Evaluation of Synthetic Faces: Human Recognition of Emotional Facial Displays

Erica Costantini[1], Fabio Pianesi[1], Piero Cosi[2]

[1] ITC-irst, 38050 Povo (TN) - Italy
{costante, pianesi}@itc.it
[2] ISTC-SPFD CNR 35121 Padova -Italy
cosi@csrf.pd.cnr.it

Abstract. . Despite the growing attention towards the communication adequacy of embodied conversational agents (ECAs), standards for their assessment are still missing. This paper reports about a methodology for the evaluation of the adequacy of facial displays in the expression of some basic emotional states, based on a recognition task. We consider recognition rates and error distribution, both in absolute terms and with respect to a human model. As to data analysis, we propose to resort to standard loglinear techniques and to information-theoretic ones. Results from an experiment are presented and the potentials of the methodology are discussed.

1 Introduction

In the last years there has been a great effort in the development of embodied conversational agents (ECAs) — i.e., artificial agents able to communicate by means of nonveRBal behaviour (gestures and facial displays) in addition to voice, and to generate emotional and conversational behaviour as a function of communicative goals and personality [1]. However, despite their growing popularity, standards for ECAs assessment are still missing. Some of the reasons for the lack a common evaluation framework for ECAs can be found in their variety and complexity. They target a large variety of domains and applications (e.g. multimodal human-computer interfaces, interactive games, tools for hearing impaired, …), serving many different purposes (virtual guides, salesmen or companions, aides to improve the intelligibility of natural and synthetic speech, to enhance the visual appearance of an interface, etc.). It seems natural that their effectiveness be measured in terms of communication abilities more than (or besides) standard usability indexes; but, it is still not clear how this can be accomplished, and to which extent this can be done in a task-independent way, to achieve generality. On the other hand, it is possible to single out different levels, including the appearance of the ECA, its 'mind model', the intelligibility of the gestures and emotional displays it produces, etc., which are expected to contribute to the ECA's effectiveness [2].

In this paper we focus on the evaluation of a specific aspect of the emotional behaviour of talking faces, i.e. *the adequacy of facial displays in the expression of some*

basic emotional states, measuring it through a recognition task. Though being quite a low-level (perceptual) assessment, it can provide important insights to developers—e.g., by allowing them to compare the performances of different faces based on different technologies, or of the same face at different stages of development. In addition, the recognisability of emotional expressions affects communicative behavior, hence its assessment is expected to be an important step towards a more comprehensive evaluation of the communicative abilities of synthetic agents.

We will start by discussing the evaluation procedure for MPEG4 facial animation players suggested by Ahlberg, Pandzic and You [3]. In §3 we attempt to improve the methodology, and introduce the details of a recognition experiment we conducted. We then discuss the data analysis and some of the results in §4, focusing on correct recognitions and on the way errors distribute across conditions. Section 5 tries to take stock of the work done, and to highlight interesting directions for future work.

2 The Linkoping approach

Ahlberg, et al [3] suggested a benchmark for measuring the subjective quality of a face model in terms of how well it can convey emotions (via facial expressions). They were interested in investigating how well a given face model can express emotions when controlled by low-level MPEG-4 FAPs (Facial Animation Parameters) captured from real persons acting out the emotions. The expressiveness was measured by the accuracy rate of human observers recognising the facial expression being shown. Subjects were given sequences of videos of human and synthetic faces expressing the 6 basic emotions of Ekman's set (fear, anger, surprise, sadness, happiness, disgust) [4]. Stimuli consisting of human faces were constructed by recording people acting different emotions through video camera (*natural* video sequences). During the recordings, the 3-D motion of the head, and a subset of 22 MPEG-4 facial feature points were tracked through head tracking equipment and IR-reflecting markers. This allowed the creation of MPEG-4 FAP files, which were then fed into two different facial animation engines to produce the *synthetic* video sequences.

The hypothesis was that the recognition rate for each synthetic face is better than the random case but worse than the human. In addition, the authors wanted to compare synthetic faces. The dispersion matrices containing the frequencies of the given judgments for each face were compared to an ideal dispersion matrix (perfect recognition) and a random matrix (random recognition). Absolute and relative scores for faces were provided, and the t-test was used to check for statistical significance of differences: The results showed significant differences between the models and the real (natural), ideal and random case; however, significant differences between the two face models were not detected.

It is an important feature of the proposed methodology that the expressiveness of MPEG4 facial animation players is compared (also) to that of the humans who provided the model: since people make recognition errors when confronted with human facial expressions, a data driven approach to face animation is expected to be capable of reproducing both the correct and incorrect recognitions of the model.

There are some limitations in the way the study was performed, though. Some are related to the reflective dots procedure used to record the data, which is very time- and effort-demanding, and requires that several constraints be met for the data to be reliable. Problems of this kind yielded some videos of low quality, which had to be discarded, producing a different number of video stimuli per each emotion and per each human model. Other possible sources of problems are: the choice of laymen as acting people, because of the low consistency and uniformity of the resulting expressions; the fact that some video were of different duration, this way introducing another disturbing factor; the resort to collective rather than individual sessions with subjects. Finally, the suggested method provides an easy way to compute absolute and relative scores. However, some technical details are not clear (one might object about the appropriateness of a t-test for testing significant differences between scores; it is not clear the reason for using a randomly generated matrix instead of the more standard procedure consisting in making hypotheses on the form of the distribution). More importantly, the suggested measure is quite a rough one, for it collapses all the information of confusion matrices in a single score. In particular, correct recognitions and errors cannot be told apart, nor it is possible to speculate about the different recognisability of the various emotions. Yet, especially for comparison and development purposes, it would be important to have finer-grained analyses addressing both dimensions — e.g., to understand how close the data-driven method (base on FAPs) mimics the human model on each of them.

3 The Experiment: Objectives and Method

For our experiment we modified the objective and methodology with respect to that described in the previous section. In particular, we attempted to improve the experimental design and the data analysis models, paying specific attention to correct recognitions and errors. As to data analysis, we exploited standard techniques for the analysis of categorical data (generalised linear models, loglinear analysis[9]) and an information theoretic approach to error analysis [10].

3.1 Objectives

The experiment aimed at evaluating expressiveness of two synthetic faces in two different animation conditions; in the first FAP files (FAP condition) recorded from actors were played; in the second, FAP files were produced from scripts specified by the developer. We will refer to the two conditions as the FAP as the rule-based (RB) one, respectively. RB was introduced because many existing talking heads exploit this mode of animation, and it seemed important to compare those approached to the data-driven one. In the data analysis we considered both the absolute merits of a given face*mode-of-animation combination (in terms of recognition rate) and its quality relative to a human model. The objective was to assess how much FAP-faces adhere to the model, and what kind of biases and idiosyncrasies, if any, the RB mode

could be responsible for. Finally, we aimed at exploring in detail the error distributions.

3.2 Method

We departed from the Ahlberg et al.'s methodology in the following respects: the neutral expression was added to the six emotions of the Ekman's set; another animation condition, the mentioned RB one, was added; recordings (and FAP files) were taken from one professional actor instead of laymen;[1] great attention was paid to recording conditions, so that we didn't have to discard any recordings, managing to have the same number of stimuli for each emotion and condition.

We used two synthetic faces (Face 1 and Face2); each was presented to subjects and evaluated in two different conditions: the face playing rule-based (RB) emotional expressions, and the face playing the FAP files extracted from the actor. This set up allowed us to: a) evaluate and compare RB approaches, whereby the specifications for expressing emotions are provided by the developer, to data-driven ones, b) pursue the task of cross-face comparison, and c) assess possible interactions between faces and mode of animation.

3.2.1 Video stimuli

Preparation of videos went through three steps: recording of an actor uttering a sentence and expressing different emotional states, production of the related MPEG-4 FAP (Facial Animation Parameters) files, and animation of the FAPs sequences using different synthetic faces.

The actor (male, 30 years old) was recorded through the Elite system [4], which uses two cameras with a frame rate of 100 Hz to capture 28 markers. Two synthetic 3D face models were used in the study, Face 1 [5] and Face2 [6]. Both faces enforce the MPEG-4 Facial Animation (FA) standard. FAPs were normalized according to the MPEG-4 FA standard, to make them speaker-independent. The point trajectories obtained from the motion tracking systems were converted into FAP streams through TRACK [8]. The FAP streams were then used to animate the synthetic faces to produce the FAP condition videos through screen capture.

The video camera recordings of the actor were digitalized and edited to be used for the Actor condition of the experiment. Finally, the rule based condition consisted in recordings obtained by playing the relevant scripts.

3.2.2 Experimental Design

A within-subjects design was adopted: subjects were presented with 3 blocks (ACTOR, FACE 1 and FACE 2) of 14 video files each, yielding a total of 42 judg-

[1] The actor was a male, while the faces we used were both female. It would have been interesting to control for cross-sex portability of FAP files, but this was not possible at the time the experiment was performed. We plan to address the issue in future studies.

ments per participant.[2] The animation conditions (RB and FAP) were appropriately randomized within the two blocks of synthetic faces.

As to emotional expressions, the videos covered the 6 emotions from Ekman's set plus 'neutral'. Each emotional state was expressed by the faces while uttering the Italian phonetically rich sentence "Il fabbro lavora con forza usando il martello e la tenaglia" (The smith works with strength using the hammer and the pincer). The audio was not presented. The task of the subjects was to identify the presented expressions by choosing from a list of the 7 available emotional states.

3.2.3 Procedure

Subjects were of 30 (15 males and 15 females) non-paid volunteers recruited at ITC-Irst. None was involved in the present project. They were given individual appointments for the experimental session in the recording lab (a silent room), and were individually tested. Before the experimental session they were given written instructions and went through a short training session to familiarize with the task. The training session exploited 4 video files for each of the three faces (total number of 12 stimuli), with different stimuli than those to be used in the experimental session. The video files (320x360, AVI file, Indeo-5.10 compression) were presented on the computer screen, through Microsoft Power Point ®. Each video file was presented only once. Each block had three different presentation orders, which were randomly created and balanced across conditions and participants. The presentation order of the three blocks was also balanced across participants.

The experimental session started immediately after the training session. Participants were asked to watch at the video files and express their judgement on a paper form, choosing from among the 7 available labels for emotional states (corresponding to the 7 presented emotional expressions). At the end of the experimental session, they given a 4 items questionnaire, aimed at collecting their feelings about the faces.

4 Results

4.1 Correct recognitions

Table 1 reports the recognition rates for each emotion and condition.

[2] The ACTOR block consisted of presentations of two series of videos from the same actor, called ACTOR1 and ACTOR2. This was done to control consistency of results with respect to the actor. Since no differences emerged, in the following our comparisons and discussion will be limited to ACTOR1, with the exception of §4.2, where both types of data from the actor are used again.

Table 1. Percentages of correct recognitions for each emotion and condition.

	ACTOR1	F1-FAP	F1-RB	F2-FAP	F2-RB
disgust	13%	20%	53%	17%	17%
happiness	97%	80%	40%	80%	77%
neutral	70%	70%	60%	53%	67%
fear	50%	17%	77%	0%	77%
anger	90%	27%	53%	7%	23%
surprise	47%	40%	87%	33%	90%
sadness	17%	7%	97%	7%	97%
All	55%	37%	67%	28%	64%

Correct recognitions were analysed by dichotomizing the responses (correct vs. wrong) and developing a multinomial logit log-linear model [9], with the correct/ wrong responses as the dependent variable, and the faces (ACTOR1, Face1 and Face2), the mode of animation (RB and FAP), and the presented emotions as the independent variables. A preliminary model selection analysis showed that the full model (including the main effects for each independent variable, and all the second, and third order interactions) was needed to adequately fit the data. Hence we developed a full logit model for the data.

Direct comparisons of the performances of the different faces in the different conditions were accomplished by computing the z-scores of the relevant log odd-ratios from the parameters of the logit model. The chosen level of confidence was $p<.01$, and confidence intervals for the z-scores were $z<-2.58$ and $z>2.58$.[3]

Results from global comparisons (ignoring differences concerning presented emotions) show that ACTOR1 has better recognition rates than both faces in the FAP mode. The scores for the comparisons between ACTOR1 and the two faces in the RB condition don't reach significance, though the one for Face1 goes very close to doing so, at the chosen level (z-score= 2.48). Finally, both synthetic faces increase their recognition rate when going from the FAP- to the RB-condition. Hence, at a global level the RB condition is closer to ACTOR1 than the RB one.

The results of a more fine grained analysis, addressing faces, mode of animation and presented emotion, are summarized in Tables 2 through Table 4.

Table 2. Significant comparisons involving at least one synt. face in the RB mode.

Anger	ACTOR1 > Face1
	ACTOR1 >Face2
Happiness	ACTOR1 > Face1
	Face2 > Face1
Disgust	Face1 > ACTOR1
	Face1 > Face2
Surprise	Face1 > ACTOR1
	Face2 > ACTOR1
Sadness	Face1 > ACTOR1
	Face2 > ACTOR1

Table 3. Significant comparisons involving at least one synt. face in the FAP mode.

Anger	ACTOR1 > Face1
	ACTOR1 > Face2
Fear	ACTOR1 > Face2

[3] See [9] for details.

Table 2 reports all the significant comparisons in which one (or more) of the synthetic face was in the RB mode; all the omitted combinations did not yield significant differences. Table 3 reports the significant results for comparisons in which one (or more) of the face was in the FAP-mode, and Table 4 informs about comparisons on a given face in the two modes of animation (RB vs. FAP).

Table 4. Significant comparisons for the same face in the two modes.

Happiness	Face1-FAP > Face1-RB
surprise	Face2-RB > Face2-FAP
	Face1-RB > Face1-FAP
fear	Face2-RB > Face2-FAP
	Face1-RB > Face1-FAP
sadness	Face2-RB > Face2-FAP
	Face1-RB > Face1-FAP

Summarizing the results, we have that:

- The RB mode improves the recognition rates of both faces of the same amount and on the same presented emotions (surprise, fear and sadness), over the FAP mode (table 4). The latter benefits only Face 1 on happiness. Hence, the RB condition is superior to the FAP one, as far as RR is concerned..
- The FAP condition does not cause much differences across the synthetic faces, nor does the RB one, a part from minor differences (Face2 is superior to Face1 on happiness, while the opposite obtains on disgust; Table 2).
- With respect to ACTOR1, the two faces in the FAP condition give either identical or poorer recognition rates (Table 3). This accords with the conclusions from the global analysis.
- With RB, ACTOR1 is still better than Face1 and Face2 on anger, and better than Face1 on happiness. The situation reverses in favor of RB for both faces on surprise and sadness (Table 2). So, the global similarity between RB and ACTOR1 we observed above, concealed important differences that tend to mutually cancel at the global level.

In conclusion, on absolute grounds the RB mode has higher recognition rates than the FAP one. With respect to ACTOR1, RB-faces do not globally differ from it, whereas ACTOR1 shows a global superiority over the FAP mode.

When we go into details, however, the picture changes somewhat: the RB mode and ACTOR1 diverge on anger, where ACTOR1 outperforms RB-faces, and on surprise and sadness, where the opposite obtains. Now, anger is the only emotion on which ACTOR1 is stably superior to all faces in all conditions, suggesting that our faces as such are bad at it (or, the actor looks angry). The superiority of the RB mode on surprise and sadness, on the other hand, suggests that the scripts of RB mode produce agents that express these emotions better (at least with respect to our actor).

Turning to the FAP mode, we should not hasten to conclude that it is ineffective. True, on the global tests they were worst than ACTOR1. The detailed analysis, however, shows that this is basically due to their poor performances on anger and, for Face2, on fear. If we discount anger on the same grounds as for the RB mode (the actor looks angry), and accept that fear is a real problem for Face2-FAP, in the re-

maining conditions the FAP mode turns out to be closer in performances to ACTOR1 (the model) than the RB one; this accords with our expectations (see Table 3).

4.2 Distributions of recognition errors

We turn now to study errors, trying to understand whether and how the way they distribute is affected by our independent variables: faces, mode of animation and presented emotions. We will not resort to the same techniques of the previous section. Log-linear analysis can be easily extended to address the greater number of response categories (7 instead of 2) that is now required; however, the limited size of our sample (30 people) would weaken our conclusions. Moreover, in this section we are interested in finding simple but powerful tools to succinctly characterize errors and their distributions, allowing for easy comparisons; loglinear techniques do not directly provide for them. Thus, we will explore an information-theoretical approach [10] that factors out various contributions to the global information/uncertainty displayed by confusion matrices, turning some of them into the tools we need. In this work we will focus on the number of confusion classes, and on the characterization of errors shared across conditions. Other important dimensions (e.g., typical error classes) will not be addressed here.[4]

Table 5 reports the global confusion matrix, showing how correct responses and errors distribute across stimuli (rows) and responses (columns).

Table 5 . Overall confusion matrix (percentages).

	Disg.	Happ.	neuter	fear	anger	surpr.	sadn.
disgust	**22%**	9%	16%	9%	8%	4%	32%
happiness	4%	**78%**	8%	1%	4%	3%	3%
neutral	2%	3%	**66%**	2%	11%	7%	9%
fear	4%	6%	8%	**44%**	14%	21%	2%
anger	5%	1%	20%	13%	**49%**	10%	3%
surprise	1%	9%	7%	13%	8%	**61%**	2%
sadness	7%	2%	18%	14%	4%	10%	**44%**

An appreciation of how errors distribute can be obtained by considering L, the mean entropy of the stimulus sequence that is not accounted for in the response sequence. For a given response category, *r*, L amounts to the (log) of the mean number of stimulus categories that cause responses to fall in *r*. Ideally, each response is induced by one and only one stimulus category (the right one), so that L=0. The converse of L is G, which informs about the (log) mean number of response categories for each stimulus category.[5]

[4] The price to pay to the information theoretic approach is that it does not come equipped with the rich inferential apparatus of other techniques. Hence we will not be able to anchor our conclusion to tests of statistical significance.

[5] $L=H_{cm}-H_{resp}$
$G=H_{cm}-H_{stim}$
where H_{stim}= entropy of the stimulus sequence, H_{rest}= entropy of the response sequence, and H_{cm}= entropy of the confusion matrix.

Table 6 reports the results in term of 2^L (number of stimulus category per response category) and 2^G (number of response category for stimulus category). As can be seen, the FAP-faces are quite different from the other combinations, having the greatest figures (more error categories) on both dimensions. We must take these data with some care, though, for L and G do not discount the distribution of errors in the confusion matrix, and are sensitive to the error rate; hence, greater error rates might give raise to larger L and G, this way biasing comparisons.

Table 6. Values of 2^L and 2^G

	2^L	2^G
ACTOR1	2.61	2.40
Face2-RB	2.67	2.46
Face1-RB	2.65	2.48
Face1-FAP	4.28	3.50
Face2-FAP	5.18	3.74

Table 7. Values of d_r and d_s

	d_r	d_s
ACTOR1	1.82	1.52
Face2-RB	2.48	1.97
Face1-RB	2.76	2.27
Face1-FAP	3.54	2.57
Face2-FAP	4.30	2.74

A more refined measure of the way responses distribute is provided by the indices d_s and d_r. The former measures the effective mean number of error (confusion) classes per stimulus, discounting the error distribution in the sense that stimuli with a low number of errors, which are spread across many response categories, contribute little. The other index, d_r, informs about the mean number of stimulus categories a response category collects confusion from. In an ideal situation, both indices should be 0.[6]

The results are reported in Table 7. Although the resulting ordering is compatible with that of Table 6, discounting the error rate reveals differences that were previously blurred. The RB faces are now somewhat farther from ACTOR1. Moreover, the variation of d_s across synthetic faces is quite limited (range: 1.97-2.75) compared to the variation for d_r (2.48-4.30). We conclude that: a) the RB faces are the closest to ACTOR1, as far as the number of error categories is concerned; b) the number of confusion categories along the stimulus dimension (d_s) is substantially stable across synthetic faces and mode of animation; c) the number of confusion categories along the response dimension (d_r) shows a clear ascending trend, when we move from ACTOR1 to RB faces and then to FAP ones. Thence, the real differences between the RB-mode and the FAP-mode on errors classes concern the way response categories collect confusions (d_r), rather than the number of error classes per stimulus category.

Suppose, now, that we want to know now how much similar is the error distribution along the stimuli dimension between two face*mode-of-animation combinations, sayFace2-FAP and Face2-RB, as a way to capture the contribution of Face2 to errors. The idea is that, to a certain extent, the errors that are shared between Face2-FAP and Face2-RB reflect Face2's properties (the way it looks, the underlying animation engine, rendering, etc.), providing us with a measure of the confusions Face2 induces, independently from the condition (FAP vs. RB) it is presented in. To this end, we resort to indices δ_s and δ_r, which are computed on the pooled confusion matrix for Face2-FAP and Face2-RB. They yield the effective fraction of errors that fall outside

[6] $d_s=2^{(G-H_\varepsilon)/\varepsilon}$
$d_r=2^{(L-H_\varepsilon)/\varepsilon}$
where L and G are as before, H_ε is the error entropy of the confusion matrix, and ε is the error rate.

the shared error categories, corrected for the overall differences in the distribution of stimuli (δ_s) and responses (δ_r). In a way, the lower these figures, the higher is the probability that a given error is due to Face2 itself, rather than to the mode of animation (or on any other intervening conditions).[7] Table 8 reports results obtained by pooling together: ACTOR1 and ACTOR2 (reported as ACTOR), Face1-FAP and Face2-FAP (FAP), Face1-RB and Face2-RB (RB), Face1-RB and Face1-FAP (Face1), Face2-RB and Face2-FAP (Face2).[8]

Table 8. Values of δ_r and δ_s for the various conditions.

	δ_r	δ_s
ACTOR	0.12	0.14
FAP	0.2	0.24
RB	0.24	0.34
Face2	0.35	0.57
Face1	0.39	0.67

Table 9. Values of δ_r and δ_s computed with respect to ACTOR1.

	δ_r	δ_s
Face1-FAP	0.22	0.43
Face2-FAP	0.30	0.56
Face1-RB	0.66	0.78
Face2-RB	0.63	0.79

Neglecting ACTOR, under the proposed interpretation the mode of animation accounts for a greater fraction of the errors than faces do (excluding ACTOR).

In Table 9 we have pooled the confusion matrices of each face*mode-of-animation combination with that of ACTOR1, and then computed δ_r and δ_s. The figures indicate the amounts of non-shared errors between each combination and ACTOR1, and inform us about how similar each combination is to ACTOR1: the lower the fraction of errors they do not share, the more similar they are. Face1-FAP is the combination with the lowest figures, hence the one sharing the greatest amount of errors with ACTOR1, closely followed by Face2-FAP. The two RB-faces are farther away, sharing fewer errors with ACTOR1

In conclusion, the analysis of errors has shown that:

- RB-faces disperse errors on fewer confusion categories than FAP-ones, in this being closer to Actor (Table 7);
- the FAP-faces share a greater amount of errors with ACTOR1 than the RB-faces (Table 9);
- in a given face*mode-of-animation combination it is the mode of animation that accounts for the greater portion of errors (Table 8).

We can interpret these results by saying that the mode of animation affects the error distribution more than the type of face (FACE1 or FACE2). In detail, the confusion categories of the RB-faces don't overlap much with those of the actor, this way determining a low number of shared errors. That is, the error distribution of the RB mode is quite distinct from that of the actor on both the stimulus and the response dimension. The FAP-faces, on the other hand, because of their greater number of error categories (Table 7), share some of them with the actor, this way explaining the higher number of common errors (Table 9). In other words, the great number of

[7] For reasons of space, we cannot report here the formulae for δ_s and δ_r. See [10] for more on this point.
[8] See fn. 2.

common errors between the FAP faces and the actor is probably a consequence of the higher dispersion of error categories in the FAP conditions.

4.3 Questionnaires

At the end of the session each participant was asked to answer 4 close ended questions asking for which face they felt the judgement task was easiest/hardest and which among the synthetic faces was the most natural/pleasant. The actor was rated as the easiest face to judge (53%), and Face1 got slightly better results than Face2. Concerning pleasantness/naturalness, Face2 was rated higher than Face1 (59% versus 41%).

5. Conclusions and Future Work

In this paper we have proposed an approach to the assessment of the identifiability of emotional expressions performed by synthetic faces through two different modes of animation: the so-called rule-based and the data-driven one. Both absolute and relative assessments were pursued, the latter by comparing the expressions of two synthetic faces to those performed by an actor. With respect to previous studies, we have adopted more refined techniques: a loglinear analysis of the data for the recognition rate, and an information-theoretic approach for error analysis. The results indicate that, in absolute terms, the RB condition is superior to the data-driven one with both faces, as far as recognition rate is concerned. In relation to the human model, however, the data-driven method matches the model better. With respect to error distribution, both the RB and the FAP mode seem to differ from the human model, though for different reasons. All these results are largely independent of the face used.

Besides allowing comparisons among different conditions, the proposed approach may directly impact on design and development choices. For instance, the fact that no major differences are exclusively due to the faces per se might suggest that the state-of-the-art of the relevant technologies is such that the appearance and other physical characteristics of the synthetic faces is presently less crucial than the way information about the relevant emotional expression is provided to them. Another possible indication is that if recognisability is the ultimate goal, then rule-based approaches seems to be appropriate: hand-written scripts allow to finely tune expressions till the desired results are obtained. On the other hand, if the focus is on ‘naturalness’, then data-driven methods are a ‘sort of good’ choice, because they produce recognition patterns close to those of the human model. However, they are still far from appropriately matching the model on error distribution, suggesting that design and development effort be focused on this aspect, in particular on reducing the number of error categories.

Turning to possible improvements, this study has not attempted to identify error categories; rather, we simply measured their numbers and common error fractions. However, information about the most common error categories, along both the stimulus and the response dimensions, would be extremely valuable to characterise how the face looks like in general. Besides this, there are factors that might affect the recogni-

tion task, which have not been addressed here. For instance: a) the sex of the synthetic face and /or of the source for the FAPs: do the two interact? Do they interact with the sex of the subjects? b) Attractiveness: synthetic faces are built to be attractive, whereas (true) human faces aren't. Has this any effect on our task? Other important directions for future investigations involve the relative importance (if any) of the upper/lower part of the face in the expression of emotions: How much does recognition deteriorate (if it does) when emotional expressions are limited to the upper/lower part of the face? Finally, the methodology could be improved by extending measurement to reaction times, this way obtaining information on the difficulty of the judgments for the subjects; and by trying to relate subjective evaluations of the faces (as in §4.3) to the results of the data analysis.

6 Acknowledgments

We wish to thanks the colleagues in the PF-STAR project who contributed valuable ideas and comments. In particular: N. Mana, M. Prete, F. Del Missier, W. GeRBino, J. Ahlberg and M. Zancanaro. Finally, we wish to to extend our thanks to C. Pelachaud, A. De Angeli e Z. Ruttkay who, as members of PF-STAR advisory board, provided important suggestions and comments on our work.

This work was done within the PF-STAR project (http://pfstar.itc.it), and partially supported by grant IST-2001-37599 from the EU.

References

1. Cassell, J., Sullivan, J., Prevost, S., Churchill, E. (Eds.): Embodied Conversational Agents. Cambridge, MA: MIT Press (2000).
2. Ruttkay, Zs., Doorman, C., Noot, H.: Evaluating ECAs - What and how?, Proceedins of the AAMAS02 Workshop on Embodied conversational agents - let's specify and evaluate them!, Bologna, Italy, July 2002.
3. Ahlberg, J., Pandzic, I. S., You, L. Evaluationg MPEG-4 Facial Animation Players, in Pandzic, I. S., Forchhimer, R. (eds), 'MPEG-4 Facial Animation: the standard, implementation and applications', 287-291, Wiley & Sons, Chichester, 2002.
4. Eckman P., Friesen W., Manual for the Facial Action Coding Systems, Consulting Psych. Press, Palo Alto (CA), 1977.
5. Ferrigno G., Pedotti A. ELITE: A Digital Dedicated Hardware System for Movement Analysis via Real-Time TV Signal Processing. In IEEE - BME-32, 943-950, 1985.
6. Cosi P., Fusaro A., Tisato G., LUCIA a New Italian Talking-Head Based on a Modified Cohen-Massaro's Labial Coarticulation Model, Proceedings of Eurospeech '03, Geneva, Switzerland, Vol. III, 2269-2272.
7. Pelachaud C., Magno Caldognetto E., Zmarich C., Cosi P., Modelling an Italian Talking Head, Proceedings of AVSP 2001, Aalborg, Denmark, September 7-9, 2001, 72-77.
8. Cosi P., Fusaro A., Grigoletto D., Tisato G., Data-Driven Methodology and Tools for Designing Virtual Agents Exploiting Emotional Attitudes (submitted to ASD 2004).
9 Agresti, A. Categorical Data Analysis. John Wiley and Sons (2002), New York

10 van Son, R.J.J.H. A Method to Quantify the Error Distribution in Confusion Matrices. Proceedings 18, 41-63. Institute of Phonetic Sciences, University of Amsterdam (1994).
11 Rosenthal, R. Judgment Studies. Cambridge University Press (1987). Cambridge