Recognising Emotions in Human and Synthetic Faces: The Role of the Upper and Lower Parts of the Face

Erica Costantini, Fabio Pianesi, Michela Prete
ITC-irst, 38050 Povo (TN) – Italy
+39 0461 314570
{costante, pianesi, prete}@itc.it

.

ABSTRACT

Embodied Conversational Agents that can express emotions are a popular topic. Yet, despite recent attempts, reliable methods are still lacking to assess the quality of facial displays. This paper extends and refines the work in [6], focusing on the role of the upper and the lower portions of the face. We analysed the recognition rates and errors from the responses of 74 subjects to the presentations of dynamic (human and synthetic) faces. The results points to the possibility of: a) addressing the issue of the naturalness of synthetic faces, and b) a greater importance of the upper part.

Categories and Subject Descriptors

H.1.2 [**Information Systems**]: User/Machine Systems – *human factors.*

General Terms

Measurement, Performance, Design, Experimentation, Human Factors.

Keywords

User study, synthetic faces, expressiveness, emotion recognition, face regions.

1. INTRODUCTION

In the last years there has been a great effort towards developing embodied conversational agents (ECAs) — i.e., artificial agents able to communicate by means of nonverbal (gestures and facial displays), and verbal behaviour, and to exhibit emotional and conversational behaviour as a function of communicative goals and personality [4]. Parallel to that, interest has raised on methodologies and protocols for evaluating ECAs. Most of the studies conducted so far have addressed the end-to-end evaluation of systems exploiting ECAs, and focusing on dimensions such as the effectiveness and quality of the resulting interaction, the ECA's believability, etc.; see [12], and [13] for a review of relevant studies and an attempt at providing a general framework.

IUI'05, January 10–13, 2005, San Diego, California, USA.

Very few works have addressed the *quality* of facial displays, and those 'low levels' dimensions that arguably determine it: lip-speech synchronisation, facial gestures signalling emotions, or emphasis, punctuation and other discourse-level regulatory characteristics. In particular, despite the current interest in the topic, and the many efforts towards endowing ECAs with the capabilities of expressing emotional states, neither benchmarks nor agreed methodologies are available to assess this dimension. Arguably, though, they would provide important information for testing and development purposes, and for comparatively evaluating different platforms. In many respects, the measurement of selected dimensions into which the 'quality of facial display' can be articulated — e.g., the recognisability of emotional expressions — can be expected to play a role similar to that played by the Word Error Rate for speech recognisers, or the recall/precision pair for Information Retrieval and Information Extraction: measures that are easy to use and interpret for development and comparative purposes.

Pursuing the development of benchmarks and protocols for the assessment of emotional expressions can also positively affect our understanding of the field itself, and help us in better designing systems exploiting ECAs. For instance, the many studies in the psychological literature addressing this topic have mostly resorted to static stimuli (pictures), with a substantial neglect of the richer and more complex dynamic stimuli — e.g., videos portraying the expression of emotions by humans or synthetic faces. Since it is the latter that we are mostly interested in for the purposes of evaluating ECAs, we can expect the research towards benchmarks and protocols to advance our knowledge, by further articulating the relevant questions, and by providing new insights. For instance, a proper understanding of the quality of a synthetic face requires that we be able to chart the effect of varying conditions on recognition judgments: the use of acted vs. spontaneous expressions as stimuli; the use of actors vs. laymen for producing stimuli; the role of gender of the face and/or of the subject; the contribution and role of different portions of the face (does recognition change when only the upper or the lower part of the face are available? If it does, how are those changes related to the recognition of the whole face?); the effect of executing concurrent tasks during the expression of the emotions, most notably, uttering something; the role of cultural and linguistic differences on recognising emotional expressions; etc.

In this paper we address some of the above mentioned issues in the context of an exploration of the methodologies for the evaluation of the expressiveness of synthetic faces. In particular, we capitalise on, and further refine a recent proposal, [6], where the emotional expressions produced by synthetic faces are assessed against those produced by humans (actors) in terms of

both the correct recognitions and the similarities/differences of the corresponding error distributions. While doing so, we will also investigate the role of the upper and lower part of the face in the recognition task, for both human and synthetic faces.

This paper is structured as follows: in the next section we discuss some of the relevant literature; then we present the study design and procedure; the remaining sections report on the results focusing on the recognition rates and the analysis of errors. The last section draws the conclusions.

2. PREVIOUS WORKS

Ahlberg et al. [2] were the first to use the expressions performed by humans as a golden standard against which to assess the expressivity of synthetic faces in a recognition task. They were interested in how well a given face model can express emotions when controlled by MPEG-4 FAPs (Facial Animation Parameters) captured from people acting the emotions. The expressiveness was measured through the recognition rate of human observers classifying stimuli. The latter consisted of videos of human and synthetic faces expressing Ekman's [8] 6 basic emotions: fear, anger, surprise, sadness, joy, disgust. Stimuli consisting of human faces were obtained by recording people (laymen) acting different emotions through video camera. During the same recordings, the 3-D motion of the head, and a subset of 22 MPEG-4 facial feature points were tracked through motion capture equipment. This way, MPEG-4 FAP files were created, which were then fed into two different Facial Animation Engines to produce the *synthetic* video sequences. The results showed significant differences between the synthetic and the human faces, but no differences between the two synthetic faces.

Ahlberg et al.'s proposal to compare the expressiveness of synthetic faces with that of humans suggests that the best synthetic face is the one whose performances are the closest to those of human faces. In a way, this provides for an operational definition of the *naturalness* of synthetic faces. Not of 'perceived' naturalness, of course, but a more 'objective' measure in terms of the similarities in human behavior across the two conditions. For these comparisons to be meaningful, however, they cannot be limited to successful recognitions, but should extend to errors and the way they distribute among confusion classes. According to the proposed notion, in fact, a naturalistic synthetic face should reproduce both the correct recognitions and the errors of the human model.

Costantini et al. [6] took up some of those concerns, and proposed a more refined procedure, in the context of a comparative evaluation of two different animation conditions for synthetic faces: a) FAP files recorded from the actor through a methodology similar to that exploited by Alhberg et., and then fed into the synthetic faces; and b) the 'proprietary' FAP files of the synthetic faces, produced from scripts specified by the developer (the Script Based, or SB-condition). The testbed consisted of expressions performed by the same professional actor who provided the FAP files. Subjects were presented with expressions of Ekman's six emotions plus the neutral one. An important difference with respect to most previous works was that the expression of each emotional state co-occurred with the utterance of 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 available. Given the categorical nature of the data, Costantini et al. exploited a log linear analysis for the recognition rate, and an information-theoretic approach for errors. The results indicated a tendency for synthetic faces in the SB condition to perform better than the human and the FAP conditions. With respect to the human model, however, the FAP method was closer to it in terms of both the recognition rates and the way errors distributed among confusion classes. According to the considerations made above, the FAP methods yielded more naturalistic faces.

Kätsyri et al. [10] compared the identification of emotional expression of a synthetic face with those of natural faces, exploiting both static (pictures) and dynamic stimuli. The latter were short (1-1.5 secs.) videos showing the expression from the neutral position till the apex. As to the human stimuli, they exploited the Cohn-Kanade's database, and a proprietary one, consisting of expressions acted by professional actors trained to perform them in accordance to Ekman's FAUs [8]. The subjects had to rate how much each expression (static or dynamic) contained/realised each of the six Ekman's basic emotions. They found that the overall levels of identification for the synthetic face were lower than the human one.

Concerning the role of the upper and the lower part of the face on emotion recognition, Bassili's [3] review of the literature on the topic noticed the absence of agreement amongst published works, most probably due to the different approaches and methodologies. The only clear fact was that different parts of the face turned out to be important for different emotions. In his own study, Bassili [3] found that the lower region of the face was associated with a higher recognition rate than the upper part on joy, surprise and disgust, whereas the reverse obtained for sadness and fear. No differences were found for anger. From this, Bassili directly concluded that the bottom of the face is 'more useful' for the recognition of joy, surprise and disgust, whereas the upper part is more useful for sadness and fear. It is not clear, however, that this particular conclusion is warranted, at least if we take 'more useful' as meaning that in the recognition task with the whole face the information provided by the bottom part of the face is more important to correct recognitions than that provided by the upper part. To establish this point, the simple comparison of recognition rates for the separate face regions does not seem to suffice. Rather, more direct ways are needed to explore the extent to which the information used during the recognition task for a given face region is re-used when the subject is confronted with the whole face.

3. THE STUDY

3.1 Objectives

The present study had two main goals: a) to further refine and advance the methodology of [6] while b) investigating the role the different regions of the face play in the recognition tasks for emotional expressions. Hence we systematically compared the performances and the error distributions obtained with whole faces (henceforth *whole*) with those resulting from the presentation of the upper (henceforth: *eyes*) and of the lower (henceforth: *mouth*) regions of the face.

3.2 Experimental Design

A within-subjects design was adopted: subjects were presented with 3 blocks (ACTOR, FACE1 and FACE2) of stimuli (video files). For synthetic faces, the proprietary SB mode of animation was used, see [6].

Each block consisted of video files covering the 6 emotions of Ekman's set plus 'neutral'. As in [6], each emotion 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 made available to subjects.

Stimuli were varied according to three conditions: 'whole', 'eyes' and 'mouth'. Hence, each block contained 3(CONDITIONS) × 7(EMOTIONS)=21 stimuli, for a total of 63 stimuli per participant.

3.3 Video stimuli

The actor (male, 30 years old) was recorded through the Elite system [9], which uses two cameras with a frame rate of 100 Hz to capture 28 markers. The video camera recordings of the actor were digitized and edited to be used for the Actor condition of the experiment.

Two synthetic 3D face models were used, Face1 [5] and Face2 [11], both enforcing the MPEG-4 Facial Animation (FA) standard. The synthetic video files were produced by screen-capturing of the synthetic faces playing the relevant script.

The stimuli for the 'eyes' and the 'mouth' conditions were obtained by cutting the videos for whole faces into two halves by means of VirtualDub. Following Vanger et al. [15], we included the chin, the lips, the nostril and the nose tip in the lower part (the 'mouth'); the nose bridge, the eyes, the eyebrows, the brow and the hair in the upper part of the face (the 'eyes'). This way, each stimulus in the 'whole' condition had one corresponding stimulus in the 'mouth', and one corresponding stimulus in the 'eyes' condition.

Table 1. Format of the video files

	whole	eyes	mouth
Actor	320×360	160×360	128×360
Face1	320×360	160×360	160×360
Face2	320×360	176×360	112×360

The video format used for presentation to the subjects were AVI file, with Indeo-5.10 compression, see Table 1. The video stimuli had different durations: 4 seconds for the human face and 7 seconds for the synthetic ones.[1]

3.4 Participants and procedure

Subjects were 74 (32 males and 42 females) students from the Department of Psychology, University of Trieste (Italy), rewarded through credits. They were individually tested in a silent lab. Before the experimental session, they were given written instructions and went through a short training session to familiarize with the task. The training session consisted in 9 stimuli, different from those of the experimental session. They were: for the Actor condition: anger × mouth, disgust × whole, neutral × eyes; for Face1: fear × mouth, neutral × whole, joy × eyes; for Face2: neutral × mouth, surprise × whole face, sadness × eyes.

The experimental session started immediately after the training one. Video files were presented on the computer screen, through Microsoft Power Point ®, with each of them presented only once. Each block had three different presentation orders that were randomly created and balanced across conditions and participants. Participants were asked to watch at each video file and to express their judgment on a paper form, choosing from the 7 available labels for emotional states (corresponding to the 7 presented emotional expressions).

4. RESULTS – RECOGNITION RATES

In this section we will first consider the role of the upper vs. the lower part of the face; then we will compare the recognitions rates across the various FACE*CONDITION combinations.

4.1 The role of the eyes and of the mouth

While discussing [3], we argued that more refined methods than the simple comparison of recognition rates should be used to address the relative importance of the upper and lower regions of the face. In this paper we take advantage of the pairing of the stimuli for the 'whole' condition with those for the 'eyes' and the 'mouth ones, and use Jaccard similarity index. For each face, and each comparison class ('eyes-whole' and 'mouth-whole') Jaccard index is defined as $d/(b+c+d)$, with respect to the crosstabulation of the data exemplified in Table 2 (0=wrong and 1=correct).

Table 2. Example of the crosstabulation used to compute the Jaccard index.

		eyes 0	eyes 1
whole	0	a	b
	1	c	d

Jaccard index is a weighted estimate of the conditional probability that a given stimulus yields a correct response on the row dimension (e.g., 'whole') given that the corresponding stimulus yields a correct response in the column dimension (e.g., 'eyes').[2] Hence, it provides a measure of the contribution of the eyes (the column dimension of Table 2) to the correct responses in the 'whole' condition (the row dimension). Table 3 reports the values of the Jaccard index for global comparisons, abstracting away from single emotions.[3]

Table 3. Values of Jaccard index for global comparisons

	Actor	Face1	Face2
Eyes-whole	.516	.430	.482
Mouth-whole	.528	.419	.352

The only significant differences concern a) Face2, where the eyes contribute more to 'whole' than the mouth, (z=3.67) and b) the fact that the contribution of the mouth to 'whole' is higher with Actor than with Face1 (z=2.89) and Face2 (z=4.75). Hence, ***for a***

[1] The differences were related to the computationally expensive process of generating the expression for synthetic faces. We could not investigate whether the durations of the stimuli had any effects on the experimental variables.

[2] It penalises the cases in which the number of successes on the column dimension is low with respect to the number of all successes, and rewards the cases when it is high.

[3] The comparison of two Jaccard indices was performed through their confidence intervals, with p<.01. Variances and standard errors were derived through the delta method [1].

Henceforth, and whenever possible, we will use standardised scores to discuss our results. Notice that a significance level of p<.01 correspond to z>|2.58|.

given face, the eyes and the mouth tend to globally contribute similar amounts of information to the correct recognitions of the whole face; the only exception is Face2, whose mouth contributes less than the eyes. At the same time, the contribution of the eyes remains globally stable across faces, whereas that of the mouth tends to decrease with synthetic faces.

Table 4 summarises the results of a finer-grained analysis, addressing the differences due to emotions (only statistically significant differences are reported).

Table 4. Comparing the contribution of the eyes (e) and of the mouth (m) to the correct responses in the 'whole' condition. Italics indicate close-to-significance data

	Actor	Face1	Face2
Disgust	=	*m>e*	=
Joy	=	m>e	m>e
Sadness	e>m	=	e>m
Fear	=	=	e>m
Anger	*m>e*	e>m	e>m
Surprise	e>m	e>m	e>m
neutral	*e>m*	=	=

With Actor, the differences between the contribution of the eyes and of the mouth to the correct recognitions in the 'whole' condition reach the chosen level of significance ($p<.01$) only for sadness ($z=3.76$) and surprise ($z=3.18$); in two other cases they go very close to doing so: $z=-2.49$, $p=.0128$ for anger, and $z=2.53$, $p=.0114$ for neutral.[4] Hence, in at least two cases the contribution of the eyes to the correct recognition in the 'whole' condition is higher than the mouth's.

With Face1 the variability increases. The mouth contributes more than the eyes on joy ($z=5.19$), whereas the eyes are more important for anger ($z=4,097$) and surprise ($z=4.79$). Finally, the comparison on disgust yields a close-to-significance value ($z=-2.485$, $p=.0130$), indicating a tendency for the mouth to be more important than the eyes on this emotion.

With Face2 the mouth contributes more to the correct responses in the whole condition on joy ($z=6.64$); the eyes' contribution is greater on sadness ($z=3.22$), fear ($z=2.65$), anger ($z=6.42$) and surprise ($z=6.33$).

In conclusion, the data for Actor ***confirms Bassili's observation that, with human faces, the importance of face regions depends on the emotion***. Concerning the synthetic faces, the increasing differences between the face regions and Actor might prove important in the assessment of their naturalness. Moreover, the frequent advantage of the eyes points to the possibility that ***the articulatory movements due to the speech negatively affect the contribution of the mouth in the recognition of complex expressions***.

4.2 Comparing recognition rates

In the next two subsections we study how successful recognitions depend on FACES, CONDITIONS and EMOTIONS. We first ran a loglinear model selection procedure. The variables CONDITIONS (recognition rates: whole=58.8%, eyes=49%, mouth=40.2%) and EMOTIONS (neutral=71.8%, joy=63.1%, sadness=55.3%, anger=50.6%, surprise=47%, fear=36%, disgust=21.5%) yielded significant main effects, but not FACES (Actor=50.5%, Face1=47.8, Face2=49.6). All the interactions were significant and needed to adequately fit the data. Therefore, a saturated logit model was computed with the correct/ wrong responses as the dependent variable, and FACES, CONDITIONS and EMOTIONS as the independent variables. Comparisons of the performances of the different FACES*CONDITION combinations were accomplished by computing the standardized log-odds ratios from the parameters of the model, along with their Wald confidence intervals (level of confidence: $p<.01$). See [1] for details.

4.2.1 Face-internal comparisons.

For each emotion and each face, we compared the recognitions in the whole condition, to those in the 'eyes' and in the 'mouth' conditions. Table 5 summarises the results (only statistically significant differences are reported).

The recognition of Actor's face in the three conditions shows a limited degree of variation: only in two cases for 'mouth', and only in one for 'eyes' do recognition rates significantly differ from those obtained with whole faces. Concerning the synthetic faces, the variation in much higher than with Actor, in most cases consisting in a superiority of 'whole' with respect to 'mouth'. Moreover, their patterns are very similar.

Table 5. Summary of the face-internal comparison of the recognitions– e=eyes, w=whole, m=mouth

	Actor	Face1	Face2
Disgust	e>w		m>w
Joy		w>e	w>e
Sadness	w>m	w>m w>e	w>m
Fear		w>m w>e	w>m w>e
Anger		w>m	w>m
Surprise	w>m	w>m	w>m
Neutral			m>w

According to these data, the recognition of emotional expressions is quite a robust phenomenon with human faces and much less so with the synthetic ones.

Table 6 summarises the results of the comparison between the recognitions in the 'eyes' and those in the 'mouth' conditions, for each emotion and each face. Bassili's [3] data are also reported (only significant results are reported).

Table 6. Comparison between the recognitions of the eyes (e) and the mouth (m)

	Actor	Face1	Face2	Bassili
Disgust	e>m	m>e	m>e	m>e
Joy		m>e	m>e	m>e
Sadness	e>m		e>m	e>m
Fear		e>m	e>m	e>m
Anger			e>m	
Surprise	e>m	e>m	e>m	m>e
Neutral				not avail.

Our data for Actor don't agree with Bassili's. In particular, Actor never shows any advantage of the mouth over the eyes. At the same time, Bassili's data show a good agreement with the synthetic faces. A possible explanation for these differences

[4] Here negative z-scores indicate that that the Jaccard index is greater for 'whole-mouth' than for 'eyes-whole'.

exploits the role of Ekman's Facial Action Units (FAUs), [8]. The scripts that produce the emotional expressions of our two synthetic faces are based on FAUs decomposition. Moreover, the resulting facial movements are linearly combined with those for lip-speech synchronisation. Hence, it can be expected that the gesture components due to FAUs be perceivable in our synthetic faces. Bassili's stimuli, on the other hand, were the non-uttering faces of actors trained on the use of FAUs. Hence, both with our synthetic faces and with Bassili's, the FAUs component was present, and we can hypothesise that this is what underlies the similarities between Bassili's data and those of our synthetic faces. Finally, the same reasoning could explain also the differences between Actor and both Bassili's data and the synthetic faces.

4.2.2 Cross-face comparisons

Let's take now Actor as a reference category against which to compare the recognitions of the emotions expressed by the other FACES, for fixed CONDITIONS. This will give us a measure of the performances of synthetic faces, and their regions, with respect to the human counterparts. Table 7 summarises the results (only statistically significant differences are reproduced).

Table 7. Comparing the human with the synthetic faces.

	whole	eyes	mouth
Disgust		Actor>Face1 Actor>Face2	Face2>Actor
Joy		Actor>Face1 Actor>Face2	
Sadness	Face1>Actor Face2>Actor	Face2>Actor	Face1>Actor Face2>Actor
Fear	Face1>Actor		
Anger	Actor>Face1 Actor>Face2	Actor>Face1	Actor>Face1 Actor>Face2
Surprise	Face1>Actor Face2>Actor	Face1>Actor Face2>Actor	
Neutral			Face2>Actor

In the 'whole' condition, the two synthetic faces behave very similarly with respect to Actor: both do better on sadness and surprise, worse on anger, and are equal to Actor on the remaining emotions. These data are consistent with those of [6].

The similarities between the two synthetic faces with respect to Actor are slightly weakened when we turn to the eyes. Both faces tend to be worse than Actor on disgust and joy, and better on surprise. Moreover, Face1-eyes does worse than Actor-eyes on anger, and Face2-eyes is superior to Actor-eyes on sadness.

With the mouth, the general pattern does not change much: the two synthetic faces behave in a very similar way with respect to Actor. Both do better than the latter on sadness, and worse on anger. Moreover, Face2 has better performances than Actor on disgust.

The two synthetic faces behave in a very similar way with respect to Actor: in most cases (8 out of 13) the different performances between Actor and one of the two synthetic faces co-occur with similar differences between Actor and the other synthetic face. Moreover, the direction of the differences remains constant: it never happens that for a given emotion Actor is superior to Face1 but inferior to Face2.

5. RESULTS - RECOGNITION ERRORS

5.1 A comparison of error distributions

In this section we study errors, investigating whether the way they distribute is affected by our independent variables: FACES, CONDITIONS and EMOTIONS. We will resort to different techniques than in 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 increased number of combinations would make hard to draw significant conclusions. Moreover, in this section we prefer to use simpler, but easier to manipulate tools to succinctly characterize and compare error distributions. To this end, we will exploit an information-theoretical approach [14] that factors out various contributions to the global information/uncertainty of confusion matrices, turning some of them into the tools we need.[5] Here, we will focus on the effective number of confusion classes, and on the amount of shared errors.

For completeness of information, Table 8 reports the global confusion matrix.

Table 8. Overall confusion matrix (percentages).

	Disg.	Joy	Neut.	Fear	Ang.	Surp.	Sad.
disgust	**21%**	3%	32%	6%	12%	6%	19%
joy	3%	**63%**	15%	2%	7%	4%	6%
neutral	4%	3%	**72%**	1%	6%	3%	10%
fear	12%	1%	6%	**36%**	10%	25%	10%
anger	8%	2%	23%	8%	**51%**	4%	5%
surprise	4%	8%	13%	14%	13%	**47%**	3%
sadness	15%	1%	8%	5%	10%	6%	**55%**

We start from indices d_s and d_r, suggested in [14].[6] The former measures the effective mean number of error (confusion) classes per stimulus, discounting (normalizing for) the error rate. When d_s increases, so do the possible confusion categories for a given stimulus category (presented emotion), hence its ambiguity. d_r, in turn, informs about the mean number of stimulus categories a response category collects confusion from, again normalizing for the error rate. A high value of d_r for a given response class, c, corresponds to a high number of stimulus categories c can collect errors from. Hence, d_r signals the amount of uncertainty as to the identity of the stimulus originating the response. Table 9 reports the values of d_s and d_r for the various FACES*CONDITIONS combinations.

Table 9. Values of d_r and d_s for various FACE*CONDITION combinations

	whole		mouth		eyes	
	d_s	d_r	d_s	d_r	d_s	d_r
Actor	2.26	2.64	2.51	3.6	3.35	3.52
Face1	2.55	3.04	1.92	2.8	2.52	3.64
Face2	2.98	3.46	2.79	3.7	2.55	3.58

[5] The price to pay to the information theoretic approach is the lack of the rich inferential apparatus that other techniques have. Hence we will not be able to anchor our conclusion to tests of statistical significance.

[6] For the formal definition and properties of d_s and d_r, see [14].

It could be expected that, since less information is available, the ambiguity of facial expressions, d_s, and the uncertainty on responses, d_r, increase in the 'eyes' and 'mouth' conditions with respect to 'whole'. On the stimulus dimension, this expectation seems to be fulfilled by Actor, but not by the two synthetic faces, whereby the stimulus ambiguity tends to decrease or to remain stable. ***In other words, the mouth and the eyes of the synthetic faces give rise to stronger biases rather than to increased uncertainty. Along the response dimension, on the other hand, the uncertainty increases in all conditions from the whole to the eyes/mouth presentation.***

To perform direct comparisons across FACE*CONDITION combinations, we exploit two other indices suggested in [14], both computed on pooled confusion matrices. For matrices M_1 and M_2 pooled into $M=M_1+M_2$ the indices δ_s and δ_r yield the effective fraction of errors in M that fall outside the error categories shared by M_1 and M_2. These indices are corrected for the overall differences in the distribution of stimuli, δ_s, and responses, δ_r, and are useful to quantify the extent to which the error distributions of two confusion matrices agree.[7] Suppose that M_1 is the confusion matrix for actor-whole and M_2 that for actor-eyes; δ_s gives the fraction of errors in the pooled matrix that do not belong to stimulus error categories that are common to the actor-whole and the actor-eyes matrices. The higher δ_s, the lower is the number of errors in common confusion classes, hence the more different are the error distributions of M_1 and M_2 along the stimulus dimension. Similarly, δ_r yields the fraction of errors that do not belong to response error categories that are common to actor-whole and actor-eyes. The relevant values for comparisons targeting the various FACE*CONDITION combinations are reported in Table 10. We limit our discussion to δ_s.

Table 10. Value of δ_r and δ_r for whole vs. mouth and whole vs. eyes comparisons

	whole-eyes		whole-mouth	
	δ_r	δ_s	δ_r	δ_s
Actor	.106	.126	.222	.307
Face1	.109	.187	.172	.316
Face2	.119	.194	.224	.389

For any face, the whole-eyes comparison yields lower δ_s values (hence, a higher number of shared errors) than the corresponding whole-mouth pair. In a way, 'eyes' accounts for a very high portion of the errors subjects make when classifying 'whole' faces. Moreover, the lower values of δ_s are those for Actor, and the higher ones those for Face2.

Hence, it is not only the case that the eyes contribute more to the correct recognitions in the 'whole' condition, as we saw above; they also account for a larger portion of errors than the mouth. Furthermore, with synthetic faces the error distributions of the eyes and the mouth differ more from that of 'whole' than with Actor.

Table 11 targets the amount of errors that are not shared between Actor and the two synthetic faces. The former shares similar amounts of errors with the two synthetic faces, especially in the 'whole' and in the 'eyes' conditions. The differences are higher on mouth, where Face2 is closer to Actor than Face1. Interestingly, the condition with the lowest figures, hence the highest amount of shared errors between the two synthetic faces and Actor is 'eyes'.

Table 11 – Errors shared between Actor and the two synthetic faces

	Actor vs. Face1		Actor vs. Face2	
	δ_r	δ_s	δ_r	δ_s
Whole	.449	.531	.425	.507
Eyes	.275	.424	.305	.405
Mouth	.397	.555	.306	.441

5.2 Error classes

In this section we analyse actual error classes, limiting our discussion to the stimulus dimensions.

Table 12 reports the most common error category for each stimulus category.[8] There is a substantial stability of the error classes for Actor, especially with 'whole' and 'eyes'. Moreover, the typical stimulus error classes for Face1-mouth and Face2-mouth are almost identical: similar stimulus classes give rise to similar confusions.

Table 12. Most frequent error categories. Boldface indicates classes that reciprocate.

	Actor			Face1			Face2		
	w	e	m	w	e	m	w	e	m
Dis	sa	sa	sa+a	n	n	n	n+sa	n	n+a
Joy	a	a	a	n	n	n	n	sa	n
Sad	n	**n**	a	d	su	d		d	d
Fe	**su**	**su**	a	d+**su**	**su**	d+sa	su	**su**	d+sa
An		d		n+f	n+f	n	sa	d	n
Su	**f**+a	**f**	n	j+**f**	**f**	n+a	j	**f**	j+n
neu	a	**sa**	sa+a		su	sa	sa	sa	sa

Finally, there are few cases of pairs of emotions that reciprocate their (main) error classes. That is, pairs in which one emotion tends to be mistaken for the other and vice versa. This is the case of fear-surprise for actor-whole: the most common error class for a fear-type stimulus is surprise, and that for surprise is fear. Five of the reciprocal class pairs involve surprise and fear (with actor-whole, actor-eyes, face1-whole, face1-eyes and face2-eyes); one involves neutral and sadness (with actor-eyes). The first pattern is particularly interesting, since it involves all the three faces in the 'whole' and in the 'eyes' conditions (with the exception of Face2-whole). On the one hand, ***this looks as another confirmation of the closeness of the 'eyes' to the 'whole'.*** On the other hand, this datum highlights a pattern of confusion that seems to be independent of the particular face, be it human or synthetic.

6. CONCLUSIONS

This paper had two goals: a) investigating a methodology to assess an important dimension that contributes to define the quality of facial displays — namely, the recognisability of emotional expressions — and b) studying the roles of facial regions in the same task. The first goal was pursued through the systematic comparison of the performances of two synthetic faces (Face1 and Face2) with those of a human model (Actor) on a

[7] For the definitions and formal properties of δ_s and δ_r, see [14].

[8] The methodology is the same as for Table 12. In some cases we reported the two highest ranking categories.

recognition task involving Ekman's six basic emotions plus the neutral expression. As observed, such a procedure makes it possible to address the issue of the naturalness of synthetic faces: the closer the synthetic face to the human model, the more natural it is. In our case, according to Table 7, both faces differ from Actor on 11 out of the 21 possible comparisons involving recognition rates. The analysis of the error distributions reported in Table 11 also indicates a substantial similarity between the two faces in their closeness to Actor. True, Face2 and Actor share more errors than Face1 and Actor, but this is mostly because of the mouth, and much less so with the eyes or the whole face. Finally, Table 5 shows that the recognition task is equally less robust with the two synthetic faces than with the actor. So it seems fair to conclude that our synthetic faces are equally 'natural' according to the proposed definition. This is as it should be, for the two faces have been developed by largely overlapping research groups, and according to very similar principles.

One might expect that at least some of the results presented here depend on the particular choice of the human model. In order for us to be able to use them to inform the development and/or improvement of synthetic faces, we need to know how strong such dependence is. For instance, one might replicate the present study by using a standard database of human expressions, whereby the degree of variation is known in advance and/or controlled for. We have collected such a data-base, comprising expressions from 8 actors, and we are currently investigating its degree of variability.

With respect to the role of facial regions in the recognition task, the eyes turn out to be more important than the mouth along many of the considered dimensions:

- they most often affect the correct recognitions of whole faces, Table 4;
- they more often yield recognition rates that are similar to those obtained with whole faces, Table 5, and higher than those of the mouth, Table 6;
- their error distributions are always closer to those of the 'whole' condition, sharing a higher number of effective error classes with it than the mouth, Table 10;
- the few pairs of reciprocal confusion classes observed occur in 'whole' and 'eyes' on similar emotion pairs (fear and surprise, see Table 12);

The comparison with the data in the literature suggests that the greater importance of the eyes could at least in part be due to the fact that our faces are uttering ones, a condition that might weaken the contribution of the mouth to the expression of emotions. It is also possible that the same factor explains some differences between our recognition rates for the human face and those in the literature. For instance, [3] reports higher recognition rates for human dynamic stimuli (above 70%). In [15] the results varied according to the used data base, but, in general, the tendency was towards higher recognition rates. As far as we know, our study is the first one to consider the role of articulatory movements in a recognition task for emotion, so firm conclusions might be premature. However, the topic is of the utmost interest, especially for applicative purposes, and deserves being pursued through the systematic comparison of face(s) in the uttering and in the non-uttering modes.

Other future lines of investigation concern the role of gender. Our sample had both males and females, but we failed in detecting any significant effect of gender. Other studies, most notably [3], did detect gender-related difference. So this issue requires more effort, addressing not only the effects of the gender of the subjects, but also the effects of the gender of the faces.

Finally, an issue that we have largely left untouched concerns the relationships between objective and subjective measures, for instance those relating to pleasantness and attractiveness, and to the perceived facility to decode emotions. Preliminary data, not presented here, suggest that the latter might be, at least in some cases, related to objective measures, such as recognition rates, but more must be done in this direction.

To this same rubric belongs the issue of the relationships between the operational notion of naturalness exploited here, and the perceived naturalness of emotional expressions. On the one hand, one might suspect the existence of some form of dependency of our measure of naturalness on the particular human model(s) exploited. This issue can be investigated by extending the study presented here to comparisons involving more than one human model. On the other hand, the investigation of the perceived naturalness of emotional expressions requires a systematic comparison of judgments concerning human and synthetic faces. By combining the sets of obtained data, we would then be able to draw conclusions about the relationship between the operational and the subjective notions of naturalness.

7. ACKNOWLEDGMENTS

This work was carried on within the PF-STAR project (http://pfstar.itc.it), and partially supported by grant IST-2001-37599 from the EU. We wish to thank K. Balci, D. Goren-Bar, N. Mana, and M. Zancanaro for their valuable ideas and comments.

8. REFERENCES

[1] Agresti, A. *Categorical Data Analysis*. John Wiley and Sons, New York. 2002.

[2] 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.

[3] Bassili, J. N., Emotion Recognition: The Role of Facial Movement and the Relative Importance of Upper and Lower Areas of the Face. In *Journal of Personality and Social Psychology*, Vol. 37, 2049-2058, 1979.

[4] Cassell, J., Sullivan, J., Prevost, S., Churchill, E. (Eds.): *Embodied Conversational Agents*. Cambridge, MA: MIT Press (2000).

[5] Cosi P., Fusaro A., Tisato G., 'LUCIA a New Italian Talking-Head Based on a Modified Cohen-Massaro's Labial Coarticulation Model'. In *Proceedings of Eurospeech '03*, Geneva, Switzerland, Vol. III, 2269-2272.

[6] Costantini, E., Pianesi, F. and Cosi, P. 'Evaluation of Synthetic Faces: Human Recognition of Emotional Facial Displays'. In E. Andre', L. Dybkiaer, W. Minker, P. Heisterkamp (eds.) *Affective Dialogue Systems*, Springer Verlag. 2004

[8] Eckman P., Friesen W., *Manual for the Facial Action Coding System*, Consulting Psych. Press, Palo Alto (CA), 1977.

[9] 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.

[10] Kätsyri, J., Klucharev, V., Frydrych, M., Sams M. ’Identification of Synthetic and Natural Emotional Facial Expressions’. In *Proceedings of AVSP'2003*, 239-244, St. Jorioz, France. 2003

[11] 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.

[12] Ruttkay, Z., Doorman, C., Noot, H. ‘Evaluating ECAs - What and How?’. In *Proceedings of the AAMAS02 Workshop on Embodied conversational agents - let's specify and evaluate them!*, Bologna, Italy. 2002.

[13] Ruttkay, Z., Doorman, C., Noot, H. ‘Embodied Conversational Agents on a Common Ground. A Framework for Design and Evaluation’. In Ruttkai, Z. and C. Pelachaud (eds.) *From Brows till Trust*.. Kluwer, Dordrecht. 2004

[14] van Son, R.. J. .J. .H.. ‘A Method to Quantify the Error Distribution in Confusion Matrices’. In *Proceedings 18*, 41-63. Institute of Phonetic Sciences, University of Amsterdam. 1994

[15] Vanger, P., Hoelinger, R., Haken, H. ’Computer aided generation of prototypical facial expressions of emotion’. In *Methods of Psychological Research Online*, Vol. 3, 25-38. 1998