*To Appear in the Proceedings of Intelligent Virtual Agents (IVA) 2006*

# Perceiving Visual Emotions With Speech

Zhigang Deng [1], Jeremy Bailenson [2], J.P. Lewis [3], and Ulrich Neumann [4]

[1] Department of Computer Science, University of Houston, Houston, TX
deng@zhigang.org
[2] Department of Communication, Stanford University, CA
[3] Computer Graphics Lab, Stanford University, CA
[4] Department of Computer Science, University of Southern California, Los Angeles, CA

**Abstract.** Embodied Conversational Agents (ECAs) with realistic faces are becoming an intrinsic part of many graphics systems employed in HCI applications. A fundamental issue is how people visually perceive the affect of a speaking agent. In this paper we present the first study evaluating the relation between objective and subjective visual perception of emotion as displayed on a speaking human face, using both full video and sparse point-rendered representations of the face. We found that objective machine learning analysis of facial marker motion data is correlated with evaluations made by experimental subjects, and in particular, the lower face region provides insightful emotion clues for visual emotion perception. We also found that affect is captured in the abstract point-rendered representation.

## 1 Introduction

Embodied Conversational Agents (ECAs) [2, 9, 20, 21, 28, 29, 36, 39] are important to graphics and HCI communities. ECAs with emotional behavior models have been proposed as a natural interface between humans and machine systems. The realism of facial displays of ECAs is one of the more difficult hurdles to overcome, both for designers and researchers who evaluate the effectiveness of the ECAs.

However, despite this growing area of research, there currently is not a systematic methodology to validate and understand how we humans visually perceive the affect of a conversational agent. As ECAs become more and more prevalent in HCI systems, understanding the usability of them as well as the significance of different face representations is clearly a priority. In this work, we aim to answer the following questions:

- Are the results from objective analysis of facial marker motion and subjective evaluation of recorded face video clips consistent?
- Does abstract point-rendered facial animation provide cues for visual emotion perception? And is it a useful representation for ECAs?

- Which emotion pairs are easily confused when people perceive emotions from visual talking faces?

To answer the above questions, we investigated the problem of 'visual perception of emotion in speech' using a multifaceted and comparative approach. An actress with markers on her face was directed to recite specified sentences with four basic emotions (neutral, happiness, anger, and sadness). A facial motion capture system captured the 3D motions of the facial markers, while a video camera simultaneously also recorded her face. We analyzed these captured objective motion data using Quadratic Discriminant Analysis (QDA) [22], and conducted subjective evaluation experiments on both the recorded face video clip and a unique rendering of just the facial markers themselves (termed the point-rendered representation in this work). We conducted the above analysis and experiments on different face regions (whole face, the upper face, the lower face). Finally, we did a comparative analysis on the objective/subjective results and considered the implications of our findings.

To our knowledge, this work is the first to investigate the consistency among objective facial marker motion, subjective evaluations of real talking face videos, and abstract point-rendered faces. By combining the objective captured motion analysis and two types of subjective evaluations, we obtain a comprehensive and multifaceted view on the problem of visual emotion perception with speech. Furthermore, we believe that using an abstract rendering of the emotions (i.e., the markers without the face) should provide insight into the ways in which people learn emotions from faces rendered from a finite set of control points. The goal of the current paper is not to learn new theoretical aspects of how people emote, but alternatively to begin to provide a methodological framework for interpreting emotions—whether those emotions are generated by physical faces, digital avatars from real-time tracked humans, or from embodied agents that use algorithms to create emotions.

The remainder of this paper is organized as follows: Section 2 reviews previous work connected emotion perception from human faces. Section 3 describes data capture and the experiment design. Section 4 describes of the analyses of the objective motion capture data. Section 5 describes the subjective evaluation that includes two subjective studies—one in which people attempted to judge emotions from the video of the actress, and one in which they attempted to perceive emotions from a rendering of just the markers themselves without the face. Section 6 gives an in-depth analysis of both the subjective and objective results and their correlations. Finally, Section 7 concludes this work and discusses implications of the results.

## 2 Background and Related Work

In the computer graphics and computer vision communities, extensive computer facial animation research has been done since Parke's seminal work [33]. Examples of these advances can be seen in work generating realistic speech animation [27, 34, 11, 6, 5, 16, 19, 24] and expressive facial animation [14, 15, 35, 8, 2, 31, 10, 4, 17, 26]. For example, the Facial Action Coding Systems (FACS) proposed by Ekman and Friesen [17] is a widely used system to represent various human expressions by

combining basic facial action units. Essa and Pentland [18] extended FACS to encode the temporal and spatial information by modeling the dynamics of facial expressions, for the purpose of analysis and recognition of facial expressions. The work of [37,38] demonstrated the success of using eigen-faces for the characterization of human faces and face recognition.

In the HCI community, researchers have conducted quite a few experiments to evaluate the effects of using ECAs as a human/machine interface. Walker et al. [25] investigated the benefits of synthetic talking faces in the context of a questionnaire study and found that talking faces made subjects spent more time, made fewer mistakes and gave more comments. Nass et al. [30] compared performance differences when people interacted with their own face or alternatively someone else's face, and found that subjects showed more positive responses when seeing their own faces. Panzdic et al. [32] evaluated and compared the performance of different synthetic talking faces for interactive services. The work of [1, 23] assesses the emotional recognizability of synthetic faces based on the FACS [17], in terms of subjective recognition rates.

Instead of only using recorded expression sequences without utterances [1, 23], Costantini et al. [12] comparatively evaluated two MPEG-4 synthetic faces (motion capture based and script-based) that spoke with emotions. Two synthetic FAP-based talking faces were evaluated under both conditions, including cross-face comparisons, and comparisons with a human actor. The results indicated that motion-capture based synthetic faces generated more natural results than script based faces. Bassili [3] investigated the role of different facial regions for emotion perception tasks only with subjective experiments. He found that the importance of a region of the face differed depending on the emotion. Costantini et al. [13] not only looked at emotion recognition rates, but also analyzed recognition errors.

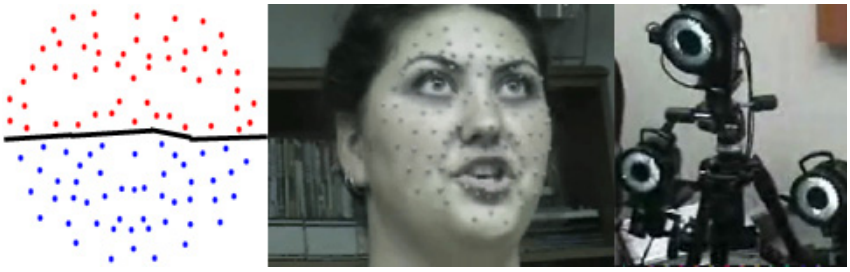

**Fig 1**. The left panel illustrates the marker layout used in the objective data analysis. Markers above the solid curve represent markers in the upper face region, and markers above the solid curve are for the lower face region. The middle panel is a snapshot of the record video. The right panel illustrates the used motion capture system.

## 3 Data Capture and Experiment Design

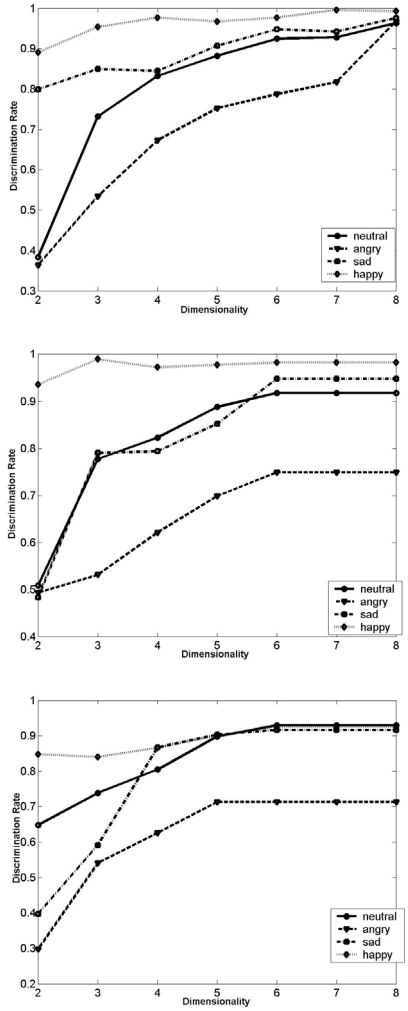


**Figure 2**: The plotting of correct discrimination rate VS reduced dimensionality. The top is for WHOLE-FACE, the middle is for UPPER-FACE, and the bottom is for LOWER-FACE.

A motion capture system was used to capture accurate 3D facial marker motions of a single actress with markers on her face (Figure 1). She was directed to recite three specified sentences four times, and each time a different emotion (from a total of four: neutral, anger, sadness, and happiness) was expressed naturally, without exaggeration. The three sentences are: "*You're truly impossible!*", "*How do you think you're going to catch up now?*", and "*So the sock struck your head and injured your pride?*" The motion capture system tracked and recovered 3D motion of every marker at a 120Hz sampling frequency. At the same time, an off-the-shelf video camera recorded her face. We collected data from a single actress (as opposed to getting a larger sample of faces) because our interest was in comparing methodologies of emotion validation, as opposed to discovering universal emotional patterns across people.

We analyzed the data in both objective and subjective ways. Objectively, we examined results from a discriminant analysis of the recorded expressive facial motion data and distributions of that data in a high dimensional motion space. We performed subjective evaluations by running experiments in which subjects identified and rated various types of stimuli based on the recorded video footage and recorded 3D marker motion without the video behind it. Audio was removed from both stimuli, and the various emotions were played in a random order for experimental subjects.

## 4 Objective Motion Discrimination

After the facial motion data were captured, we preprocessed the motion data by aligning motion frames with a chosen reference frame. The alignment included translation (anchoring a nose marker to be the center of each frame) and head motion removal using a statistical shape-analysis method [7]. Hence, all motion frames were located in a uniform coordinate system.

The aligned expressive facial motion data were processed to generate three groups: whole face motion data (WHOLE-FACE), upper face motion data (UPPER-FACE), and lower face motion data (LOWER-FACE). Each of the above three groups has 12 motion capture sequences: 3 sentences*4 expressions = 12. Each frame of UPPER-FACE is composed of only the motions of the markers in the upper face region (the red markers in the left panel in Figure 1), and each frame of LOWER-FACE is for the motions of the markers in the lower face region (the blue markers in the left panel in Figure 1). For each group, Principal Component Analysis (PCA) was applied to reduce original high dimensional motion vectors (concatenating 3D motions of markers in one frame into one vector) into low dimensional vectors.

To observe how these four expressive motions were further discriminated while the reduced dimensionality was varied, a Quadratic Discrimination Analysis (QDA)

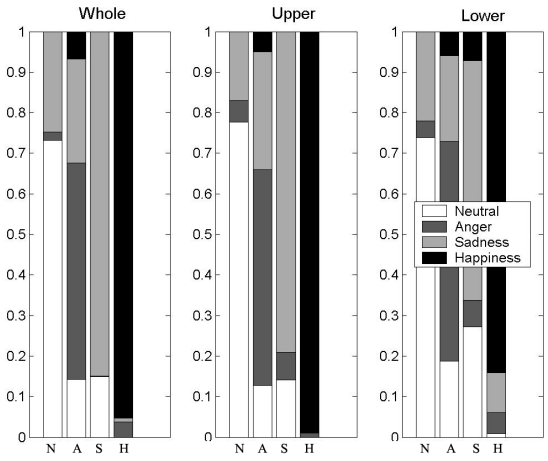


**Figure 3**. Plot of emotion confusion matrices of objective data analysis. Here N denotes "Neutral", A denotes "Anger", S denotes "Sadness", and H denotes "Happiness". The left is for the whole face, the middle for the upper face region, and the right for the lower face region.

method [22] was used to discriminate expressive motion. As illustrated in Figure 2, there is a big jump in discrimination rate when the dimensionality is increased from two to three. In all three cases (*WHOLE-FACE*, *UPPER-FACE* and *LOWER-FACE*), happiness was the easiest emotion to classify and anger was the most difficult one.

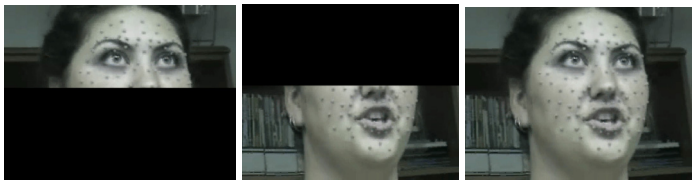

**Figure 4**. Three versions are used in the subjective evaluation.

| | Whole Face | | | | Upper Face | | | | Lower Face | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | N | A | S | H | N | A | S | H | N | A | S | H |
| **N** | 0.732 | 0.020 | **0.248** | 0 | 0.777 | 0.053 | **0.170** | 0 | 0.739 | 0.041 | **0.220** | 0 |
| **A** | 0.142 | 0.535 | **0.256** | 0.067 | 0.128 | 0.532 | **0.290** | 0.050 | 0.188 | 0.541 | **0.211** | 0.060 |
| **S** | **0.149** | 0.001 | 0.850 | 0 | **0.142** | 0.068 | 0.790 | 0 | **0.272** | 0.065 | 0.592 | 0.071 |
| **H** | 0 | 0.038 | 0.008 | 0.954 | 0 | 0.011 | 0 | 0.989 | 0.008 | 0.053 | 0.098 | 0.841 |

**Table 1**: Emotion confusion matrices of QDA objective data analysis. Bold numbers highlight these easily confused visual emotion pairs.

Based on the above QDA results, we calculated emotion confusion matrices that represent how expressive motions were confused with each other. Figure 3 and Table 1 show the confusion matrices (the reduced dimensionality is 3). In all three cases (the whole face, the upper face and the lower face) anger was easily misidentified as sadness, and sadness and neutral were easily confused with each other. Bold numbers in Table 1 highlight these easily confused visual emotion pairs.

**Figure 5.** Illustrations of three versions of rendered point motion snapshots (whole face points, upper face points, and lower face points)

## 5 Subjective Experiment Evaluations

In this section, we used two types of video clips: video clips (*video faces*) directly recorded during the capture session, and 3D point motion clips (*point-rendered faces*) made by simply rendering captured markers' 3D motion. For the ordinary video clips, we first removed audio. Then, for each video clip, we made three versions: a video clip only showing the upper face region (*UPPER-VIDEO*), a video clip only showing the lower face region (*LOWER-VIDEO*), and one showing the full face (*WHOLE-VIDEO*). The partitioning scheme of a face is illustrated in Figure 4. These black

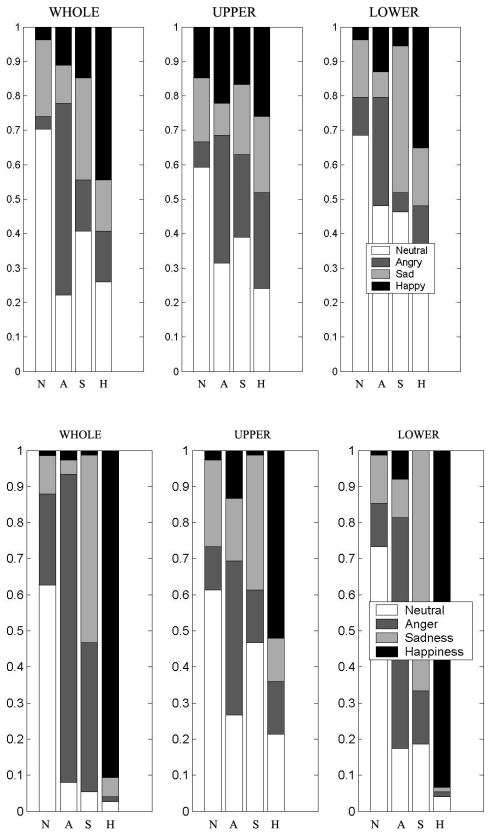


**Figure 6**. The illustration of the confusion matrices of face regions for different emotions in the POINT REGION EXPERIMENT (top) and the VIDEO REGION EXPERIMENT (bottom).

masks were superimposed by video editing tools. The subjective evaluation of these video clips is referred to as the ***video region experiment***. For the clips based on only the markers, we directly plotted marker motion using Matlab to generate point-motion clips (Figure 5). As in other conditions, there was no audio. Then we made three versions of point-motion video by only showing motion of specified markers: a motion clip only showing the upper face markers (*UPPER-POINT*), a motion clip only showing the lower face markers (*LOWER-POINT*), and the one showing the full face markers (*WHOLE-POINT*). The subjective evaluation of point motion clips is referred to as the ***point region experiment.***

A subjective evaluation experiment was set up in a university classroom. A projector was used to project the video onto a large screen. The experiment was composed of several sub-experiments: video region experiment (including UPPER-VIDEO, LOWER-VIDEO, and WHOLE-VIDEO), and point region experiment (including UPPER-POINT, LOWER-POINT, and WHOLE-POINT). Within each sub-experiment, corresponding clips were randomly played to a total of 25 subjects. The same group of subjects participated in all of the above experiments. The 25 subjects were undergraduate

| | **Whole Face** | | | | **Upper Face** | | | | **Lower Face** | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | N | A | S | H | N | A | S | H | N | A | S | H |
| **N** | **0.627** | 0.253 | 0.107 | 0.013 | **0.613** | 0.120 | 0.240 | 0.027 | **0.733** | 0.121 | 0.133 | 0.013 |
| **A** | 0.080 | **0.853** | 0.040 | 0.027 | 0.267 | **0.427** | 0.173 | 0.133 | 0.173 | **0.640** | 0.107 | 0.080 |
| **S** | 0.053 | 0.413 | **0.521** | 0.013 | 0.467 | 0.147 | **0.373** | 0.013 | 0.186 | 0.147 | **0.667** | 0 |
| **H** | 0.027 | 0.013 | 0.053 | **0.907** | 0.213 | 0.147 | 0.120 | **0.520** | 0.041 | 0.013 | 0.013 | **0.933** |

**Table 2**: Emotion confusion matrices of the *video region experiment*.

| | Whole Face | | | | Upper Face | | | | Lower Face | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | N | A | S | H | N | A | S | H | N | A | S | H |
| **N** | **0.704** | 0.037 | 0.222 | 0.037 | **0.593** | 0.074 | 0.185 | 0.148 | **0.685** | 0.111 | 0.167 | 0.037 |
| **A** | 0.222 | **0.556** | 0.111 | 0.111 | 0.315 | **0.370** | 0.093 | 0.222 | 0.481 | **0.315** | 0.074 | 0.130 |
| **S** | 0.407 | 0.148 | **0.296** | 0.149 | 0.389 | 0.241 | **0.204** | 0.166 | 0.463 | 0.056 | **0.425** | 0.056 |
| **H** | 0.259 | 0.148 | 0.148 | **0.445** | 0.241 | 0.279 | 0.222 | **0.259** | 0.296 | 0.185 | 0.167 | **0.352** |

**Table 3**: Emotion confusion matrices of the POINT REGION EXPERIMENT.

student volunteers who were majoring in various disciplines such as engineering and psychology, and fourteen were female. In each experimental session, a group of participants saw each emotion clip once, and attempted to determine its perceived emotion from four possible options (neutral, anger, sadness and happiness). At the same time, he/she filled in a confidence level for this choice on a scale from 1 to 5 (1 meant "*random guess"* and 5 meant "*very sure"*). They could choose the same response option as often as they wanted.

We plotted emotion recognition rate versus facial regions to see what the associations were between visual emotion perceptions and upper or lower facial regions when the audio was removed. Figure 6 and Table 2-3 illustrate the confusion matrices for the point region experiment and for the video region experiment. As we can see from Fig. 6, for neutral, sadness and anger, the point region and video region results were consistent in some places. Since the point motion clips did not provide as much information as the videos, the perception rate on point motion clips was lower than that of corresponding video clips (bold numbers in Table 2-3). However, emotion recognition and confusion patterns (Figure 6) were similar across the two different experiments, indicating that subjects were in fact inferring meaningful emotional cues from the limited information

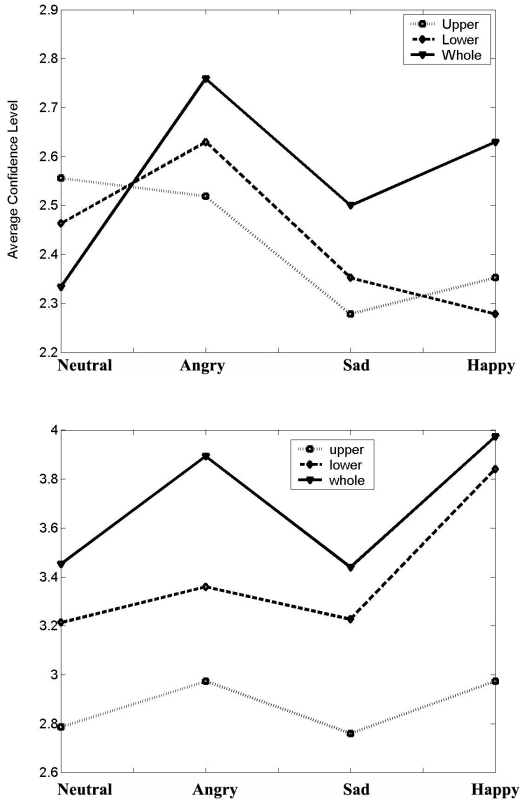


**Figure 7**. The plotting of average confidence level vs emotion types in the POINT REGION EXPERIMENT (top) and the VIDEO REGION EXPERIMENT (bottom). As we can see from the figure, their patterns are similar.

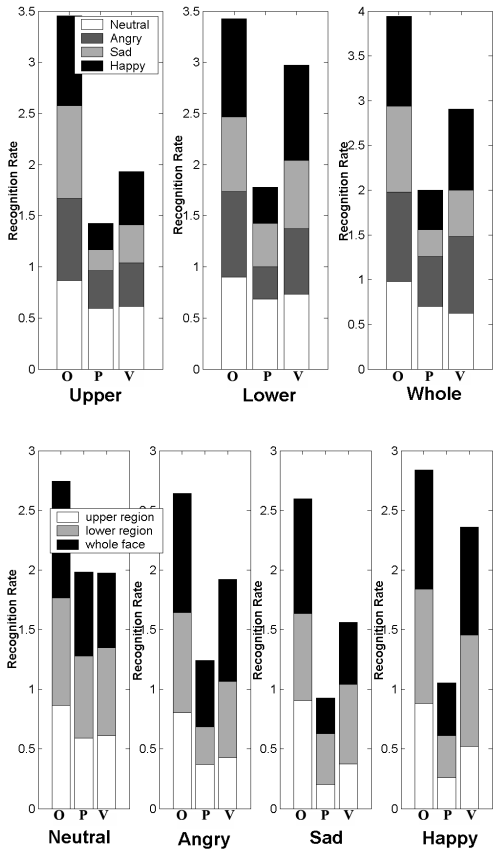


**Figure 8.** (Top) comparisons of different face regions on emotional perception. (Bottom) emotion perception rate depended on emotion type. Here O = QDA mocap data objective analysis, P = point region experiment, and V = video region experiment.

provided by the point clouds.

Both the point region experiment and video region experiment (Figure 6) again confirmed that sadness and neutral were easily confused. It is interesting to notice that happiness was the most easily recognized emotion in the video clips, but was more difficult to recognize from the point motion clips.

Figure 7 illustrates the average confidence level in the video region experiment (right) and the average confidence level in the point region experiment (left). For all emotions, in both the video and point region experiments, generally the average confidence level of the whole face was higher than that of the lower face region, and the average confidence level for the lower face region was higher than the upper face region. As we can see from Figure 7, the point region experiment results share generally similar patterns to the video region experiment.

## 6 Comparative Analysis

Figure 8 compares the importance of different facial regions in terms of three experimental conditions (motion capture data objective analysis, point region experiment, and video region experiment). As we can see from Figure 8, generally, the mocap objective analysis approach recognizes emotions best, except for the

neutral expression. Most likely, subjects put "neutral" as their default choice when unsure, and this caused the high recognition rate for neutral in the point and video region experiments. The right panel of Figure 8 illustrates that recognition rates depend on the emotion, with happiness and anger being more recognizable than sadness, while the neutral emotion achieves the highest perception accuracy.

We further investigate the differences between the previous confusion matrices (Table 1-3) generated by three different evaluation approaches. Table 4 shows

| | **Whole Face** | | | | **Upper Face** | | | | **Lower Face** | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | N | A | S | H | N | A | S | H | N | A | S | H |
| **N** | 0.105 | **-0.23** | **0.141** | -0.01 | 0.164 | -0.07 | -0.07 | -0.03 | 0.006 | -0.08 | 0.087 | -0.01 |
| **A** | 0.062 | **-0.32** | **0.216** | 0.040 | -0.14 | 0.105 | 0.117 | -0.08 | 0.015 | 0.099 | 0.104 | -0.02 |
| **S** | 0.096 | **-0.41** | **0.329** | -0.01 | **-0.33** | -0.08 | **0.417** | -0.01 | 0.086 | -0.08 | -0.08 | 0.071 |
| **H** | -0.03 | 0.025 | -0.05 | 0.047 | **-0.21** | -0.14 | -0.12 | **0.469** | -0.03 | 0.040 | 0.085 | -0.09 |

**Table 4**: Difference matrices obtained by subtracting emotion confusion matrices of the *video region experiment* from that of *QDA objective motion analysis*. Significant differences are highlighted with bold.

| | **Whole Face** | | | | **Upper Face** | | | | **Lower Face** | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | N | A | S | H | N | A | S | H | N | A | S | H |
| **N** | -0.08 | **0.216** | -0.12 | -0.02 | 0.02 | 0.046 | 0.05 | -0.12 | 0.048 | 0.01 | -0.03 | -0.02 |
| **A** | -0.14 | **0.297** | -0.07 | -0.08 | -0.05 | 0.057 | 0.08 | -0.08 | **-0.30** | **0.325** | 0.033 | -0.05 |
| **S** | **-0.35** | **0.265** | **0.225** | -0.14 | 0.078 | -0.09 | **0.169** | -0.15 | **-0.27** | 0.091 | **0.242** | -0.05 |
| **H** | **-0.23** | -0.14 | -0.09 | **0.462** | -0.03 | -0.13 | -0.10 | **0.261** | **-0.25** | -0.17 | -0.15 | **0.581** |

**Table 5**: Difference matrices obtained by subtracting emotion confusion matrices of the *point region experiment* from that of *video region experiment*. Significant differences are highlighted with bold.

difference matrices by subtracting the emotion confusion matrices of the video region experiment from that of QDA objective motion analysis. Several significant differences in its whole face section are due to the visual confusions between sad and angry emotions. In the upper face section, QDA objective motion analysis did a much better job than the video region experiment in discriminating sadness and happiness from the upper face motion. Given that there are relatively small amounts of motion in the upper face when humans are making sad or happy speech, experimental subjects often visually misjudged it as neutral. In the lower face section, we were surprised to see a high consistency between QDA objective data analysis and the video region experiment. It suggests that experimental subjects indeed received insightful emotion clues from the lower face region, as such, their emotion confusion matrices were well matched.

Table 5 shows the difference matrices between the point region experiment and video region experiment. Since the abstract point-rendered faces provided less spatial information and less face details than the video faces, we expected the point region experiment to result in a lower recognition rate than the video region experiment. Highlighted differences in the whole face section of Table 5 confirmed our assumption. The upper face section of Table 5, consistent with that of Table 4, shows that subjects had more problems correctly differentiating sadness and happiness from the upper face region alone. Also, except for errors mistaking emotions for neutral,

the lower face section of Table 5 shows that emotion confusion matrices of the lower face region of the point-rendering faces had a similar confusion pattern to video faces. It again confirms that the lower face region (abstract point-rendering representation or real video faces) does provide insightful emotion clues for visual emotion perception (confusion) patterns that are consistent with QDA objective facial motion analysis (Table 4).

## 7 Discussion and Conclusions

In this paper, we investigated how people perceive visual emotions with speech. This work is the first to evaluate the consistency among objective facial motion data analysis, subjective perception of faces in video, and subjective perception of abstract point-rendered faces. We found that the objective discrimination results of motion capture data are highly correlated with those of subjective evaluation experiments. In particular, the lower face region (across multiple representations in rendering) provides insightful clues for visual emotion perception. We also found that anger was easily visually perceived as sadness, and sadness and neutral were difficult to visually discriminate.

In comparison with previous literature [3, 13], the findings of this study are consistent on some points; for instance, our study verifies the importance of the lower face region for most of the emotions [3] and pairs of angry-sad and neutral-sad are easily confused [13]. Our new findings of this work include the fact that the lower face region incorporates important clues for visual emotion perception, because it can generate similar emotion confusion matrices (patterns) among its different face representations. This finding can be further exploited to evaluate expressive embodied conversational agents.

We are aware that the four basic emotions (neutral, anger, sadness, and happiness) studied in this work are not enough to cover comprehensive ECAs, e.g., fear, interest, and disgust were not covered in this work. In future work we plan on examining a wider array of emotions. Furthermore, a major limitation of the current work is that we only studied a single actress when generating the emotions. In order to fully generalize our findings, we plan on replicating this work with a number of actors/actresses. However, the purpose of the current work was to validate a number of methodologies, and given our limited stimulus set, we found excellent convergence in our methods.

There are many questions that remain open. As described in the section on objective motion analysis section, the emotion visual-confusion matrices of the objective analysis are highly consistent with the subjective evaluation experiments. This observation indicates that the objective emotion visual-confusion matrices (especially the lower face region) could be used as a useful benchmark to evaluate synthetic expressive facial animation. For example, for a synthetic expressive speech animation, the same QDA [22] can be applied to the 3D motion dataset (Figure 1) on the animated face. If the new generated objective visual-emotion confusion matrices are close to the ones illustrated in Figure 3 and Table 1, (Measurement could be the Absolute Expressive Performance (AEP) and the Relative Expressive Performance

(REP) [1].), then it could suggest that the synthetic expressive animated face sequence is close to a real human. In sum, the current work provides the beginnings of establishing a system of methodologies to evaluate emotions generated by digital ECAs.

## Acknowledgements

Special thanks go to Joy Nash and Murtaza Bulut for capturing motion data, and Kimy Tran, Stephanie Parker, and Erica Murphey for helping with the subjective evaluation experiments.

## References

1. Ahlberg, J., Pandzic, I.S., and You, L: Evaluating MPEG-4 Facial Animation Players, *MPEG-4 Facial Animation: the standard, implementation and applications*, Pandzic, I.S. and Forchhimer, R. (eds.), 2002, 287-291.
2. Andre, E., Rist, M., and Muller, J: Guiding the User through Dynamically Generated Hypermedia Presentations with a Life-like Character, *IUI '98, pp.* 21-28.
3. Bassili, J.N: Emotion Recognition: The Role of Facial Movement and the Relative Importance of Upper and Lower Areas of the Face, *Journal of the Personality and Social Psychology*, 1979(37), 2049-2058.
4. Blanz, V., Basso, C., Poggio, T., and Vetter, T: Reanimating Faces in Images and Video, *Computer Graphics Forum*, 22(3), 2003
5. Brand, M: Voice Puppetry, *Proc. of ACM SIGGRAPH 1999*, ACM Press, pp. 21-28.
6. Bregler, C., Covell, M., and Slaney, M: Video Rewrite: Driving Visual Speech with Audio, *Proc. of ACM SIGGRAPH'97*, ACM Press, pp.353-360.
7. Busso, C., Deng, Z., Neumann, U., and Narayanan, S: Natural Head Motion Synthesis Driven by Acoustic Prosody Features, *the Journal of Computer Animation and Virtual Worlds*, 16(3-4), July 2005, pp.283-290.
8. Cassell, J., Pelachaud, C., Badler, N., Steedman, M., Achorn, B., Becket, T., Douville, B., Prevost, S., and Stone, M: Animated Conversation: Rule-Based Generation of Facial Expression, Gesture and Spoken intonation for Multiple Conversational Agents*, Proc. of ACM SIGGRAPH'94*, ACM Press, 413-420.
9. Cassell, J., Sullivan, J., Prevost, S., and Churchill, E: Embodied Conversational Agents, MIT Press, 2000.
10. Chuang, E.S., Deshpande, H., and Bregler, C: Facial Expression Space Learning, *Proc. of Pacific Graphics'02*, 2002, 68-76.
11. Cohen, M.M., and Massaro, D.W: Modeling Coarticulation in Synthetic Visual Speech, Magnenat-Thalmann, N., Thalmann, D. (Editors), *Models and Techniques in Computer Animation*, Springer Verlag, 1993, 139-156.
12. Costantini, E., Pianesi, F., and Cosi, P: Evaluation of Synthetic Faces: Human Recognition of Emotional Facial Displays, *Affective Dialogue Systems*, Dybkiaer, L. Minker, W. and Heisterkamp, P. (eds.), 2004.
13. Costantini, E., Pianesi, F., Prete, M: Recognising emotions in human and synthetic faces: the role of the upper and lower parts of the face, *Proc. of IUI'05*, ACM Press (2005), 20-27.

14. Deng, Z., Neumann, U., Lewis, J.P., Kim, T.Y., Bulut, M., and Narayanan, S: Expressive Facial Animation Synthesis by Learning Speech Co-Articulation and Expression Space, *IEEE Transaction on Visualization and Computer Graphics*, 12(6), Nov/Dec, 2006.
15. Deng, Z., Bulut, M., Neumann, U., and Narayanan, S: Automatic Dynamic Expression Synthesis for Speech Animation, *Proc. of IEEE Computer Animation and Social Agents 2004*, July 2004, pp. 267-274.
16. Deng, Z., Lewis, J.P., and Neumann, U: Synthesizing Speech Animation by Learning Compact Speech Co-Articulation Models, *Proc. of Computer Graphics International 2005*, June 2005, pp. 19-25.
17. Ekman, P., Friesen, W.V: Unmasking the Face: A Guide to Recognizing Emotions from Facial Clues, Prentice-Hall, 1975
18. Essa, I. A. and Pentland, A.P: Coding, Analysis, Interpretation, and Recognition of Facial Expressions, *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 19(7), 1997, 757-763
19. Ezzat, T., Geiger, G., and Poggio, T: Trainable Videorealistic Speech Animation, *ACM Trans. Graph.*, 21(3), 2002, 388-398.
20. Gratch, J. and Marsella, S: Evaluating a Computational Model of Emotion, *Journal of Autonomous Agents and Multiagent Systems*, 11(1), pp. 23-43.
21. Gratch, J., Rickel, J., Andre, E., Badler, N., Cassell, J., and Petajan, E: Creating Interactive Virtual Humans: Some Assembly Required, *IEEE Intelligent Systems*, July/August 2002, pp. 54-63.
22. Hastie, T., Ribshirani, R., and Friedman, J: The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Springer-Verlag, 2001.
23. Katsyri, J., Klucharev, V., Frydrych, M., Sams, M: Identification of Synthetic and Natural Emotional Facial Expressions, *Proc. of AVSP'2003*, 2003, 239-244.
24. Kshirsagar, S., and Thalmann, N.M: Visyllable Based Speech Animation, *Computer Graphics Forum*, 22(3), 2003.
25. Walker, J. H., Sproull, L., and Subramani, R: Using a human face in an interface, *Proc. of CHI'94*, ACM Press (1994), 85-91.
26. Lee, Y., Terzopoulos, D., and Waters, K: Realistic modeling for facial animation, *Proc. of ACM SIGGRAPH'95*, ACM Press (1995), 55-62.
27. Lewis, J.P: Automated lip-sync: Background and techniques, *J. of Visualization and Computer Animation*, 1991, 118-122.
28. Lewis, J.P., Purcell, P.: Soft Machine: A Personable Interface, *Proc. of Graphics Interface 84*, pp. 223-226.
29. Marsella, S. and Gratch, J: Modeling the Interplay of Plans and Emotions in Multi-Agent Simulations, *Proc. of the Cognitive Science Society*, 2001.
30. Nass, C., Kim, E.Y., and Lee, E.J: When My Face is the Interface: An Experimental Comparison of Interacting with One's Own Face or Someone Else's Face, *Proc. of CHI'98*, ACM Press (1998), 148-154.
31. Noh, J.Y., and Neumann, U: Expression Cloning, *Proc. of ACM SIGGRAPH'01*, ACM Press (2001), 277-288.
32. Pandzic, I.S., Ostermann, J., and Millen, D: User evaluation: synthetic talking faces for interactive services, *The Visual Computer*, 1999(15), 330-340.
33. Parke, F: Computer Generated Animation of Faces, *Proc. ACM Nat'l Conf.* ACM Press (1972), 451-457.
34. Pelachaud, C., Badler, N., and Steedman, M: Linguistic Issues in Facial Animation, *Proc. of Computer Animation'91*, 1991.
35. Pelachaud, C., Badler, N. and Steedman, M: Generating Facial Expressions for Speech, *Cognitive Science*, 20(1), 1994, 1-46.

36. Rist, M., Andre, E., and Muller, J: Adding animated presentation agents to the interface. *IUI '97: Proc. of Intelligent user interfaces*, ACM Press (1997), 79-86.
37. Sirovich, L. and Kirby, M: Low-dimensional procedure for the characterization of human faces, *J. Opt. Soc. Am. A*, 4(3), 1987 March, 519-524
38. Turk, M. A. and Pentland, A.P: Face Recognition Using Eigenfaces, *IEEE CVPR 91*, 586-591.
39. Uttkay, Z., Doorman, C., and Noot, H: Evaluating ECAs - What and How? *Proc. of the AAMAS02 Workshop on Embodied Conversational Agents*, 2002.