ISCA Archive
http://www.isca-speech.org/archive

AVSP 2003 - International Conference
on Audio-Visual Speech Processing
St. Jorioz, France
September 4-7, 2003

Identification of synthetic and natural emotional facial expressions

Jari Kätsyri
Vasily Klucharev
Michael Frydrych
Mikko Sams

Laboratory of Computational Engineering
Helsinki University of Technology, Finland
{katsyri,vasily,frydrych}@lce.hut.fi, Mikko.Sams@hut.fi

Abstract

Identification of emotional expressions of a Talking Head (TH) were evaluated and compared to that of natural faces. In addition, the effect of static (pictures) and dynamic (video sequences) stimuli was studied. Natural stimuli consisted of six basic emotional expressions (anger, disgust, fear, happiness, sadness and surprise). Two expression sets were selected from both Ekman-Friesen facial affect pictures [1] and Cohn-Kanade database [2]. In addition, two new natural expression sets were recorded in our laboratory. Synthetic expressions were created by our new TH [3], both with and without facial texture. Preliminary results indicate that the TH expressions, except fear, were identified as expected. Overall level of identification of TH stimuli was below those of natural ones. Of all used stimuli, happiness was identified the best and fear the worst. Natural static and dynamic expressions were identified equally well. However, the dynamic expressions of the TH were identified significantly more accurately than the static ones.

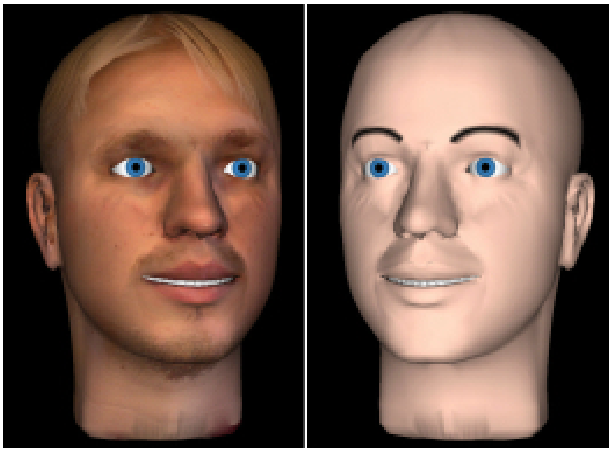

Figure 1 A Talking Head developed at HUT. Picture on the left shows a textured and picture on the right a non-textured version.

1. Introduction

A high-quality database of emotional expressions is important for both studies of emotion perception and for development of synthetic faces. Pictures of facial affects collected by Ekman and Friesen [1], is an example of such a database that contains pictures of several actors posing six basic emotion expressions (anger, disgust, fear, happiness, sadness and surprise). This database was created in 1970's and has remained popular since then. It has been used in neurocognitive studies of facial expression recognition [4]. An important reason for its popularity is that the identification of facial expressions it contains was originally evaluated and verified [1].

Ekman-Friesen database contains only still pictures of expressions. Other databases containing videos of expressions have been created later. One of them is the Cohn-Kanade database [2] of Carnegie Mellon University. Unlike the Ekman-Friesen database, it is available free of charge for non-profit use. The videos in this database have been FACS coded. FACS (Facial Action Coding System) [5] is a comprehensive classification system, where pictures or videos containing facial expressions are coded as Action Units (AUs) by trained human coders. The Cohn-Kanade database contains videos of amateurs posing both single AUs and basic expressions. The database has been used, e.g., in studying automatic recognition of facial expressions [6], but to our knowledge identification of the facial expressions has not been evaluated by human subjects.

If their quality is high enough, synthetic facial expressions created with the Talking Heads (THs) [8] could be used for various applications, including studies of facial expression recognition, instead of natural stimuli. THs have been used earlier in studies of audiovisual speech perception [9, 10]. An important advantage in their use is that they are fully controllable. THs with facial expression research have similar advantages: the stimuli can be created easily, they are fully controllable and contain no superfluous events which are difficult to avoid when using natural stimuli. Stimuli containing, e.g., different head orientations and gaze directions can be created easily. Typically, existing THs express six basic expressions. Expressions are typically based either on MPEG-4 SNHC standard [11] or facial expression "prototypes" described with FACS AU combinations [8].

In the present study, we evaluated identification of facial expressions of natural faces and of our new TH [3] (Figure 1). Textured and non-textured versions of the TH are otherwise

Expression	Created prototype
Anger	AU4+5+7+15+24
Disgust	AU9+10+17
Fear	AU1+2+4+5+7+20+25
Happiness	AU6+12+25
Sadness	AU1+4+7+15+17
Surprise	AU1+2+5+25+26

Table 1 FACS prototypes for modelled expressions.

identical, but a picture of a real face has been mapped on the surface of the generic TH. In the expression mechanism of the TH a subset of FACS AUs has been modeled. The six basic expressions are defined as FACS prototypes based on literature [5, 8, 12, 13, 14], as shown in Table 1. Also blends of basic expressions have been created and implemented, but they were not evaluated in this study.

We aimed at finding out the quality of the expressions of our TH in comparison to those found in available natural databases. We were also interested in seeing whether identification of static and dynamically presented expressions differ and whether texture influences the identification of the TH expressions.

2. Methods

2.1. Subjects

Subjects were 55 (37 males, 18 female; 20-29 years old) Finnish students from the Helsinki University of Technology (HUT) who participated in the experiment as a part of their studies.

2.2. Stimuli

Stimuli were static and dynamic, synthetic and natural facial expressions. Dynamic stimuli were short (1-1.5 s) video sequences showing the expression from neutral position to its apex (maximum). A video set contained six basic expressions recorded from one person. However, one set contained expressions from several persons as described below. A picture set contained the six basic expressions and a neutral face.

Synthetic videos and pictures were created by our TH. There were two sets of TH expressions: non-textured and textured.

We also constructed a new HUT video expression database. Six basic expressions were recorded from two actors. The expressions were based on the same expression prototypes as were used in the expression model of TH (see Fig. 1). Both actors were certified FACS coders, and trained in controlling facial muscles associated with FACS AUs.

The used natural expressions were from the Ekman-Friesen (EF), the Cohn-Kanade (CK) and the HUT databases. Two well-identified picture sets of basic expressions were selected from the Ekman-Friesen database. Each set contained six expressions of one person. Two sets of both pictures and videos were selected from the CK database. Pictures were frames from the videos where expression was at its apex. Unfortunately we did not find sets of all basic expressions suitable to our study from the same actor, so we chose two suitable items for each basic expression. Two sets of videos and pictures contained items from total of five persons. The HUT database contains two full video sets and pictures of expressions from two persons. In total, there were four video sets of natural stimuli from two databases (CK and HUT) and six picture sets from three databases (EF, CK and HUT).

2.3. Procedure

The subjects were divided into two groups. For one group of 28 subjects, only pictures were shown. This group saw both emotional and neutral facial pictures. In total this group saw 56 pictures (8 picture sets × 7 expressions). The other group of 27 subjects saw the dynamic stimuli, but also pictures of EF faces. The second group saw in total 48 videos and pictures (6 video sets × 6 expressions + 2 picture sets × 6 expressions).

Stimuli were presented with Presentation software (version 0.53) [15]. The order of pictures and videos was randomized. The subjects had to evaluate how much each stimulus contained each of the basic emotional expressions: happiness, sadness, surprise, anger, disgust and fear. In addition, they were asked to evaluate the naturalness of the stimuli. Evaluations were made on a scale from 1 to 7 (1 = totally disagree, 4 = uncertain, 7 = totally agree) and given by keyboard. The questions were shown one by one in a randomized order while the stimulus was shown. Stimuli were presented in a loop until all the questions were answered. There was no time limit for responses.

Before the actual experiment the subjects participated in a training session consisting of two randomly selected stimuli.

2.4. Analysis

Evaluation of an expression on the used scale is influenced by their intensity. Because the intensities of the used emotional expressions were not matched across datasets, the evaluations were classified according to the "winner expression category", i.e. according to the expression a subject gave the highest rating. Ekman and Friesen [1] used a similar approach. If a subject gave the same rating to several expressions, the stimulus was not classified to any expression category. In total, 20% of all the stimuli (18% excluding neutral pictures) were not classified.

One-way repeated measures analyses of variance (ANOVAs) were used to compare number of correct identifications between different databases and between different expressions. These analyses were conducted separately for pictures and videos. Two-way mixed design ANOVA was used to compare results for databases between pictures and videos. A two-way ANOVA was conducted to compare results for TH expressions between pictures and videos. Original evaluations for neutral pictures of textured and non-textured TH were compared with an additional one-way ANOVA. In all these cases, planned comparisons were used

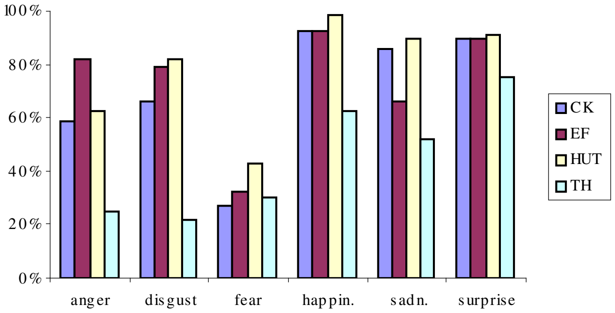

Figure 2 Identification of static expressions of TH and actors from different databases.

to compare results, either between expressions, databases or evaluations. Sign tests were used, separately for pictures and videos, to compare results for expressions between textured and non-textured TH.

3. Results

3.1. Pictures

Figure 2 shows how subjects identified the static expressions. Results from two stimulus sets have been averaged for each database. In general, subjects identified the expressions as expected. Scores for the intended (by the experimenters) expressions were higher than that for the other expressions with all natural pictures. However, the fearful expression of TH was more often (34%) identified as surprise.

	anger	disg.	fear	happ.	sadn.	surpr.
anger	--	**0.280**	0.000	0.000	0.001	0.000
disgust	**0.280**	--	0.000	0.000	0.015	0.000
fear	0.000	0.000	--	0.000	0.000	0.000
happiness	0.000	0.000	0.000	--	0.004	**0.922**
sadness	0.001	0.015	0.000	0.004	--	0.005
surprise	0.000	0.000	0.000	**0.922**	0.005	--

Table 2 p-values for differences between identification results of static expressions. Identification results have been averaged over all databases. Non-significant values are highlighted.

Table 2 shows significance values for differences between static expressions. Identification of both happiness and surprise is significantly better than that of all other expressions. The difference between them is not significant. Identification of fear is significantly poorer than that of the other expressions. There was no significant difference between identification of anger and disgust or happiness and surprise.

In general, HUT expressions were identified better than CK expressions (F(1,27)=6.92, p=0.011). There were no significant differences in identification of HUT and EF databases, or EF and CK databases. Identification of TH expressions was much worse than that of natural faces (F(1,27)=151.47, p<0.0001).

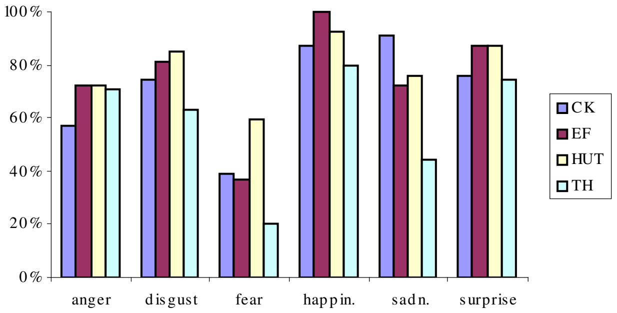

Figure 3 Identification of dynamic expressions of TH and actors from different databases.

3.2. Videos

Identification of dynamic expressions is shown in Figure 3. Results for Ekman pictures are also included for comparison. All natural expressions were identified as expected, but TH's fearful expression was identified more often (37%) as surprise.

	anger	disg.	fear	happ.	sadn.	surpr.
anger	--	**0.115**	0.000	0.000	**0.577**	0.010
disgust	**0.115**	--	0.000	0.006	**0.307**	**0.307**
fear	0.000	0.000	--	0.000	0.000	0.000
happiness	0.000	0.006	0.000	--	0.000	**0.079**
sadness	**0.577**	**0.307**	0.000	0.000	--	0.042
surprise	0.010	**0.307**	0.000	**0.079**	0.042	--

Table 3 p-values for differences between identification results of dynamic expressions. Identification results have been averaged over all databases. Non-significant values are highlighted.

Table 3 shows significance values for differences between dynamic expressions. Identification of happiness is significantly better than that of all other expressions, except for surprise. Identification of fear is significantly poorer than that of the other expressions. There was no significant difference between identification of surprise and disgust or anger, sadness and disgust.

In general, expressions were identified better from HUT than from CK database (F(1,26)=5.67, p=0.025). There were no significant differences in identification of HUT and EF databases, or EF and CK databases. Identification of TH

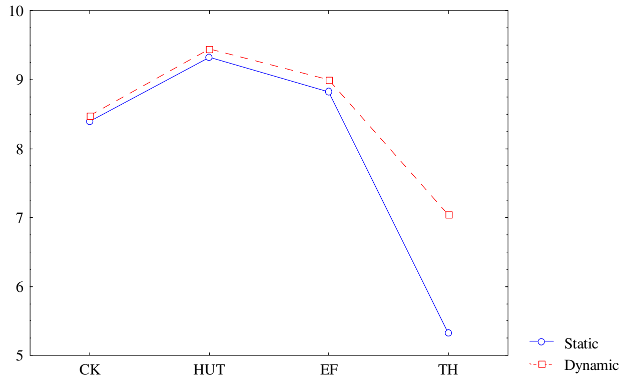

Figure 4 Number of correct responses for pictures and videos of different databases.

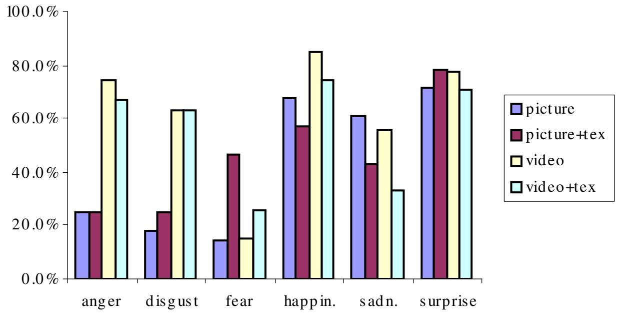

Figure 5 Identification of expressions from pictures and videos of textured and non-textured version of TH.

expressions was significantly worse than that of natural ones (F(1,26)=57.05, p<0.0001)

3.3. Pictures vs. videos

Figure 4 shows identification of static and dynamic expressions of TH and actors from the different databases. Difference in identification of static and dynamic expressions was significant (F(3,159)=4.76, p<0.004). This difference was due to much better identification of dynamic than static expressions of TH (F(1,53)=13.15, p<0.001). Expressions from the different databases were equally well recognized with static and dynamic stimuli.

The identification of dynamic and static TH expressions was significantly different (F(5,265)=7.25, p<0.0001). There is a clear increase (cf. Figure 5) in identification of anger (F(1,53)=23.51, p<0.0001) and disgust (F(1,53)=17.93, p<0.0001).

3.4. Textured vs. non-textured TH

Figure 5 depicts identification of textured and non-textured TH expressions. Only significant difference was found with static fearful expression, which was identified better when textured (Sign test, p=0.027). With dynamic expressions there were no signficant differences.

Figure 6 shows average scores (original scale from 1 to 7) for textured and non-textured TH neutral pictures. Textured neutral pictures got significantly higher scores for anger (F(1,27)=9.77, p=0.004), disgust (F(1,27)=4.46, p=0.044) and fear (F(1,27)=13.75, p<0.0001).

4. Discussion

Both CK and HUT expressions were found to be competitive with those in EF database. EF database is a good reference, because it has become almost a standard in the field of facial expression research.

In comparison to the evaluation study made by Ekman and Friesen [1], the facial expressions of their database were identified dramatically less accurately in this study. This may be explained by different experiment setup. In the original study the subjects saw stimulus for 10 seconds before evaluating it, whereas in the present study the evaluations were given while examining the stimulus. In addition, in the present study the questions were presented in a random order and the subjects were not asked to be consistent between their ratings for the same stimulus. Our decision of classifying all evaluations from a subject into one "winner expression" could also be criticized. This kind of analysis reduces the effect of expression's intensity on the scores, but on the other hand, it also omits other important information, such as how unambiguously the expression was evaluated.

The TH expressions, except fearful ones, were identified as intended, especially from videos. The recognition was still clearly worse in comparison to all natural stimuli. Taking both natural and synthetic stimuli in consideration, dynamic and static expressions of happiness were identified clearly better than other ones. Good identification of happiness is in accordance with hypothesis of top-level discrimination between happy and unhappy expressions [7].

Identification of fear from both natural and synthetic stimuli was clearly the worst. In addition, the fearful expression of TH was clearly unsuccessful, because it was identified more often as surprised than afraid. These results are compatible with earlier results [7] considering the poor identification of fear and it's confusion with surprise in classification tasks involving typical basic expression labels. The reason why fear is often misinterpreted as surprise may be in the perceptual similarities between expressions. In both, eyes are wide open, eyebrows are raised and mouth is opened, even if there are differences in these effects [13] (also cf. Table 1). On the other hand, Adolphs [7] reports that fear is fairly easily discriminated from other expressions in tasks involving comparison of similarity between two pictures. It seems that the source of confusion is not only the similar feature set. In our normal daily life we probably see surprised expression much more often than fearful ones. When expressions are relatively similar, it is a good strategy to classify them according to their likelihood in a certain situation.

TH's dynamic angry and disgusted expressions were identified much better than the static ones. Such a difference was not found with natural stimuli. TH expressions were of

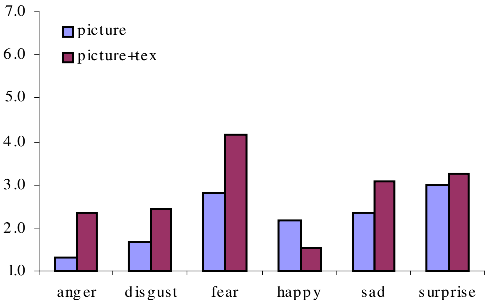

Figure 6 Average evaluations of TH's textured and non-textured neutral expressions.

quite low intensity and it may be that seeing the static expressions did not give enough hints for recognizing the intended expressions. However, with videos the transition from neutral picture probably provided a baseline for the comparison of the apex value, and hence made identification of expression easier.

It is uncertain why textured static expression of fear was recognized better than the non-textured one. This result may be due to effects of facial texture even on the neutral face. Same model of eyes was used both with textured and non-textured TH, so adding the facial texture may have changed contrast between facial mesh and eyes. With textured version the white part of the eye may have been more pronounced, thereby creating a slight appearance of fear. Seeing the movement dynamics in videos may have cancelled this effect.

5. Conclusions

Comparison of how the six basic expressions of our TH were identified in comparison to natural stimuli was the main issue in this study. TH's fearful expression was evaluated as being more surprised than afraid. All the other expressions were identified as expected. However, identification of expressions from our TH were below that of natural expressions. Part of the poor identification may be explained by the low intensity of the modeled expressions. However, also the emotion model needs improvement.

Seeing the movement of facial expression from neutral face to the apex of expression improved the identification of expression drastically with TH. With natural stimuli there was no such consistent effect. This implicates that seeing the dynamics of facial expression is important when the facial expression is of low intensity or when there are other factors that make expression more difficult to recognize.

Considering both natural and synthetic stimuli, happy expressions were identified the best and fearful expressions clearly the worst. These results were in accordance with earlier findings.

6. Acknowledgements

This study was supported by the Academy of Finland grant 44897 to the Center of Excellence of Computational Science and Engineering and grants 49881 and 49900 and to M. Sams. J. Kätsyri is supported by the Ministry of Education fund to the graduate school HeCSE.

7. References

[1] Ekman, P. and Friesen, W., *Pictures of Facial Affect*, Consulting Psychologists Press, Palo Alto, CA, 1978.

[2] Kanade, T., Cohn, J. F., Tian, Y., ``Comprehensive Database for Facial Expression Analysis", *4th IEEE Int. Conference on Automatic Face and Gesture Recognition Proc.*, 2000, p 46 - 53.

[3] Frydrych, M., Kätsyri, J., Dobšík, M., Sams, M., ``Toolkit for animation of Finnish Talking Head", *AVSP 2003*.

[4] Calder, A. J., Lawrence, A. D., Young, A. W. "Neuropsychology of fear and loathing", *Nature Reviews Neuroscience 2 (5)*, 2001, pp. 352-63.

[5] Ekman, P., Friesen, W., Hager, J., *Facial Action Coding System*, Consulting Psychologists Press, Palo Alto, CA, 1978.

[6] Cohn, J. F., Zlochower, A., Lien, J., and Kanade, T. "Automated face analysis by feature point tracking has high concurrent validity with manual FACS coding", *Psychophysiology*, 36, 1999, p 35-43.

[7] Adolphs, R. "Recognizing emotion from facial expressions: psychological and neurological mechanisms", *Behavioral and Cognitive Neuroscience Reviews 1*: 21-61.

[8] Prevost, S. and Pelachaud, C., *Talking Heads: Physical, Linguistic and Cognitive Issues in Facial Animation*, Course Notes for Computer Graphics International 1995, Leeds, UK, 1995.

[9] Olives, J.-L., Möttönen, R., Kulju, J. and Sams, M., ``Audio-visual speech synthesis for finnish". *AVSP 1999 proc.*, USA, 1999, pages 157-162.

[10] Vatikiotis-Bateson, E., Kroos, C., Kurarate, T., Munhall, K. G., Rubin, P., Yehia, H. C., ``Building talking heads: Production based synthesis of audiovisual speech.", First IEEE-RAS Int. Conference on Humanoid Robots proc., MIT, Cambridge, MA.

[11] Tekalp, M. and Ostermann, J. "Face and 2D Mesh Animation in MPEG-4", *Image Communication Journal*, Tutorial Issue on MPEG-4 Standard, Elsevier, 2000.

[12] Ekman, P., Friesen, W., Hager, J., *Facial Action Coding System: Investigator's Guide*, Consulting Psychologists Press, Palo Alto, CA, 1978.

[13] Ekman, P., Friecen, W., *Unmasking the face. A guide to recognizing emotions from facial expressions*, Consulting Psychologists Press, Palo Alto, CA, 1975.

[14] Faigin, G., *The Artist's Complete Guide to Facial Expression*, Watson-Guptill, New York, 1990.

[15] Neurobehavioral Systems. Webpage: http://nbs.neuro-bs.com/