching TV commercials as well as positive and negative affective pictures. Recordings were analyzed with EMO Scan to infer emotional valence from facial expressions. Inferred valence proved to be a valid predictor of picture valence. Regarding the TV commercials we found that results from facial expression analysis yield results that are similar to explicit ratings. This holds for both time series analysis and overall, time-aggregated results. The conclusion is that EMO Scan is a reliable tool to assess the valence of responses to emotional stimuli also in Chinese respondents.*

Keywords— *Emotion Capturing; Facial Coding; Affective Computing; Ad Testing; Intercultural Studies; China*

1 Introduction

1.1 Automatic analysis of facial expressions

Inferring emotions from facial expressions is not only deeply rooted in human nature (and that of other mammals) but also a topic for scientific exploration since the times of Darwin (1872). With the development of the Facial Action Coding System FACS in the 1970ies (Ekman and Friesen, 1978), a tool for the systematic annotation of the fundamental actions of facial muscles—the so-called Action Units (AU)—has become available. Since then, facial expressions have been used as an important data source not only in fundamental psychological science, but also in many applied research fields, for instance, for pain detection in medicine (e.g., Prkachin, 2009), to monitor client-therapist interaction in psychotherapy (e.g., Rasting et al., 2005), for lie detection in forensic or security contexts (e.g., ten Brinke and Porter, 2012, using a simplified coding system), or for ad testing in marketing (e.g., Debraix, 1995). However, manual FACS coding is very laborious (ratio of coding time to real time can often be 100:1; Prkachin, 2009, p. 182), limiting application in commercial settings to relatively few studies, few cases and short observation periods. More recently, however, with the help of powerful machine learning algorithms, huge progress was made in the detection of facial expressions through intelligent software. Automatic analysis of facial expressions became, and still is, a very active research area in the signal processing and affective computing community (see Martinez and Valstar, 2016; Pantic and Rothkrantz, 2003; Zeng et al., 2009). Numerous systems for the assessment of emotions from facial expressions have been developed to either recognize discrete basic emotions such as "anger" or "happiness" directly or to infer Action Units, which are subsequently used to infer affective states or basic emotions.

[*] Corresponding author, anja.dieckmann@gfk-verein.org

[†] GfK Verein, Fundamental Research, Nuremberg, Germany

[‡] Fraunhofer Institute for Integrated Circuits, Erlangen, Germany

[§] University of Geneva, Swiss Center for Affective Sciences, Switzerland

Citation: Anja Dieckmann, Matthias Unfried, Jens Garbas, Marcello Mortillaro (2017). Automatic analysis of facial expressions in an advertising test with Chinese respondents. GfK Working Paper Series, No. 5 / 2017.

With the advent of such automatic systems for facial expression analysis, the interest from practitioners was revived. Market research and in particular the domain of advertisement pretesting, where objective assessment of emotions is of particular interest, is a promising domain for application, and several companies have started using such systems (e.g., GfK, 2013; Ipsos, 2014; WPP, 2012). Because of the economic importance of Asia, there is demand for software that can be reliably used in these countries. However, most systems are predominantly trained on the basis of Western/Caucasian faces, so it remains unclear to what extent automatic emotion detection can be generalized to Asian respondents.

1.2 GfK EMO Scan

In a typical market research interview, emotional expressions are usually very subtle. EMO Scan was developed to classify such subtle expressions along one core dimension of emotional experience, that is, valence (Fontaine et al., 2007). Valence refers to the overall affective evaluation (Shuman et al., 2013) or hedonic experience (Russell, 2003) of a feeling—more simply put, its pleasantness—and has been widely studied in psychology, mostly using self-reports. However, self-reports are usually given post-hoc when subtle and fleeting emotional experiences may have been forgotten; when self-reports are given concurrent to other activities, like watching an ad, they may impact emotions and thus bias results (cf., e.g., Ariely, 1998). Thus, real-time acquisition of valence over time is almost impossible with self-report methods, especially when changes happen very rapidly, so automated continuous facial expression analysis for valence detection is highly desirable.

EMO Scan is a computer-vision system that has been developed for automated assessment of emotional valence from facial expressions (Garbas et al., 2013). Its core algorithms are based on the SHORE (Sophisticated High-speed Object Recognition Engine) computer vision library developed by Fraunhofer IIS (Ruf et al., 2011). SHORE makes use of machine learning with Real-AdaBoost (Schapire and Singer, 1999) for generating face classification models from an extended Census feature space. Census features encode brightness variations in a local 3x3 pixel neighborhood; being invariant to illumination changes they lead to very robust classifiers for practical applications. SHORE, in addition, uses extended Census features for representing larger image structures (Küblbeck and Ernst, 2006). The efficient software implementation allows for real-time detection and analysis of many faces even on mobile platforms. SHORE has been extended and retrained for valence analysis within EMO Scan. A new face database containing happy, angry, disgusted, sad, frightened, surprised, and neutral labels has been created mostly from TV talk show footage exhibiting spontaneous reactions. The labels were mapped to positive and negative valence classes according to the "Circumplex Model of Affect" (Russell, 1980). The data has been further enhanced with faces exhibiting the most relevant FACS Action Units for positive (AU12) and negative (AU4) valence from other data sets. In sum, more than 13,000 images were used for training two classifiers: one for positive and one for negative valence. For obtaining a continuous valence score, both classifier outputs are fused. This is done by subtracting the score of the negative valence classifier from the score of the positive valence classifier. Further, subject-dependent bias is removed by subtracting a mean value generated during a short calibration period during which participants are, for instance, exposed to a black screen (for more details, see Garbas et al., 2013).

EMO Scan is designed to work robustly under realistic market-research conditions, such as different kinds of consumer equipment and diverse lighting conditions (Unfried and Iwanczok, 2016). The main use case is ad pretesting. The system captures the emotional response discernible from the respondent's face while she watches a TV commercial.

In a validation study with a German sample, average valence scores inferred from faces recorded during exposure to a subset of pictures from the International Affective Picture System (IAPS; Lang et al., 2008) predicted picture valence with accuracy between 78 and 100% depending on threshold (Garbas et al., 2013). Subsequently, EMO Scan has been made available to commercial application by GfK SE. However, its cross-cultural applicability remained an open question. The present study will therefore examine to what extent EMO Scan is able to identify valence in facial expression in a non-Western culture, namely China.

1.3 Emotion and culture

Most scientists agree that basic emotions are universally recognized (as shown, e.g., in a meta-analysis by Elfenbein and Ambady, 2002). Also, the capacity to experience core affect is considered universal (Mesquita, 2003; Russell, 1991). However, scientists also agree that culture affects emotion expression, for instance by defining which expressions are socially acceptable. Especially cultural differences between Asian collectivistic and Western individualistic societies have been the focus of many investigations (e.g., Masuda et al., 2008). Focusing on East Asian countries, already early studies in the 1970ies suggest that culture-specific display rules may lead to "masked" expressions (Ekman and Oster, 1979). For instance, Japanese respondents showed lower level of expression of negative emotions like anger, disgust, fear and sadness (Ekman, 1972; Biehl et al., 1997). Japanese respondents also perform worse than Westerners at identifying these emotions (Biehl et al., 1997). It was argued that the expression of these emotions is socially undesirable in Japan. More recently, it was suggested that lower levels of expression in Japanese respondents even hold for positive emotions (Safdar et al., 2009).

Turning to China, a study using pictures to evoke emotions found that 3-year-old mainland Chinese girls showed fewer smiles and disgust expressions than European-American girls (Camras et al., 2006). Chinese girls adopted by European-American families also differed from Mainland Chinese girls, further supporting the role of culture in emotion expression.

Overall, people in China report lower frequency and intensity scores of positive and negative affect compared to US and Australia (Eid and Diener, 2001). The authors explain this result with cultural differences. Anthropological research suggests that emotions may be considered less relevant in Chinese compared to US American culture (Potter, 1988): While acknowledging emotions as aspects of individual experience, they are not considered a legitimate rationale for actions nor seen as significant basis for social relationships (p. 185). Moreover, in traditional Chinese medicine, excessive emotions are considered as causes for illnesses (Zhang et al., 2015), and their inhibition is believed to be beneficial (Lin, 1983).

Of course, masking or suppressing facial expressions can represent a major problem for automatic analysis of facial expressions. However, eye region expressions are harder to control than mouth (cf., Yuki et al., 2007). Thus, the eye region may provide more reliable and sensitive signals when expression is controlled. Along these lines, Jack et al. (2009) found in a study of eye movements that East Asian observers pay more attention to the eyes than to the mouth as compared to Western Caucasian obsversers. As a consequence, East Asian subjects are more likely to confuse fear and surprise, as well as disgust and anger—emotions whose expression is similar, respectively, in the eye region. Anecdotally, the focus on the eye region is reflected in Asian emoticons that depict different emotion expressions through different representations of the eyes (rather than the mouth as in Western emoticons; Yuki et al., 2007).

Beyond basic emotions, cultural differences increase. Most differences are attributed to the collectivistic nature of Asian societies, in contrast to Western individualism (e.g., Masuda et al., 2008). Also, the stimuli that elicit positive and negative emotions can differ between cultures. To give an example from marketing, a TV commercial that is funny in one country may very well be seen as offensive in another country. But, regardless of cultural background, people's responses to stimuli are expected to differ in the fundamental emotional appraisal dimension of valence—which is the dimension that is targeted by EMO Scan. So, if we can show that EMO Scan can validly distinguish positive and negative expressions in Asian faces, identifying culturally suitable motifs for marketing communication could represent an important application case.

2 Method

2.1 Participants

In order to evaluate if EMO Scan is also a valid method for the inference of emotional states from facial expressions of Asian participants we conducted a study in Shanghai, China. In total, 209 participants completed the survey: 104 male, 105 female. Mean age was 35.6 years (SD = 11.55 years), with a range of 18 to 59 years. Participants were recruited by telephone from the local GfK consumer panel.

2.2 Design and Procedure

Participants were invited to individual rooms by an interviewer and placed in front of a computer screen with a webcam on top. The interviewer informed the participant about the procedure and asked for her consent to take video recordings that would be used in a research project. Left and right of the desk soft light was created by soft boxes to ensure that the respondent's face was well lit and clearly visible for later analysis.

The whole procedure consisted of a web-based questionnaire that was programmed for self-administration but the interviewer remained in the room in case there were any questions.

After the introductory information and general instructions, sociodemographic questions and questions about attitude towards TV commercials were asked. Then, participants were exposed to a block of 5 TV commercials (TVC) in randomized order during each of which respondents' faces were recorded with the webcam. Onset of the TV commercial was synchronized with onset of webcam recording by the questionnaire software. Videos were recorded with 25 frames per second. Each TV commercial was preceded by 3 seconds of black screen exposure for post-hoc calibration, and followed by 2 seconds of black screen for capturing potential aftereffects.

After the commercial block, the TVCs were shown for a second time, now individually and immediately followed by a series of diagnostic questions about the TV commercial and the advertised product. As the main focus of the present paper is on the results for affective pictures (see below), we will restrict analysis to pleasantness ratings and overall liking for the TV commercial as well as liking for the advertised product. Subsequently, TV commercials were shown individually for a third and last time, now followed by appraisal ratings for each of the different scenes contained in a clip.

In the next section, participants saw a slide show consisting of 32 affective pictures in a random sequence, drawn from the International Affective Picture System (IAPS; Lang et al., 2008), 16 of positive, and 16 of negative valence (see Table 1). Erotic pictures and pictures of extreme violence or injuries were excluded so that the range of motifs was well within what is common in marketing communication. Each affective picture was preceded by 1 second of black screen for post-hoc calibration and then remained visible for 6 seconds. During each exposure, respondents' faces were recorded with the web cam.

After the slide show, participants were shown the 32 pictures a second time in the same order as before. They were asked to rate their pleasantness ("I personally felt that the picture was . . . :") on a bipolar 11-point scale from "very unpleasant" (-5) to "very pleasant" (+5).

Finally, participants were asked to evaluate the questionnaire and the interview situation. After completing the questionnaire, they received their show-up fee from the interviewer and were bid good-bye. Overall, one session lasted about 50 minutes.

2.3 Analysis of video recordings

All video recordings were analyzed with the EMO Scan software, resulting in one net score per video frame that indicates valence of facial expressions. To account for differences in physiognomy (e.g., differences in the appearance of faces due to wrinkles, relative position of facial features such as eyebrows, etc.), valence scores were baseline-corrected: The mean valence during the baseline period—that is, the 3 seconds of black screen preceding each TV commercial and the 1 second of black screen preceding each IAPS picture—was subtracted from the valence scores. Thus, valence scores should be interpreted as valence compared to baseline facial expressions, so they indicate in which direction and to what degree a scene or a picture changes the valence of the viewers' emotional experience relative to immediately before the exposure.

3 Results

3.1 IAPS

With the exception of erotic pictures (which had been excluded from our subset), valence responses to IAPS pictures have been shown to be comparable between Chinese and Western respondents (Gong and Wang, 2016; Huang et al., 2015). This is confirmed in the present study by the respondents' mean ratings of the pictures (see Table 1): Positive pictures received mean valence rating of 2.73 on average, with a minimum mean valence of 1.91. Negative pictures received mean valence rating of -2.45 on average, with a maximum mean valence of -1.55. Thus, positive and negative pictures clearly elicit opposite responses in terms of self-reported valence on average. Our main question is to what extent we can predict normative picture valence (i.e., the valence according to IAPS manual) as well as subjective valence (as indicated by the indivdual picture ratings) from the results of the automatic analysis of facial expressions with EMO Scan.

To test this, valence scores during the 6-second exposure time of each picture were averaged, resulting in one baseline-corrected valence score per picture and respondent. Then, the individual mean valence scores were averaged across participants, resulting in one average valence score per picture. A numerically positive value (i.e., a value greater than zero) indicates that the picture, on average, induced an increase in valence relative to baseline, whereas a negative value suggests that the picture led to a decrease in valence. We can now evaluate for how many of the IAPS pictures we can predict normative valence from the sign of the mean EMO Scan valence scores. For comparison, for each percentage reported in the subsequent paragraph, we also report in parentheses the corresponding percentage from Garbas et al. (2013) that was achieved with German respondents.

Based on the sign of the mean valence scores of the 32 pictures, a hit rate of 87.5% (vs. 75.0% in German sample) is achieved (computed as the fraction of correct predictions; see Table 1). However, several pictures (e.g., pictures showing nice landscapes or sports scenes) result in rather small valence scores close to zero, indicating that valence of facial expressions hardly changed relative to baseline. Focusing on a subset of pictures that produced stronger responses might yield higher predictive accuracy. If the focus is narrowed on the 11 pictures with valence scores more extreme than zero +/- 1 standard deviation (SD of the 32 mean valence scores = 2.90), a hit rate of 100% is achieved (vs. 100% in German sample based on 7 pictures). Overall, the correlation between mean valence ratings for the IAPS pictures and mean valence scores is r = .75 (compared to r = .62 in German sample).[1]

Additionally, we compare the individual valence ratings for pictures with the results of the EMO Scan valence classifier because the perceived valence could differ both from normative valence from the IAPS manual and between individual respondents. Similar to Garbas et al. (2013) we estimated three linear models

$$y_{i,j} = \mathbf{x}_{i,j}^{\top}\boldsymbol{\beta} + \varepsilon_{i,j},$$

where $y_{i,j}$ is the rating of picture j of individual i, $\mathbf{x}_{i,j}$ is a vector of independent variables, and $\varepsilon_{i,j}$ denotes the idiosyncratic error.

In all models, the vector $\mathbf{x}_{i,j}$ contains the averaged valence per picture for each respondent and the interaction of the averaged valence and a dummy variable ("neg") indicating whether the normative valence of a picture according to the IAPS manual is positive or negative. Additionally, we controlled for the valence classification according to the IAPS manual using a dummy variable for negative pictures in model (1).

In the model (2) we included a picture fixed effect instead of the normative IAPS classification. Finally, we controlled for individual unobserved heterogeneity in model (3) estimating a typical fixed effects model. The standard errors shown in Table 2 are cluster-robust controlling for individual error terms being correlated.

Regression results in Table 2 are almost identical to the results from Garbas et al. (2013). The signs of all coefficients are the expected ones. The average valence inferred by EMO Scan is positively connected with the average rating, which means that a higher valence measured with EMO Scan is accompanied with higher, more positive, individual ratings, and vice versa. The explanatory power of the valence variable hardly differs for positive and negative classified pictures since the interaction

[1] Note, however, that mean valence scores for the German sample were computed based on a global calibration phase before picture exposure, whereas for the Chinese sample in the present study each picture was calibrated based on individual calibration phases that preceded *each* picture. A reanalysis of the data from Garbas et al. (2013) according to the same principle of individual, picture-specific calibration yields the following, slightly higher accuracies for the German sample: a hit rate for all pictures of 78.1%, and a correlation between ratings and valence scores of r = .70.

				Predictions	
Motive	Normative valence (acc. to IAPS)	Average rating (-5 to +5)	Mean valence of facial expressions (SD=2.90)	Valence in correct direction?	Valence in correct direction for pictures with values more extreme than ±1 SD?
Baby	POS	3.46	7.49	1	1
Dollars	POS	3.13	3.69	1	1
Seal	POS	2.72	1.97	1	
Skiing	POS	2.62	-2.17	0	
Rabbits	POS	2.77	4.38	1	1
Sunset	POS	2.98	-0.27	0	
Skydivers	POS	2.72	0.23	1	
Rafting	POS	2.29	0.77	1	
Water slide	POS	2.95	1.20	1	
Ice cream	POS	2.77	1.41	1	
Fireworks	POS	2.89	1.66	1	
Water skiing	POS	2.38	0.35	1	
Puppies	POS	2.97	1.72	1	
Couple in ocean	POS	2.6	2.92	1	1
Jaguars	POS	1.91	0.64	1	
Kissing couple	POS	2.57	0.63	1	
Cockroaches	NEG	-3.19	-1.47	1	
Garbage	NEG	-2.67	-0.99	1	
Spider on arm	NEG	-1.78	-3.67	1	1
Snakes	NEG	-3.35	0.14	0	
Rat	NEG	-3.4	-4.91	1	1
Aggressive dog	NEG	-1.55	-3.66	1	1
Headache	NEG	-1.87	-0.71	1	
Car wreck	NEG	-2.8	-4.18	1	1
Sinking ship	NEG	-2.94	-3.59	1	1
Burning stuntman	NEG	-1.96	-2.67	1	
Exhaust fumes	NEG	-1.61	0.18	0	
Skulls	NEG	-3.29	-3.18	1	1
Dentist treatment	NEG	-2.07	-5.94	1	1
Boy in despair	NEG	-1.92	-0.73	1	
Soldiers	NEG	-2.38	-1.95	1	
Spider	NEG	-2.49	-2.66	1	
HITRATES				**0.88**	**1.00**

Table 1: Predictions of normative valence for 32 IAPS pictures based on inferred valence of facial expressions.

	Rating		
	(1)	(2)	(3)
valence(mean)	0.005***	0.004**	0.004**
	(0.002)	(0.002)	(0.002)
valence(mean)*neg	-0.005*	-0.004	-0.004
	(0.003)	(0.003)	(0.003)
intercept	2.731***	-3.228***	–
	(0.075)	(0.134)	–
neg	-5.2***	–	–
	(0.138)	–	–
control picture	–	yes	yes
		(partly sig.)	(partly sig.)
control respondent	–	–	yes
	–	–	(partly sig.)
Observation	6449	6417	6209
adj. R-squared	0.674	0.698	0.729

Clustered standard errors in parentheses
*** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$

Table 2: Results of regression analysis: Linear estimation of individual ratings by using individually averaged valence from EMO Scan and information from IAPS manual.

term is not significant (or only at the ten percent level) and additionally, the sum of the coefficients for valence and the interaction term equals zero. Similar to the German validation study the coefficient for valence is highly significant in all models and robust towards further control variables. This means that valence measured by EMO Scan explains variance in picture ratings beyond the variance explained by normative IAPS classification also for Chinese respondents.

3.2 TV commercials

Participants have been exposed to five different TV commercials, selected to elicit the following responses:[2]

(a) chocolate bar (positive, amusement)

(b) baby care (positive, warmth, tenderness)

(c) sweets (mixed, amusement but possibly also rejection)

(d) automotive (negative, boredom)

(e) anti drink & drive campaign (negative, shock, sadness)

Valence of facial expressions during ad exposure is compared to self-reported overall pleasantness of the TV commercial (11-point bipolar scale), overall ad liking (5-point bipolar scale) and agreement with the statement "This commercial makes me like the product featured in the commercial" (5-point unipolar scale).[3]

Besides averaging valence of facial expressions during exposure, two more options for deriving a summary affect metric over time are proposed. For that, individual valence results are aggregated to 1 Hz. That is, the 25 values corresponding to 1 second of the exposure (as videos were recorded with 25 frames per second) are averaged, so that only one average valence value per second remains. Based on the 1-Hz data sets, the "peak-end value" of each respondent for a particular commercial is computed. It is based on the peak-end rule (Fredrickson and Kahneman, 1993), according to which global evaluations of affective experiences can be "predicted by an unweighted combination of the most extreme affect experienced during the episode and by affect at the end moments" (p. 54). Accordingly, we compute the peak-end value by averaging the most extreme 1-Hz valence score (that shows the greatest absolute difference from zero) and the average valence of the last 5 seconds of stimulus exposure (including 1 second of black screen). The second, related score for aggregating valence is "slope", defined as the regression coefficient of the linear trend across all seconds from stimulus onset to end, including the first second of the concluding black screen. The idea is based on a finding by Ariely (1998) that shows that, at least for negative experiences, trend is an important predictor of overall retrospective evaluation. Average values of the three ratings (pleasantness, ad liking, product liking) and the three valence scores (average, peak-end value, slope) for the five TV commercials are reported in Table 3.[4]

Results from facial expression analysis and self-report agree when it comes to distinguishing negative from positive clips: The commercials with the lowest pleasantness ratings also have the lowest facial valence. However, from the two positive clips, the positive-warm clip received higher ratings than the funny clip, while facial valence results point in the opposite direction. This is most likely due to the intensity of emotion expression, which is naturally stronger when laughing in amusement than when possible showing only a slight quiet smile when exposed to warm, tender content. Also, the mixed emotion ad produced quite high facial valence scores but received relatively low ratings. Here, strong amusement responses may have outweighed more subtle signs of confusion or rejection. These discrepancies need to be kept in mind when interpreting EMO Scan results and strong amusement responses must be treated with a certain level of caution.

Comparing the different—albeit highly correlated—aggregate facial valence scores, it needs to be noted that the pattern described above holds for all of them. But the peak-end value offers the highest degree of differenti-

[2] The ads have been selected upon recommendation of two Chinese market research experts.

[3] As said above, many more diagnostic questions about the ad have been asked. The three items were selected from the frame questionnaire for their presumed connection to the valence of the experience during ad exposure.

[4] Product liking is missing for the anti drink & drive TV commercial. This is because the commercial has been taken from a public campaign against drunk driving, so there was no advertised product. Note that this may also explain why overall ad liking is relatively high for this spot: While the experience during exposure may have been unpleasant, with shocking and sad events depicted, most viewers probably support the well-intended purpose of the ad.

	Questionnaire results			**Facial expression results**		
	Average pleasantness rating*	Average overall ad liking**	Average agreement "made me like product"***	average valence	peak end value	slope
Chocolate bar (positive-amusement)	2.57	3.15	3.90	6.78	24.16	0.40
Baby care (positive-warm)	2.67	3.29	3.87	-1.39	1.60	-0.01
Sweets (mixed)	1.84	2.64	3.42	3.62	14.86	0.90
Automotive (negative-boredom)	1.43	2.48	3.33	-6.75	-7.09	-0.20
Anti drink & drive (negative-sadness)	-0.18	2.89		-6.03	-3.17	-0.11

* on a scale from -5=very unpleasant, to 5=very pleasant

** on a scale from 1=disliked very much, 5=liked very much (inverted relative to original scale)

*** on a scale from 1=do not agree at all, to 5=totally agree

Table 3: Average rating results and time-aggregated results from facial expression analysis for the five test commercials.

ation between the ads and the highest face validity, while its agreement with self-report in terms of ranking is the same as for the average valence.

In Table 3 only aggregated scores across time are considered. This, however, means that a lot of information is lost, especially the temporal resolution that is considered one of the greatest strengths of automated analysis of facial expressions. To evaluate the course of emotional valence changes over time of exposure, we compare EMO Scan results and scene-to-scene ratings for each commercial.

Figure 1 shows the mean valence time course resulting from EMO Scan along with the mean scene-to-scene pleasantness ratings. Each new scene in the TV commercial has been shown to respondents post-hoc as pictures and rated on a bipolar 11-point scale from "very unpleasant" (-5) to "very pleasant" (+5).

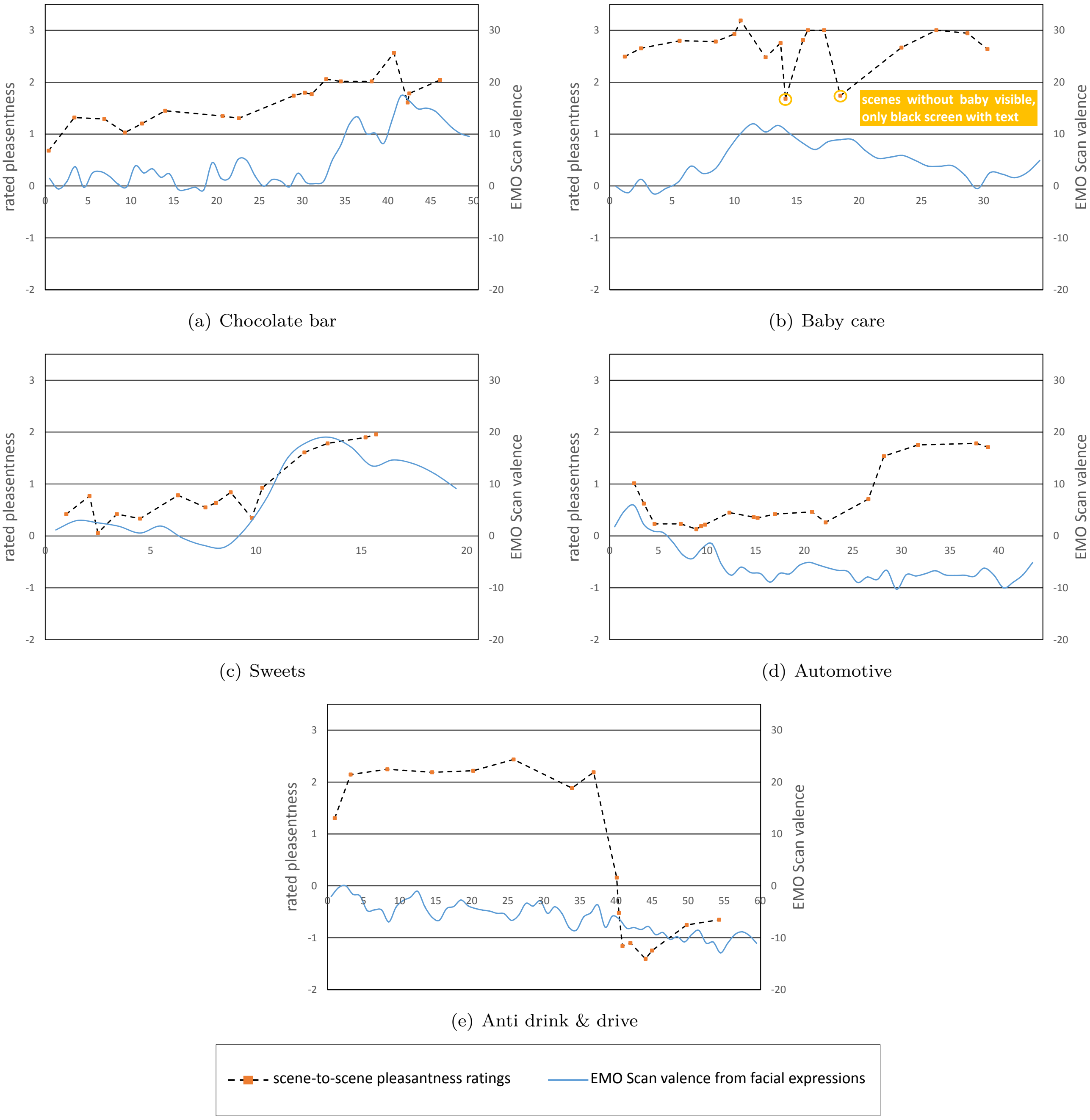
For the humorous chocolate bar and sweets ads, these scene-to-scene ratings and the EMO Scan valence clearly show a similar time course. For the baby care and the anti drink & drive ad there seems to be a common pattern, at least when ignoring two short scenes in the baby care commercial in which only text in front of black background and no baby is shown (highlighted in yellow in the respective figure). An exception is the automotive commercial where we observe opposing trends.

The patterns that can be observed in Figure 1 are confirmed by the correlations between scene-to-scene ratings and EMO Scan valence (see Table 4). The Pearson correlation coefficient for the car commercial is negative. As described, the commercial is quite repetitive and probably boring, so the lack of positive facial response is not surprising; however, in the last third of the commercial, a car and the car brand become visible. Cars are a desired good in China, with ownership rapidly increasing (ChinaAutoWeb, 2014), so respondents may have been reluctant to give low pleasantness ratings for the respective scenes, leading to a discrepancy between facial responses and ratings.

Clip	Correlation EMO Scan valence with scene-to-scene ratings
Chocolate bar	0.59
Baby care	
- all scenes	-0.10
- without black-screen scenes	0.31
Sweets	0.83
Automotive	-0.27
Anti drink & drive	0.61

Table 4: Pearson correlation coefficients between scene-to-scene valence ratings and EMO Scan valence. For computation we matched scene-to-scene rating and EMO Scan valence according to the time stamp and correlated the rating of each scene with the average EMO Scan valence scores observed during the second that immediately followed the respective scenes.

Additionally, the overall correlation for baby care is negative, although rather small. Considering the ratings for the baby care commercial in more detail, the two scenes in the middle of the commercial with no baby but only black screen and text visible stand out: they receive clearly lower pleasantness ratings than all the other scenes. When these two scenes are eliminated the correlation for the baby care commercial increases from -0.1 to 0.31. It makes sense that the two scenes score lower in pleasantness when rated in isolation and out of context, compared to scenes depicting a cute baby. When viewed within the flow of the commercial video, in con-



(a) Chocolate bar

(b) Baby care

(c) Sweets

(d) Automotive

(e) Anti drink & drive

Figure 1: Scene-to-scene pleasantness rating and corresponding valence time course inferred from facial expression.

trast, the two text screens seem to be well integrated and accepted, and do not interrupt the overall positive facial expressions.

4 Discussion

The goal of the present study was to evaluate whether automatic facial expression analysis with EMO Scan offers valid results also for Asian respondents. Focusing on IAPS, valence scores automatically inferred from facial expressions are lower compared to a previous study with German participants. Nevertheless, when using the sign of the score to predict whether the normative valence of an IAPS picture is positive of negative, an impressive hit rate of 87.5% is achieved, which can be increased to 100% when focusing on the subset of stimuli that produce relatively strong facial responses. Thus, accuracy is even higher than in the previous study with a German sample. Surprisingly, given that cross-cultural studies find that expression of negative emotion is socially undesirable in Asian collectivistic cultures (e.g., Biehl et al., 1997), higher accuracy is due to better identification not only of positive but also of negative IAPS pictures. Thus, at least in our computer-based interview procedure, negative facial responses seem to have been quite clearly discernible for the automatic system.[5]

Further research should focus on increasing the variability of physical features in the respondents face including skin color and presence of facial hair, because weaker contrast in the face may make it more difficult to detect facial features such as eyebrows. Re-training of the computer-vision algorithms with more population-specific training data might become necessary in these cases.

Also, it would be interesting to extend on these findings and explore cultures in which the facial expression of emotions is potentially more controlled than in China, such as Japan (Biehl et al., 1997; Safdar et al., 2009).

5 Conclusion

We showed that EMO Scan produces sufficiently high predictive validity also for Chinese participants. Similar to the validation study with Western respondents it should be refrained from interpreting small valence values close to zero because of their limited validity. Considering the results from the TV commercials, we showed that EMO Scan is associated with, e.g., ad liking. Also the temporal change of the valence during the exposure reflects the scene-to-scene valence ratings self-reported by the participants. Thus, we conclude that EMO Scan can be used to infer emotional valence of responses to TV commercials in China, both in terms of overall response as well as on the individual scene level.

[5] As explained above, EMO Scan is based on extensive machine learning resulting in a classifier working on local feature patterns that do not necessarily have a semantic meaning (21025 features are evaluated per 30x30 image block). Therefore, we do not know whether facial responses in our present sample semantically differ from the previous study in Germany. Annotations of Action Units would be needed to explore possible differences in the way positive and negative valence is expressed in German vs. Chinese respondents. Because the total recording time is far too long for manual annotations, future work should either select a subsample of recordings for annotations or work with automatic AU detection software (for a promising approach, cf. Hassan et al., 2016).

Acknowledgements

We thank John Wittenbraker, Markus Iwanczok, Alice He and Vicky Lai for helpful input and support in study design, stimuli selection and data collection. We thank Regina Schreder for support in data analysis and graphic editing for the paper.

References

Ariely, D. (1998). Combining experiences over time: The effects of duration, intensity changes and on-line measurements on retrospective pain evaluations. *Journal of Behavioral Decision Making*, 11(1):19–45.

Biehl, M., Matsumoto, D., Ekman, P., Hearn, V., Heider, K., Kudoh, T., and Ton, V. (1997). Matsumoto and Ekman's Japanese and Caucasian Facial Expressions of Emotion (JACFEE): Reliability data and cross-national differences. *Journal of Nonverbal Behavior*, 21(1):3–21.

Camras, L., Chen, Y., Bakeman, R., Norris, K., and Cain, T. (2006). Culture, ethnicity, and children's facial expressions: A study of European American, Mainland Chinese, Chinese American, and adopted Chinese girls. *Emotion*, 6(1):103–114.

ChinaAutoWeb (2014). *China's auto fleet expands to 154 million vehicles.* Retreived from http://chinaautoweb.com/2014/11/chinas-auto-fleet-expands-to-154-million-vehicles/.

Darwin, C. (1872). *The Expression of the Emotions in Man and Animals.* John Murray, London. Retrieved from http://darwin-online.org.uk/.

Debraix, C. (1995). The impact of affective reactions on attitudes towards advertisements and the brand: A step towards ecological validity. *Journal of Marketing Research*, 17(4):470–479.

Eid, M. and Diener, E. (2001). Norms for experiencing emotions in different cultures: Inter- and intranational differences. *Journal of Personality and Social Psychology*, 81(5):869–885.

Ekman, P. (1972). Universals and cultural differences in facial expressions of emotions. In Cole, J., editor, *Nebraska Symposium on Motivation*, pages 207–282. University of Nebraska Press, Lincoln, NB.

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

Ekman, P. and Oster, H. (1979). Facial expressions of emotion. *Annual Review of Psychology*, 30:527–554.

Elfenbein, H. and Ambady, N. (2002). On the universality and cultural specificity of emotion recognition: A meta-analysis. *Psychological Bulletin*, 128(2):203–235.

Fontaine, J., Scherer, K., Roesch, E., and Ellsworth, P. (2007). The world of emotions is not two-dimensional. *Psychological Science*, 18(12):1050–1057.

Fredrickson, B. and Kahneman, D. (1993). Duration neglect in retrospective evaluations of affective episodes. *Journal of Personality and Social Psychology*, 65(1):45–55.

Garbas, J., Ruf, T., Unfried, M., and Dieckmann, A. (2013). Towards robust real-time valence recognition from facial expressions for market research applications. In *Proceedings of 2013 Humaine Association Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 570–575.

GfK (2013). *GfK boosts ad success with ground-breaking application of facial coding* [Press release]. Retrieved from http://www.gfk.com/de/insights/press-release/gfk-boosts-ad-success-with-ground-breaking-application-of-facial-coding-1/.

Gong, X. and Wang, D. (2016). Applicability of the International Affective Picture System in Chinese older adults: A validation study. *PsyCh Journal*, 5:117–124.

Hassan, T., Seuss, D., Wollenberg, J., Garbas, J., and Schmid, U. (2016). A practical approach to fuse shape and appearance information in a Gaussian facial action estimation framework. In *Proceedings of 22nd European Conference on Artificial Intelligence (ECAI 2016), 285*, pages 1812–1817.

Huang, J., Xu, D., Peterson, B., Hu, J., Cao, L., Wei, N., Zhang, Y., Xu, W., Xu, Y., and Hu, S. (2015). Affective reactions differ between Chinese and American healthy young adults: a cross-cultural study using the international affective picture system. *BMC Psychiatry*, 15:60.

Ipsos (2014). *Ipsos selects Realeyes to provide facial coding emotional response metrics for deeper insights into consumer emotions* [Press release]. Retrieved from https://www.ipsos.com/sites/default/files/news_and_polls/2014-04/6467.pdf.

Jack, R., Blais, C., Scheepers, C., Schyns, P., and Caldara, R. (2009). Cultural confusions show that facial expressions are not universal. *Current Biology*, 19(18):1543–1548.

Küblbeck, C. and Ernst, A. (2006). Face detection and tracking in video sequences using the modified census transformation. *Image and Vision Computing*, 24(6):564–572.

Lang, P., Bradley, M., and Cuthbert, B. (2008). International affective picture system (IAPS): Affective ratings of pictures and instruction manual. Technical report A-8, University of Florida, Gainesville, FL.

Lin, T. (1983). Psychiatry and Chinese culture. *Western Journal of Medicine*, 139(6):862–867.

Martinez, B. and Valstar, M. (2016). Advances, challenges, and opportunities in automatic facial expression recognition. In Kawulok, M., Celebi, E., and Smolka, B., editors, *Advances in Face Detection and Facial Image Analysis*, pages 78–80. Springer, Cham, Switzerland.

Masuda, T., Ellsworth, P., Mesquita, B., Leu, J., Tanida, S., and de Veerdonk, E. V. (2008). Placing the face in context: Cultural differences in the perception of facial emotion. *Journal of Personality and Social Psychology*, 94(3):365–381.

Mesquita, B. (2003). Emotions as dynamic cultural phenomena. In Davidson, R., Goldsmith, H., and Scherer, K., editors, *The Handbook of the Affective Sciences*, pages 871–890. Oxford University Press, New York.

Pantic, M. and Rothkrantz, L. (2003). Toward an affect-sensitive multimodal human-computer interaction. *Proceedings of the IEEE*, 91(9):1370–1390.

Potter, S. (1988). The cultural construction of emotion in rural Chinese social life. *Ethos*, 16(2):181–208.

Prkachin, K. (2009). Assessing pain by facial expression: Facial expression as nexus. *Pain Research and Management*, 14(1):53–58.

Rasting, M., Brosig, B., and Beutel, M. (2005). Alexithymic characteristics and patient-therapist interaction: A video analysis of facial affect display. *Psychopathology*, 38(3):105–111.

Ruf, T., Ernst, A., and Küblbeck, C. (2011). Face detection with the sophisticated high-speed object recognition engine (SHORE). In Heuberger, A., Elst, G., and Hanke, R., editors, *Microelectronic Systems: Circuits, Systems and Applications*, pages 237–246. Springer, Berlin.

Russell, J. (1980). A circumplex model of affect. *Journal of Personality and Social Psychology*, 39(6):1161–1178.

Russell, J. (1991). Culture and the categorization of emotions. *Psychological Bulletin*, 110(3):426–450.

Russell, J. (2003). Core affect and the psychological construction of emotion. *Psychological Review*, 110(1):145–172.

Safdar, S., Friedlmeier, W., Matsumoto, D., Yoo, S., Kwantes, C., Kakai, H., and Shigemasu, E. (2009). Variations of emotional display rules within and across cultures: A comparison between Canada, USA, and Japan. *The Canadian Journal of Behavioural Science*, 41(1):1–10.

Schapire, R. and Singer, Y. (1999). Improved boosting algorithms using confidence-rated prediction. *Machine Learning*, 37(3):297–336.

Shuman, V., Sander, D., and Scherer, K. (2013). Levels of valence. *Frontiers in Psychology*, 4(261):1–17.

ten Brinke, L. and Porter, S. (2012). Cry me a river: Identifying the behavioral consequences of extremely high-stakes interpersonal deception. *Law and Human Behavior*, 36(6):469–477.

Unfried, M. and Iwanczok, M. (2016). Improving signal detection in software-based facial expression analysis. *GfK Verein Working Paper Series, No. 1 / 2016*.

WPP (2012). *Millward Brown and Affectiva deliver new way to test emotional responses to ads* [Press release]. Retrieved from http://www.wpp.com/wpp/press/2012/jan/30/millward-brown-and-affectiva-deliver-new-way/.

Yuki, M., Maddux, W., and Masuda, T. (2007). Are the windows to the soul the same in the East and West? Cultural differences in using the eyes and mouth as cues to recognize emotions in Japan and the United States. *Journal of Experimental Social Psychology*, 43(2):303–311.

Zeng, Z., Pantic, M., Roisman, G., and Huang, T. (2009). A survey of affect recognition methods: Audio, visual, and spontaneous expressions. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 31(1):39–58.

Zhang, H., Zhang, X., Liang, X., Lai, H., Gao, J., Liu, Q., and Wang, H. (2015). Introduction on the Emotion-Will Overcoming Therapy (EWOT): A novel alternative approach of psychological treatment from Chinese medicine. *Chinese Medicine*, 6(2):75–82.