How to Make a Robot Smile? Perception of Emotional Expressions from Digitally-Extracted Facial Landmark Configurations

Caixia Liu[1,2], Jaap Ham[1], Eric Postma[2], Cees Midden[1],
Bart Joosten[2], and Martijn Goudbeek[2]

[1] Human-Technology Interaction Group,
Department of Industrial Engineering and Innovation Sciences,
Eindhoven University of Technology, Eindhoven, The Netherlands
{c.liu,j.r.c.ham,c.j.h.midden}@tue.nl
[2] Tilburg Center for Cognition and Communication,
Tilburg University, Tilburg, The Netherlands
{e.o.postma,b.joosten,m.b.goudbeek}@tilburguniversity.edu

Abstract. To design robots or embodied conversational agents that can accurately display facial expressions indicating an emotional state, we need technology to produce those facial expressions, and research that investigates the relationship between those technologies and human social perception of those artificial faces. Our starting point is assessing human perception of core facial information: Moving dots representing the facial landmarks, i.e., the locations and movements of the crucial parts of a face. Earlier research suggested that participants can relatively accurately identity facial expressions when all they can see of a real human full face are moving white painted dots representing the facial landmarks (although less accurate than recognizing full faces). In the current study we investigated the accuracy of recognition of emotions expressed by comparable facial landmarks (compared to accuracy of recognition of emotions expressed by full faces), but now used face-tracking software to produce the facial landmarks. In line with earlier findings, results suggested that participants could accurately identify emotions expressed by the facial landmarks (though less accurately than those expressed by full faces). Thereby, these results provide a starting point for further research on the fundamental characteristics of technology (AI methods) producing facial emotional expressions and their evaluation by human users.

Keywords: Robots, Emotion, Facial expression, Facial landmarks, FaceTracker, Perception.

1 Introduction

To design robots or embodied conversational agents that can accurately display facial expressions indicating an emotional state, we need technology to produce those facial expressions, and research that investigates the relationship between those technologies and human social perception of those artificial faces. Our starting point is to assess

S.S. Ge et al. (Eds.): ICSR 2012, LNAI 7621, pp. 26–34, 2012.

human perception of core facial information as represented by the dynamics of facial landmarks: the locations and movements of the crucial parts of a face.

Humans are very good at social perception based on dot patterns. In a famous study by Johansson [2], participants watched videos of lights attached to the joints of walking people against an otherwise black background. Participants were able to identify familiar persons from their gait as reflected in the dynamics of the lights. We adopted the idea to use point lights to determine such core, but minimal information required to recognize emotional expressions.

Also, Bassili suggested that people can relatively accurately identity facial expressions when all they can see of a real face which was painted white stickers representing that core information (although less accurate than recognizing full faces) [1]. He performed such a study by painting the face of confederates black and applying a uniform pattern of about 100 white dots on their faces. He found participants to be quite well able to identify the six basic emotions (happiness, sadness, surprise, fear, anger and disgust) from the facial expressions as revealed by the dynamics of the white dots. His results are reproduced in table 1. The rows list the displayed emotions and the columns the emotions reported by the participants. The entries in the table represent the percentages of correct recognition of the emotions for the fully displaced faces (left of the slash) and their landmark representations (right of the slash).

Table 1. Bassili's results of emotion recognition (reproduced from [1])

		Reported emotion					
		Happiness	**Sadness**	**Fear**	**Surprise**	**Anger**	**Disgust**
Displayed emotion	**Happiness**	**31/31**	6/13	0/6	38/31	19/6	6/13
	Sadness	13/0	**56/25**	0/25	0/13	0/0	31/37
	Fear	0/0	0/19	**69/6**	13/25	6/13	12/37
	Surprise	0/0	0/6	6/0	**94/75**	0/6	0/13
	Anger	6/0	0/13	0/0	0/6	**50/6**	44/75
	Disgust	0/6	0/6	0/19	0/19	6/6	**88/57**

In the current study we investigated the accuracy of recognition of similar manipulated facial landmarks (compared to accuracy of recognition of full faces), but now used FaceTracker[1] software [4], see also [3], [5], [6], [7], [15] to produce the configurations of facial landmarks. The main research question is how well participants recognize emotional expressions from facial-landmark videos as compared to the full-face videos. The goal of our study is twofold. The first goal is to replicate Bassili's study with state-of-the-art facial-expression recognition software. The second goal is to determine if the facial landmarks generated by the facial-expression recognition software contain sufficient information to be able to recognize emotions. Furthermore, we will also extend earlier research by investigating the fundamentals of the perception of the full faces versus the facial landmarks. That is,

[1] FaceTracker is not commercially available. Jason Saragih provided us with his software upon our request.

we will investigate and compare valence judgments and arousal judgments that people make about the full faces and the facial landmarks.

2 Digital Extraction Landmarks and Hypothesis

To be able to present participants with facial-landmarks videos representing facial expressions, we need information about the locations and movements of the crucial parts of a face while that face is displaying facial expressions. These locations might be calculated based on models of human faces and facial expressions [9], [10], but they can also be extracted from real human facial expressions [3]. In the current research, we used FaceTracker software [4], [15] to extract this information from full-face videos of actors displaying basic emotions (happiness, sadness, fear, disgust and anger). Extracting this information from full-face videos is relatively complex. That is, human full faces are highly non-rigid objects causing their appearances to vary with expressions. FaceTracker uses a facial landmark registration algorithm to locate and track designated facial locations, so-called landmarks, e.g., the positions of eyebrows, eyes, nose, mouth, teeth and the profile of the full face. Applied to our video dataset, FaceTracker yields the coordinates of the 66 landmarks for each frame. We used these coordinates to create facial-landmark videos consisting of 66 white dots against a black background. Using these facial-landmark videos, we investigated whether people can identify facial expressions presented in a facial-landmark video as well as facial expression presented in a full-face video showing a facial expression on which that facial-landmark video was based. Future technology that produces facial expressions in robots and other artificial social agents can use this information to produce facial expression on an artificial social agent to ameliorate human-robot interaction, or interaction of humans with other artificial social agents.

In line with Bassili (1978), first, we hypothesize that people can relatively accurately identity facial expressions based on full-face videos showing full faces of human actors (H1). Second, we hypothesize that people can also relatively accurately identity facial expression based on facial-landmark videos showing faces consisting of moving dots (generated through face-tracking software; H2). Third, we hypothesize that people can identify facial expressions better based on full-face videos, than based on facial-landmark videos (H3). Fourth, we also expect that participant's judgments about the valence and arousal levels of emotions displayed by full-face videos will show a strong correlation with participant's judgments about the valence and arousal level of emotions displayed by facial-landmarks videos (H4).

3 Method

Participants and Design. Sixteen participants participated in this study. None of them were familiar with or involved in facial expression related topics. All participants spoke Dutch as their mother language, and were students at Eindhoven University of Technology. Their average age was 25.6 years old (*SD* =10.46). Each participant was presented ten full-face videos (two actors, one male and one female, each expressing five emotions in five separate videos), and also ten facial-landmark videos (based on two different actors, one male and one female, each expressing five emotions in five separate videos. These two actors should be different from the previous two actors who were used in full-face videos to prevent interference as a result of identification of the actor). Half of the observers identified the emotions of the full-face videos first and the facial-landmark videos second. The other observers identified the emotions of the facial-landmark videos first and the full-face videos second. Overall, all the videos were shown to all the participants the same number of times, thereby counterbalancing the stimuli of the experiment. The design was a within-subject design (each participant was confronted with both conditions namely full-face videos and facial-landmark videos) and the dependent variable was recognition accuracy or classification performance. So, even though one participant saw different actors for the full-face videos and the facial-landmark videos, overall, all participants saw the same set of actors for the full-face videos and the facial-landmark videos. Thereby, the final results were not influenced by the different acting skills of different actors. The full-face videos of actors that exhibit emotional expressions were part of the GEMEP corpus [11], [12].

Stimulus Materials. For the full-face videos, we used four actors (two male and two female). Of each actor, we used five short videos (average length is 2 seconds), each showing the face and upper torso of an actor, while the actor acted as if he or she experienced an assigned emotion. That is, for each actor, we used five videos representing either happiness, sadness, fear, disgust, or anger.

Based on these full-face videos, we constructed the facial-landmark videos by applying the FaceTracker software to generate 66 landmarks consisting of locations indicating eyebrows, eyes, nose, mouth and the face profile, based on which we could construct a facial-landmark video with white points on a black background. Each facial-landmark video was based on one full-face video. So, we employed four (different actors) times five (different emotions) full-face videos of actors expressing emotions, and four times five facial-landmark videos of the same emotions.

Within each video (full-face video or facial-landmark video), we arranged video segments such that the emotion expression was displayed three times. Each participant was shown full-face videos of two actors, for each of which five videos were shown (expression the five basic emotions), and also three times five facial-landmark videos based on the other two actors in our set of four actors (to prevent interference as a result of identification of the actor).

Procedure. Participants participated individually, in a cubicle that contained a computer and a display. All instructions and stimulus materials were shown on the computer display, and the experiment was completely computer controlled. Each participant was instructed that he or she would be shown several short videos of faces expressing emotions, and that sometimes it would be the full-face video, and sometimes it would be the facial-landmark video. Also, participants were instructed that after each video they would be asked three questions about the emotion expressed by the face. Each of these three questions was explained. Each participant was presented ten full-face videos, and, on different screens, also ten facial-landmark videos (see Figure 1). Half of the observers were presented with the full-face videos first and the facial-landmark videos second. The other observers were presented with the facial-landmark videos first and the full-face videos second. Each set of five emotions displayed those five videos in a different random order.

For each of the videos, participants were first shown the video and then, on the next page, asked the three questions. In the first question, the participant was asked to identify the emotions expressed in the video by selecting one of six options ('happiness', 'sadness', 'fear', 'disgust', 'anger' and 'don't know'). In the second and third questions, the participant was asked to rate the valence level of the expressed emotion (1 = negative, to 7 = positive, or 'don't know'), and the arousal level of the expressed emotion (1 = low arousal, to 7 = high arousal, or 'don't know'). All the questions could only be answered one by one and a participant could not return to an earlier question.

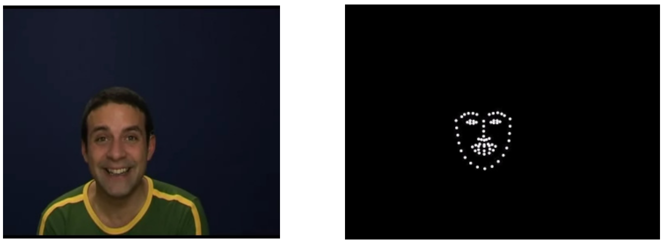

Fig. 1. An example of a frame of a full-face video and a frame of a facial-landmark video

4 Results

Under ideal circumstances, the full-face videos should yield high accuracy in the recognition of expressed emotions. If this high level of accuracy is achieved, it is possible to compare accuracy in the recognition of specific emotions on the basis of facial-landmark videos, and on the basis of full-face videos.

Table 2. Participant's identification of emotion displayed on full-face videos or facial-landmark videos

		Reported emotion					
		Happiness	Sadness	Fear	Anger	Disgust	Don't know
Displayed emotion	**Happiness**	**91/53**	0 / 3	3 / 6	0 / 9	0 / 0	6 /28
	Sadness	0 / 3	**66/38**	6 /16	13/ 6	0 / 9	16/28
	Fear	0 /22	13/9	**78/ 9**	0 / 3	9 /16	0 /41
	Disgust	0 / 9	29/16	3 / 9	**65/13**	0 /16	3 /38
	Anger	0 /16	0 / 0	0 / 9	0 / 6	**100/56**	0 /13

As can be seen in Table 2, the numbers represent the percentage of participants who responded with the column label when shown a video of the emotion described by the row label; Numbers on the left of the slash are for responses to the full-face video, and those to the right are for the response to the facial-landmark videos. Each row represents the responses of 16 participants in each of the two conditions.

However, the results of our full-face video conditions did not always reach this ideal. The average percentage of accuracy across the five emotions was 80%, and ranged from 100% (anger) to 65% (disgust), where 16.7% accuracy would be expected by chance. The average accuracy level did, however, differ significantly from that expected by chance, $\chi^2(1) = 39.75$, $p < .0001$. The results suggested that participants relatively accurately identified facial expressions based on full-face videos showing full faces of human actors supporting our first hypothesis (H1).

Also, the results supported our second hypothesis (H2). That is, results suggested that participants also relatively accurately identified facial expressions based on facial-landmark videos showing faces consisting of moving dots (H2). The mean accuracy level for the recognition of emotions displayed in the facial-landmark videos was 33.8% and ranged from 56% (anger) to 9% (fear). This average accuracy level was greater than expected by chance, $\chi^2(1) = 3.81, p < .05$.

In line with Bassili (1978) but now using computer-generated facial-landmark videos, results supported our third hypothesis in suggesting that participants identified facial expression videos better based on full-face videos, than based on facial-landmark videos(H3). That is, a comparison of the overall accuracy rate on the full-face videos and facial-landmark videos revealed that the former was significantly higher than the latter, $\chi^2(1) = 22.20$, $p < .001$. These results suggest that facial motion information (dots) can be useful in the differentiation of emotions, but the addition of other kinds of information might provide considerable help in the task.

Furthermore, because the emotions displayed by our actors may not have been optimally recognizable, we ascertained whether the structure of errors under full-face videos condition and facial-landmark videos conditions were similar. The correlation between responses given for full-face videos and facial-landmark videos (as shown in each cell of Table 2) was $r = .77$; $p < .001$. This indicated that participants' errors were far from random, and the errors they made in the facial-landmark videos were similar to those they made in the full-face videos.

Fourthly, the results also supported our fourth hypothesis by suggesting that participants' judgments about the valence and arousal levels of emotions displayed by full-face videos strongly correlated with participants' judgments about the valence and arousal level of emotions displayed by facial-landmark videos (H4). That is, the correlation between valence evaluations given for full-face videos and facial-landmark videos (as shown in Table 3) was $r = .91$; $p < .05$, and for participants arousal evaluations of the full-face videos and the facial-landmark videos was $r = .93$; $p < .05$.

Table 3. Participant's evaluation of the valence and arousal levels of the facial emotional expressions

		Reported emotion	
		Valence (Negative - Positive)	**Arousal** (Low - High)
Displayed emotion	**Happiness**	6.4/4.9	5.8/4.6
	Sadness	2.4/3.2	3.2/2.8
	Fear	2.0/3.8	5.7/4.1
	Disgust	1.9/3.6	5.7/4.3
	Anger	1.3/2.8	6.6/5.8

As can be seen in Table 3, the numbers represent the average levels of participants' valence and arousal evaluations when shown videos of the emotion described by the row label; Numbers on the left of the slash are for responses to the full-face videos, and those to the right are for the response to the facial-landmark videos. Each row represents the responses of 16 participants in each of the two conditions.

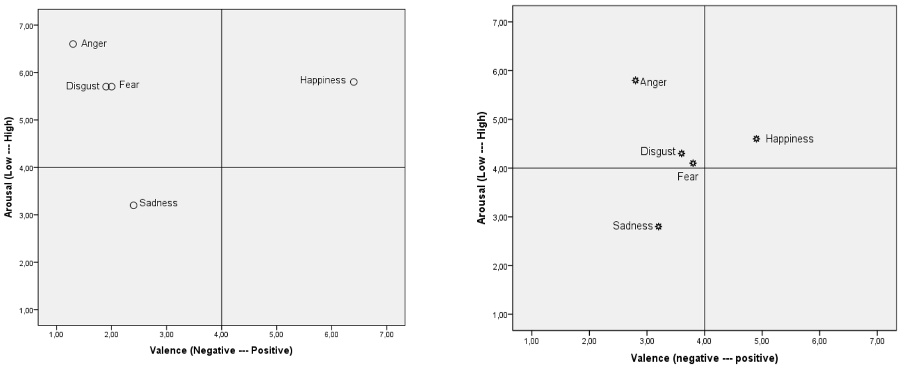

Fig. 2. The location of each emotional expression evaluated in the full-face videos (left) and facial-landmark videos (right), represented on a 2-dimensional space of valence (x-axis) and arousal (y-axis)

The two plots in Figure 2 illustrate the location of each emotional expression in a 2-dimensional space of valence (x-axis) and arousal (y-axis). The left plot represents the full-face videos and the right plot represents facial-landmark videos. Figure 2 suggests that emotional expressions of full-face videos and facial landmark videos are judged to be in the same coordinate area of this 2-dimensional space of valence and arousal.

5 Discussion

In line with Bassili (1978), results suggested that participants could accurately identify emotions expressed by dot faces (though less accurately than those expressed by full faces). Thereby, these results suggested that the algorithm used for extracting landmarks from full-face videos showing actor faces expression emotions can be used to produce facial-landmarks of which the expressed emotions can be identified relatively accurately. Thereby the facial-landmark video information based on this algorithm might be used to create more complete faces (e.g., robot faces, or faces of other artificial social actors), the emotional expressions of which might also be relatively accurately recognized.

Another conclusion that might be drawn from the current results is that facial-landmark videos might be sufficient for recognizing and identifying emotions expressed. This suggests that other information present in full-face videos (e.g., skin movement, skin color, details of eyes, mouth, etc.) is not necessary for identification of expressed emotions. Future research might investigate how to use this information about landmark movement to produce artificial facial expressions, and whether other information present in artificial faces should necessarily express congruent emotions.

In general, the current results suggest that future research might use the FaceTracker software to extract facial landmarks from a full face. Future research might investigate its algorithm methods and extend it to increase recognition accuracy and investigate ways to use its information to render more accurate emotional expression for robots or avatars.

At the same time, results suggested that participants identified the facial-landmark videos less accurately than the full-face videos. This suggested that full-face videos may contain additional information necessary for accurate emotion identification. This kind of information might be related to the landmarks and their movement. It might for example be the case that the algorithm we used was suboptimal, or that the number of dots was suboptimal. Also, this kind of information might be related to other factors. For example, some elements not present in the facial-landmark videos may have been important for optimal emotion identification. For example, information (expressing the same emotion as the moving dots) present in skin color, skin movement, details of elements of faces might help to increase accuracy of identification of emotion expressions in artificial faces even further.

Furthermore, future research might investigate not only algorithms for extracting landmark information from face videos, but could also investigate models to map the dots information to generate the related facial expression of artificial social actors. Those models could integrate modeling of landmark dot movements for accurate emotion expression with modeling of other facial elements and their role in improving emotion expression.

So, how to make a robot smile? The current research suggested that landmark information extracted from a human face expressing an emotion (information about location and movement of the elements of that face) can be enough information for a human perceiver to make an accurate judgment about the emotion expressed. That information might be enough to, for example, make a robot smile. To design robots that can accurately display facial expressions indicating an emotional state, we need technology to produce those facial expressions, and research that investigates the relationship between those technologies and human social perception of those artificial faces. The current research investigated a potential core element of such technology—landmarks extracted from human faces, and assessed the accuracy of human perception of that technology.

References

1. Bassili, J.N.: Facial motion in the perception of faces and of emotional expression. Journal of Experimental Psychology: Human Perception and Performance 4, 373–379 (1978)
2. Johansson, G.: Visual motion perception. Scientific American 232, 76–88 (1975)
3. Saragih, J.M., Lucey, S., Cohn, J.F., Court, T.: Real-time avatar animation from a single image. Automatic Face & Gesture (2011)
4. Saragih, J., Lucey, S., Cohn, J.: Deformable model fitting by regularized landmark mean-shift. International Journal of Computer Vision 91, 200–215 (2011)

5. Lucey, S., Wang, Y., Saragih, J., Cohn, J.: Non-rigid face tracking with enforced convexity and local appearance consistency constraint. International Journal of Image and Vision Computing 28, 781–789 (2010)
6. Saragih, J., Lucey, S., Cohn, J.: Face alignment through subspace constrained mean-shifts. In: IEEE International Conference on Computer Vision, pp. 1034–1041 (2009)
7. Saragih, J., Lucey, S., Cohn, J.: Deformable model fitting with a mixture of local experts. In: International Conference on Computer Vision (2009)
8. Fong, T., Nourbakhsh, I., Dautenhahn, K.: A survey of socially interactive robots. Robotics and Autonomous Systems. 42, 143–166 (2003)
9. Alexander, O., Rogers, M., Lambeth, W., Chiang, M., Debevec, P.: Creating a photoreal digital actor: the digital Emily project. In: 2009 Conference for Visual Media Production, pp. 176–187 (2009)
10. Yang, C., Chiang, W.: An interactive facial expression generation system. Springer Science Business Media. Mutimed Tools Appl. (2007)
11. Bänziger, T., Mortillaro, M., Scherer, K.R.: Introducing the Geneva Multimodal Expression corpus for experimental research on emotion perception. Emotion (2011) (advance online publication), doi:10.137/a0025827
12. Bänziger, T., Scherer, K.R.: Introducing the Geneva Multimodal Emotion Portrayal (GEMEP) corpus. In: Scherer, K.R., Bänziger, T., Roesch, E.B. (eds.) Blueprint for Affective Computing: A Sourcebook, pp. 271–294. Oxford university Press, Oxford (2010)
13. Xiao, J., Chai, J., Kanade, T.: A closed-form solution to non-rigid shape and motion recovery. International Journal of Computer Vision 2, 233–246 (2006)
14. Breazeal, C.: Designing sociable robots. MIT Press, Cambridge (2002)
15. Lucey, P., Lucey, S., Cohn, J.F.: Registration invariant representations for expression detection. In: 2010 International Conference on Digital Image Computing: Techniques and Applications, pp. 255–261 (2010)