Automated Measurement of Spontaneous Surprise

Bart Joosten[1], Eric Postma[1], Emiel Krahmer[1], Marc Swerts[1], Jeesun Kim[2]

[1]*TiCC, Tilburg University, Tilburg, The Netherlands*

[2]*MARCS Auditory Laboratories, University of Western Sydney, Sydney, Australia*

1. Introduction

Humans exhibit a rich variety of facial expressions reflecting their emotional state. As pointed out by [6], automatically measuring facial expressions to infer emotional states is a challenging task, amongst others because there are no uniquely defined representational units of facial expressions for the basic emotions. In natural circumstances, spontaneous facial expressions tend to be subtle and highly context-dependent, which makes their formal representation very difficult. We have adopted an approach that performs computational analyses on video data acquired from carefully designed experiments for the measurement of facial expressions (see, e.g. [4]). This paper reports on our approach to measure spontaneous surprise. Two main obstacles in automatically measuring spontaneous surprise are of a behavioral and a technical nature. The behavioral obstacle is to elicit and identify spontaneous surprise. The technical obstacle is to develop a computational method to measure the facial expressions associated with spontaneous surprise. In order to deal with the first obstacle two of the authors (MS and JK) developed a novel experimental paradigm to elicit spontaneous surprise in humans. The paradigm is described in Section 2. The second obstacle is addressed by the Spontaneous Surprise Measurement method presented in Section 3. The setup for the experimental evaluation of the method is described in Section 4 and the results are reported in Section 5. Finally, Section 6 is a concluding discussion.

2. Eliciting nonverbal expressions of surprise

According to [3], surprise is prototypically displayed by three facial actions: raising of the eyebrows, opening of the mouth, and widening of the eyes. However, when researchers try to elicit surprise expressions in participants in experimental studies, the prototypical visual actions are rarely shown (e.g., [7]). This may be due to the lack of a proper experimental setting. We tried to improve the setting by designing a memory task experiment in which lexically similar utterances are elicited in a neutral and in a surprise context. To keep participants unaware of our intentions, they are presented with a cover story that the experiment measures the how context and reading words aloud affects the number of words memorized. The experiment continues in three stages: imagination, verbalization, and recall.

1. Imagination stage. Participants imagine words that fit a specific context (e.g. “organs of the human body” in the neutral condition versus “favorite food items among Dutch children” in the surprise condition)
2. Verbalization stage. Participants read aloud each of a sequence of 10 words (of either of the two contexts) displayed on a screen. In both conditions, the target word “liver” is one of displayed words.
3. Recall stage. Participants have to recall as many words as possible.

Crucially, we elicit the target word (liver) in two conditions: a neutral condition in which the word clearly fits in the context “organs of the human body” and a surprise condition in which the word is highly unexpected “favorite food items among Dutch children” (liver is clearly not a favorite food item for Dutch children). Participants see the words on a screen using a hidden camera that is positioned behind the computer screen so that facial expressions can be clearly captured. We made such hidden recordings of 27 Dutch participants all of Dutch descent (students at Tilburg University who took part for course credit). In most participants, the paradigm results in clearly different behavior in both experimental conditions (neutral versus surprise).

3. The Spontaneous Surprise Measurement method

The Spontaneous Surprise Measurement (SSD) method aims at measuring surprise by focusing on the first of the three actions proposed by Ekman [3]: the raising of the eyebrows. To that end, the SSD method consists of three steps: the identification of landmarks, the texture analysis of the eyebrow region, and the classification of surprise. The landmark-identification step is realized by using the Constrained Local Model [8]. Given a video frame, it returns the locations of a number of predefined facial landmarks, such as, the nose tip, corners of the eyes, and the eyebrows. The texture analysis is restricted to an image patch covering the facial region of the eyebrows and is performed using a multi-scale Gabor filter-bank [1]. For each image position (pixel), the Gabor filter-bank returns N x M energy values representing the presence of oriented visual structure of a certain thickness (spatial frequency), where N represents the number of orientations and M the number of spatial frequencies. Finally, the classification maps the (aggregated) energy values onto the binary classes *frown* and *non-frown* and subsequently to *surprise* and *no surprise*.

4 Experimental set-up

In this section we describe the data set (4.1) and the experimental settings of the three components of the SSM method (4.2-4.4).

4.1 Data set

Our dataset was collected in the experiment described in Section 2. All participants were recorded on frontal-face video in 24-bit color (resolution 640 by 480 pixels at a frame rate of 30 fps). The video fragments selected for our data set started immediately following the presentation of the neutral and surprise-evoking target. This resulted in a set of 54 videos (27 for each condition). The recordings are approximately four-seconds long, varying from 119 to 122 frames in total. Figure 1 shows an example of one participant in both conditions, directly after verbalizing the target word.

Careful visual inspection of the video fragments revealed five distinctive facial expressions in the surprise condition, viz. (1) eyebrow frowning, (2) eyebrow raising, (3) widening eyes (4) mouth opening, and (5) brief head retraction. Of these expressions, (2), (3) and (4) correspond to Ekman's actions. The presence and prominence of each of the expressions differed from participant to participant, but eyebrow frowning (1) was the most prevalent expression overall. Therefore, we manually annotated the frames for all participants who displayed the eyebrow frown (11 in total). The frowns had an average length of 17.1 frames σ^2=16.1. In addition to the labeling of frown and no-frown frames, we labeled consecutive sequences of frames as neutral or surprise. All frames are labeled neutral, except for those consisting of three consecutive frames labeled "frown". These sequences are labeled "surprise".

Figure 1. Illustration of a participant recorded in the neutral condition (left) and surprised condition (right).

4.2 Landmark identification

We employed FaceTracker as described in [8] on each video in the data set. The FaceTracker software contains a pre-trained CLM which automatically places a grid of 66 facial landmarks on each video frame. The output of the analysis is a collection of 66 coordinates per frame. After manual inspection 8 videos appeared to have insufficient (correct) fittings. These were excluded from further analysis. We selected a region of interest (ROI) based on the center between the inner eyebrow landmarks. At this center we extract a patch of 201 by 201 pixels. To compensate for in-plane rotations of the head, in each frame the patch is rotated so that both eyes are centered at the same row of pixels. Subsequently, the rotated patch is down-sampled to a size of 51 x 51 pixels.

4.3 Texture analysis

Frown detection is based on the output of Gabor filters convolved with the patches extracted around the inner eyebrow landmarks. For this purpose we constructed a 6 by 12 log-Gabor filter bank consisting of filters with 6 scales and 12 orientations. We used Kovesi's [5] Matlab tools to compute the convolutions (minimum wavelength 3 pixels, scaling factor of 1.8 between successive filters, and filter bandwidth 1.45). A margin of 5 pixels wide is excluded from the convolved image patch to remove image border artifacts. The Gabor energy for each scale and orientation combination is computed by summing over the entire convolved image. As a consequence, our feature space consists of 72 dimensions, each representing the total Gabor energy of a filter at a specific scale and orientation. A large value in any of these scale-orientation combinations signals the presence of a visual contour at a particular scale (thickness) and orientation. Given the specific thickness and orientation of frowns, we expect the Gabor filters with corresponding scale and orientation to generate large energy values.

4.4 Classification

The manual annotation of frown and no-frown frames resulted in 205 frown frames and 5343 no-frown frames. In order to automatically differentiate between the two classes of frames we explored two techniques for classification, viz. Linear Discriminant Analysis (LDA) and Support Vector Machines (SVM). Training and testing was performed on a leave-one-subject-out validation scheme. This entails that we first trained our classifiers on a balanced number of frames (i.e., all 'frown frames' and an equal number of randomly selected 'non-frown frames') of both classes, excluding one participant's frames and subsequently tested the classifiers on the left out participant. This procedure is repeated for each participant and the summed score of all the iterations represents the classifier's predictive value. Since we have only 205 frown frames each training set contains at most 410 frames.

LDA linearly maps labeled instances from our feature space to a 1-dimensional space. The linear transformation matrix is constructed in such a way that it tries to maximize the separation between the classes. A maximal separation in one dimension, readily leads to the identification of a threshold value separating both classes. Unseen instances are mapped to the same 1-dimensional space and classified using the aforementioned threshold.

The SVM classifier operates on the 72-dimensional Gabor feature space. We used the freely available LIBSVM [2] as our SVM implementation. We choose a radial basis function as a kernel and used a grid search to find optimal values for the parameters C and γ. In our case $C = 2^{15}$ and $\gamma = 2^{-5}$.

We evaluate the performance of surprise classification performance of the LDA and SVM classifier separately.

5 Results

Application of the SSM method to the videos in the data set yielded results at three levels, corresponding to the three components of the method. At the level of landmark detection, visual inspection revealed the FaceTracker to be quite successful in adequately positioning the 66 landmarks at the faces of the participants. In some cases, there were some misalignments along the cheek regions, but because the texture analysis does not rely on these

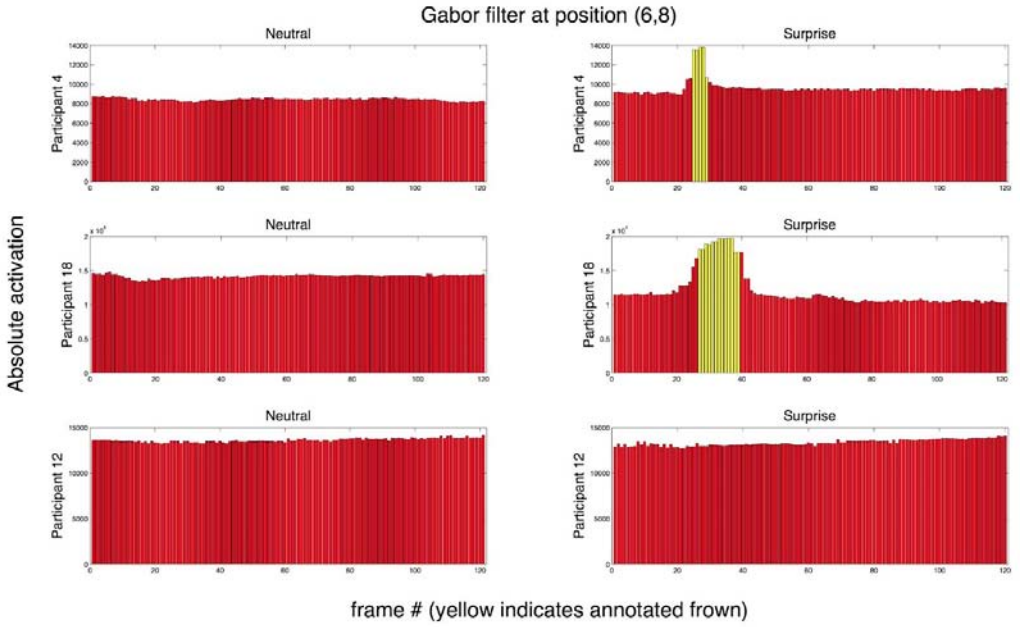

Figure 2. Gabor filter activation of filter at position (6,8) of three participants in both the neutral and surprise condition. Participants 4 and 18 clearly show a distinctive activation at the manually annotated 'frown frames' (depicted in yellow), while participants 12 lacks a frown of surprise.

landmarks, they do not affect the overall performance of the SSM method. The performance of the texture analysis component is illustrated in Figure 2 for the video sequences of three participants in both conditions: participant 4 (top row), participant 18 (middle row), and participant 12 (bottom row). The six panels in the figure show the energy (output) value of one of the Gabor filters (the filter at orientation 6 and scale 8) as a function of frame number of the video. The videos of the neutral and surprise conditions are displayed on the left and right, respectively. All frames during which participants frown have been manually labeled and are indicated by yellow bars, all others by red bars. The frowning of participants 4 and 18 is associated with an elevated energy value of the Gabor filter. As a consequence, the filter output appears to signal frowning. Participant 12 does not frown in the surprise condition and, hence, does not result in an elevated Gabor energy value.

An impression of the current leaving-one-subject-out performances of the SSM method as determined with the LDA and SVM classifiers is given in Tables 1 and 2. Tables 1a and 1b contain the confusion tables listing the frown-detection and surprise-detection results respectively for the LDA classifier. Similarly, tables 2a and 2b contain the detection results for the SVM classifier. Overall, the confusion tables show that no-frowns are often detected as frowns and that neutral sequences are often detected as surprise sequences. The higher scores on the main diagonal indicate that the SSM method is reasonably successful in detecting spontaneous surprise.

Tables 1a and 1b. Frown and surprise classification performance for the LDA classifier.

		predicted	
		no-frown	frown
actual	no-frown	4040	1303
	frown	78	127

		predicted	
		neutral	surprise
actual	neutral	13	10
	surprise	5	18

Tables 2a and 2b. Frown and surprise classification performance for the SVM classifier.

		predicted	
		no-frown	frown
actual	no-frown	3677	1666
	frown	72	133

		predicted	
		neutral	surprise
actual	neutral	13	4
	surprise	10	19

6 Concluding discussion

We have developed a method for the measurement of Spontaneous Surprise that is reasonably successful at detecting frowning as part of one of the key facial actions for surprise (raising the eyebrows). Admittedly, our method was evaluated solely on Western-European participants. (An interesting experiment would be to evaluate our method on subjects with a different ethnical background). Further optimizing the texture-analysis and extending our method with motion information of the facial actions may lead to an improvement in performance. However, our aim is to yield improvement by incorporating domain knowledge about the facial features signaling surprise. More specifically, we want to extend the SSM method by including measurements of the two other facial actions proposed by Ekman, i.e., opening of the mouth, and widening of the eye. In addition, "mouth opening" and "brief head retraction" will also be examined.

References

1. Berezhnoy, I.J., Postma, E.O., Herik, H.J. van den (2007). Computer Analysis of Van Gogh's Complementary Colours. *Pattern Recognition Letters* **28**, 703-709.
2. Chang, C.-C., Lin, C.-J. (2011). LIBSVM : a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology* **2**,1-27. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm
3. Ekman, P., Keltner, D. (1997). Universal facial expressions of emotion. In U. Segerstrale, P. Molnar (Eds.) *Nonverbal communication: Where nature meets culture*, 27-46.
4. Joosten, B., Amelsvoort, M.A.A. van, Krahmer, E.J., Postma, E.O. (2011). Thin slices of head movements during problem solving reveal level of difficulty. In G. Salvi, J. Beskow, O. Engwall, S. Al Moubayed (Eds.), *Proceedings of the international conference on audio-visual speech processing* (AVSP 2011) 85-88.
5. Kovesi, P.D. (2012): *MATLAB functions for computer vision and image analysis* http://www.csse.uwa.edu.au/~pk/research/matlabfns/. Accessed February 2012.
6. Pantic, M., Rothkrantz, L. (2000). Automatic analysis of facial expressions: the state of the art, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **22**, 1424-1445.
7. Reisenzein, R., Bördgen, S., Holtbernd, T., Matz, D. (2006). Evidence for strong dissociation between emotion and facial displays: The case of surprise. *Journal of Personality and Social Psychology* **91**, 295-315.
8. Saragih, J., Lucey, S., Cohn, J. (2011) Deformable model fitting by regularized landmark mean-shifts, *International Journal of Computer Vision* **91**, 200-215.