Evaluation of Real-World and Computer-Generated Stylized Facial Expressions

Article *in* ACM Transactions on Applied Perception · November 2007

DOI: 10.1145/1278387.1278390 · Source: OAI

Evaluation of Real-World and Computer-Generated Stylized Facial Expressions

CHRISTIAN WALLRAVEN and HEINRICH H. BÜLTHOFF
Max Planck Institute for Biological Cybernetics
DOUGLAS W. CUNNINGHAM
Max Planck Institute for Biological Cybernetics and WSI-GRIS
JAN FISCHER
WSI-GRIS
and
DIRK BARTZ
Visual Computing (ICCAS)

The goal of stylization is to provide an abstracted representation of an image that highlights specific types of visual information. Recent advances in computer graphics techniques have made it possible to render many varieties of stylized imagery efficiently making stylization into a useful technique, not only for artistic, but also for visualization applications. In this paper, we report results from two sets of experiments that aim at characterizing the perceptual impact and effectiveness of three different stylization techniques in the context of dynamic facial expressions. In the first set of experiments, animated facial expressions are stylized using three common techniques (brush, cartoon, and illustrative stylization) and investigated using different experimental measures. Going beyond the usual questionnaire approach, these experiments compare the techniques according to several criteria ranging from subjective preference to task-dependent measures (such as recognizability, intensity) allowing us to compare behavioral and introspective approaches. The second set of experiments use the same stylization techniques on real-world video sequences in order to compare the effect of stylization on natural and artificial stimuli. Our results shed light on how stylization of image contents affects the perception and subjective evaluation of both real and computer-generated facial expressions.

Categories and Subject Descriptors: I.3.7 [**Computer Graphics**]: Three-Dimensional Graphics and Realism—*Animation*; J.4 [**Computer Application**]: Social and Behavioural Sciences—*Psychology*

General Terms: Experimentation

Additional Key Words and Phrases: Evaluation of facial animations, avatar, facial expressions, psychophysics, perceptually adaptive graphics, stylization

ACM Reference Format:

Wallraven, C., Bülthoff, H. H., Cunningham, D. W., Fischer, J., and Bartz, D. 2007. Evaluation of real-world and computer-generated stylized facial expressions. ACM Trans. Appl. Percpt. 4, 3, Article 16 (November 2007), 24 pages. DOI = 10.1145/1278387.1278390 http://doi.acm.org/10.1145/1278387.1278390

Authors' addresses: Christian Wallraven and Heinrich H. Bülthoff, Max Planck Institute for Biological Cybernetics, Tübingen, Germany; Douglas W. Cunningham, Max Planck Institute for Biological Cybernetics and WSI-GRIS, University of Tübingen, Tübingen, Germany; Jan Fischer, WSI-GRIS, University of Tübingen, Tübingen, Germany; Dirk Bartz, Visual Computing (ICCAS), University of Leipzig, Leipzig, Germany.

1. INTRODUCTION

Stylized or nonphoto-realistic (NPR) rendering has attracted much interest in the computer graphics community over the last decade and has established itself firmly alongside the quest for increasing realism. Techniques that allow automatic creation of images that convey complex meaning and support high degrees of abstraction "with a few brush strokes" have applications ranging from illustration to information visualization to artistic expression.

One of the major challenges in designing stylization algorithms lies in identifying *principled ways* for creating such images. These principled ways depend, of course, crucially on the goal at hand: creating an image so that it efficiently conveys specific information or so that it conforms to particular aesthetic principles, these are two vastly different goals and require two different solutions. Even when focusing on a particular task, it is often unknown which visual information is needed to support this task: If faced with the task of rendering a facial expression such that it is easily recognizable, no one can clearly describe exactly what information is necessary or sufficient in order to perceive a thoughtful smile.[1] Finally, in some cases, it is also difficult to evaluate or measure the success of a particular technique. Although questionnaires and other introspective measures are commonly used and offer quick and easy answers, they provide only rather indirect insights. For example, one might simply ask observers "Is this an effective technique for rendering facial expressions?" and get valid data about what the observers *think* about the effectiveness of a technique. Reflections about the potential effectiveness of a technique, however, is not necessarily the same thing as actually measuring its effectiveness. For that, one needs a direct measure.

In this paper, we conduct a detailed evaluation of stylized rendering. More specifically, we examine the degree with which three stylization techniques, which greatly differ in terms of their visual impression, are able to effectively render facial expressions. We also evaluate the stylization techniques on both animated facial expressions and real-world video footage of facial expressions. A comparison of both types of stimuli will yield insights into the domain specificity of the techniques; we might assume, for example, that computer animated faces because of their smoothness are affected differently by stylization than real-world sequences, which show much more inherent variability on the pixel level. Another central goal of this paper is to compare and contrast several experimental measures ranging from introspective ratings to task-specific performance characteristics. This allows us not only to contrast the evaluation criteria, but also to paint a more complete picture of the impact of the different stylization techniques in the context of facial expressions.

2. RELATED WORK AND MOTIVATION

In this section, we briefly discuss related work in stylized rendering techniques, evaluation of those techniques, as well as psychophysical research on perception of facial expressions. In addition, we also state how this paper aims to advance on issues raised in each context.

2.1 Stylization Techniques

Artistic and illustrative stylization have been areas of very active research for several years. Strothotte and Schlechtweg [2002] have published a good survey of methods used in the field. While artistic and illustrative-stylization techniques usually remove some detail from the original image, in some applications they can actually convey the relevant information better than unprocessed data. This principle was dubbed "functional realism" in Ferwerda [2003]. The following brief discussion lists three selected classes of stylized rendering algorithms that have inspired the work in this study.

[1] This study focuses on stylized renderings of complex, natural images rather than on the highly abstracted semiotics of iconography.

In DeCarlo and Santella [2002], a technique for *cartoonlike stylization* of photographs is presented that uses a combination of color segmentation and edge detection. The areas in which the stylization is applied are selected by frequency of fixations determined through eye tracking of users looking at an image. The resulting images consist of uniformly colored image regions on top of which edges are painted to emphasize contours. A second class of algorithms creates *brush-stroke* stylization of images and videos [Haeberli 1990; Litwinowicz 1997]. These are often used in painterly rendering as they give the impression that the images were painted using a paintbrush. A third class of algorithms is based on *half toning*, where the goal is to transform a grayscale or color continuum into black-and-white hatching [Freudenberg et al. 2002]. Images created using half toning resemble sketches done with a pen by capturing the underlying shading in the image using cross-hatching.

2.1.1 *Aim.* The aim of this paper is not to develop additional stylization techniques, but to evaluate the effectiveness of three existing algorithms—one from each class. Such a perceptual evaluation will provide valuable information for the areas of computer animation and NPR on how to create easily recognizable, stylized animations.

2.2 Evaluation Approaches

Evaluations of stylized or NPR techniques have been conducted using three main approaches, which will be discussed in the following.

In the first approach, which is based on introspection, users (or experts) are asked for their impression of some aspect (e.g., effectiveness [Agrawala and Stolte 2001]) of the abstracted imagery. This approach is easy to apply (often using questionnaires) and analyze, which might explain why it is, perhaps, the most common evaluation method not only for stylized techniques, but also for computer graphics in general (see, e.g., Stokes et al. [2004]). Its main drawback is that it has limited validity and generalizability as the desired information may not be readily and reliably accessible by introspection.

The second approach—most often used in human factors studies—evaluates the performance of users in a task-dependent context (for examples, see Fischer et al. [2006a], Gooch and Willemsen [2002], Gooch et al. [2004], Wallraven et al. [2005]). Gooch et al. [2004], for example, evaluated the impact of a line-drawing stylization method on identification and learning of faces. They found that the stylization faces were just as easily identified as photographs. Critically, they also found that users learned novel faces *faster* when they were stylized than when they were real photographs. This shows that abstracting the correct information not only results in a more efficient data representation, but also in more effective processing. In a more recent paper, Winnemöller et al. [2006] conducted two user studies that aimed at evaluating the proposed stylization technique[2] using naming and a memory game tasks. In the first task, which followed the same protocol as Gooch et al. [2004], participants were significantly faster at naming stylized images than nonstylized images. Accuracy in both tasks, however, was at ceiling so that a potential impact of stylization on this measure could not be analyzed. In the second task, participants played a memory game using either stylized or nonstylized faces—again, participants were quicker to finish the memory game with stylized faces. Taken together, these studies suggest that stylization provides crucial benefits for static stimuli in tasks with a heavy memory load—in this paper, we would like to go beyond the static dimension and concentrate on the impact of stylization on dynamic stimuli. In another application of task-dependent evaluation, Fischer et al. [2006a] investigated the use of stylization for creating a consistent augmented reality environment. Currently, the placement of virtual objects in real video results in, sometimes obviously, noticeable visual artifacts. Psychophysical experiments showed, however, that if both the virtual objects and the real scene were stylized,

[2] This technique is similar to the “cartoon” stylization used in this paper.

participants failed to distinguish between real and virtual objects, thus demonstrating the usefulness of abstraction. The major disadvantage of this second evaluation approach is that the large number of potential tasks makes it near impossible to measure performance on every level. For techniques that are designed with a specific task in mind, however, such a direct, task-specific evaluation approach is, of course, to be preferred.

In a recent paper, Santella and DeCarlo 2004 presented an interesting third approach to evaluation of stylized images. This approach is based on eye movements that are known to reflect not only overt, but also covert, cognitive processes. In their study, statistical analyses of fixation clusters were conducted to show how different NPR techniques guide and capture the users' gaze. Although the results seem promising, it is unclear exactly what the method is measuring and how it compares to the other approaches. Furthermore, data acquisition (which requires eye-tracking equipment), analysis, and interpretation are difficult and less than straightforward.

2.2.1 *Aim.* In this study, we will take an *integrative* approach to evaluating stylized imagery by collecting both introspective *and* task-specific data in order to paint a more complete picture. More specifically, we will investigate effectiveness of stylized, dynamic facial expressions through evaluating a battery of measures: these include introspective questionnaires, direct comparisons, recognition performance, perceived intensity, and perceived sincerity.

2.3 Perception of Facial Expressions

Facial expressions have been extensively studied in the cognitive sciences over the last few decades (for a recent review, see Adolphs [2002]). Studies by Ekman [1972], for example, suggest that there are seven universally recognized facial expressions (anger, contempt, disgust, fear, happiness, sadness, and surprise). In addition, facial expressions have been shown to provide a rich nonverbal communication channel that is able to alter the meaning of what is being said, to provide emphasis to spoken words, or to control the flow of a conversation (see Bull [2001]). Recently, a series of papers ([Cunningham et al. 2003, 2004, 2005], Wallraven et al. [2004]) has started to characterize the visual information that drives the recognizability, intensity, and believability of conversational facial expressions. In Wallraven et al. [2005] this research was used in an initial set of experiments to evaluate the perceptual realism of several 3D animation methods. Since these animation methods allow full control over important information channels used in facial expressions (such as internal motion of the face, rigid head motion, shape, and texture), they provide an ideal tool for highly controlled psychophysical experiments. After determining how perceptually realistic the animations are, one of the animation methods was used to investigate the relative contribution of shape, texture, and motion to the recognizability and perceived sincerity of conversational expressions. In all experiments, a strong influence of dynamic information (both rigid head motion and nonrigid facial motion) was found. Whereas shape and texture manipulations showed only little influence on recognizability, intensity and sincerity were more affected by these dimensions. Overall, the results determined the differential contribution of a variety of perceptual characteristics and animation methods to the perception of facial expressions and demonstrated the benefits of a close coupling of psychophysical and computer animation methods.

2.3.1 *Aim.* Analyses of the visual information that is emphasized by different stylization techniques will not only help to evaluate the effectiveness of the stylization techniques, but will shed further light on the processing of facial expressions.

3. STYLIZED FACIAL EXPRESSIONS

In the following, we first briefly review the original sequences, that is, the facial animation system and the video corpus of facial expressions that were used as source material for the stylization experiments.

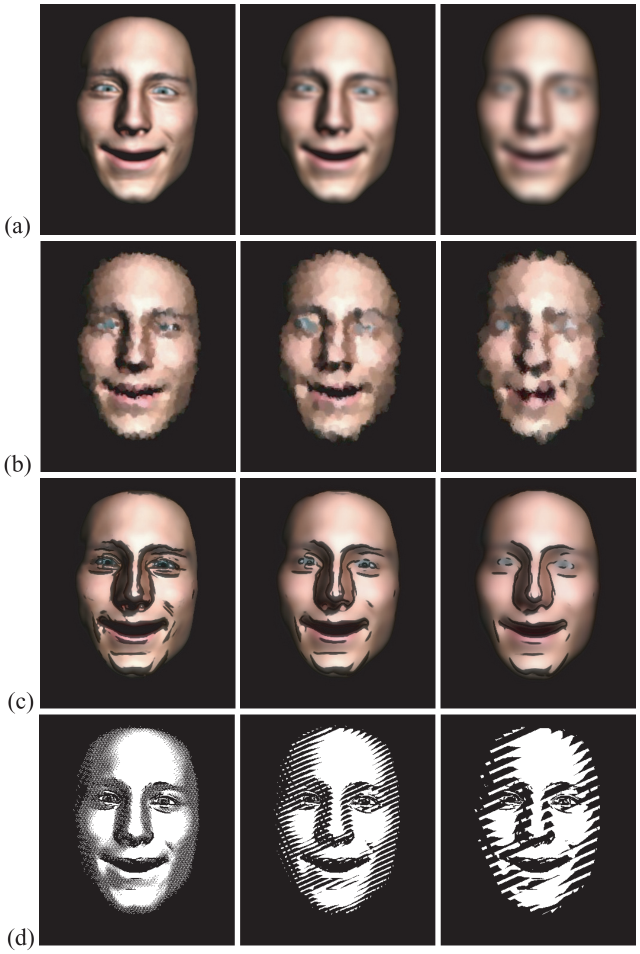


Fig. 1. Stylization techniques used in this paper for an animated happy expression and all resolution levels. (a) Standard avatar, (b) brush-stroke, (c) cartoon, and (d) illustrative stylization.

We then discuss the three different stylization techniques that were applied to these animated expressions in order to create stylized sequences. These specific techniques were selected as each enhances or decreases the importance of quite different image characteristics (such as color, edges, or motion continuity) to a different degree. In addition, for each technique, we also determined a suitable parameter that allowed us to manipulate the level of detail contained in the image. This was done in order to investigate the impact of increasing abstraction on the effectiveness of each technique.

3.1 Original Video Sequences

3.1.1 *The Avatar.* The avatar that was used in this paper is based on the design by Breidt et al. [2003] and was introduced in Wallraven et al. [2005] (see Figure 1a). It is based on a combination of high *spatial* resolution 3D scans of peak expressions, together with high *temporal* resolution motion capture of the facial deformation during these expressions. Both scan and motion capture data for these expressions are taken from a trained actor using a method-acting protocol that ensures very "natural" expressions. Scans of peak expressions are first put into correspondence using a manually designed control mesh in order to create a basis set of *morphable meshes*. From the motion capture data (captured with 72 markers), nonrigid motion is extracted and used to specify linear detectors for the expression-specific deformations in the face. The detectors provide the weights that drive the basis set of morph channels. Finally, eyes and teeth geometry are added to the scans and anchored to the rigid head motion. The movements of the eyes are created by fixating them on the virtual camera throughout the expression sequence. This corresponds closely to the real eye movements made by the actor during the recordings. The whole pipeline results in a realistic animation based on the amplitude and timing of marker motion in the motion capture data. The expression sequences used in this study were the same as in Wallraven et al. [2005] (except for the addition of eyes and teeth).

3.1.2 *Real-World Video Sequences.* The recordings of the expression sequences were done with the VideoLab of the Max Planck Institute [Kleiner et al. 2004]. This system is a custom-designed recording setup with six digital cameras arranged in a semicircle around the actor/actress at a distance of about 1.5 m. These cameras are fully synchronized and each of them is connected to a dedicated computer onto which the video data is streamed.

For this experiment, we took video sequences of facial expressions from the *same actor* whose data was used for creating the 3D scans and motion capture recordings for the avatar. Using the same method-acting protocol, nine facial expressions were recorded (see Experiment 3). All sequences were recorded at PAL video resolution of 768 × 576 pixels and at a temporal resolution of 25 frames/s. In order to make the sequences comparable to our avatar animation, which did not use a torso, we used an image-based stereo motion-tracking algorithm to clip and postprocess the video sequences (for a detailed description of this manipulation technique, see Cunningham et al. [2005] and Kleiner et al. [2004]; example images are shown in Figure 2a). Finally, we equalized the average image intensities across animated and real-world sequences and resized all stimuli such that the visual angle subtended by the face in each image was equal for both stimulus classes.

3.2 Stylization Techniques

3.2.1 *Brush-Stroke Stylization.* The brush-stroke stylization used in this study (Figures 1b and 2b) is a painterly style where the output images are composed of a number of small brush strokes. The algorithm used for achieving this effect was presented in Fischer et al. [2005a]. Briefly, a 2D sampling grid is generated in a one-time preprocessing step. The grid remains fixed throughout the processing of consecutive input images and consists of an array of sampling point records. Each sampling point record contains the 2D position of the point and additional information about the brush stroke that is to be painted there. The point position is based on a regular grid with a horizontal and vertical grid spacing. Each brush-stroke location is randomly displaced from this initial regular grid position. In addition, the radius of the Brush stroke is randomly generated, with a user-definable random number range. Finally, a random color offset is computed for each brush stroke. The image stylization process samples the input image by reading pixel colors at the sampling point positions in a random order (this random order is determined in a preprocessing step and remains constant for all images). The color offset is then added to each pixel color and the resulting RGB components are clamped to the valid color number range. Each brush stroke is drawn as a textured square with a side length of the stored stroke radius centered at the brush-stroke location—this radius introduces a natural resolution scale. During brush-stroke rendering, alpha blending is enabled to achieve partial transparency for overlapping brush strokes.

3.2.1.1 *Characteristics.* The brush-stroke stylization preserves local colors in the image, with only a limited random color offset added to the input pixel. Depending on the selected brush stroke radius, however, small or medium-sized regions are masked in the output image. The discrete sampling of

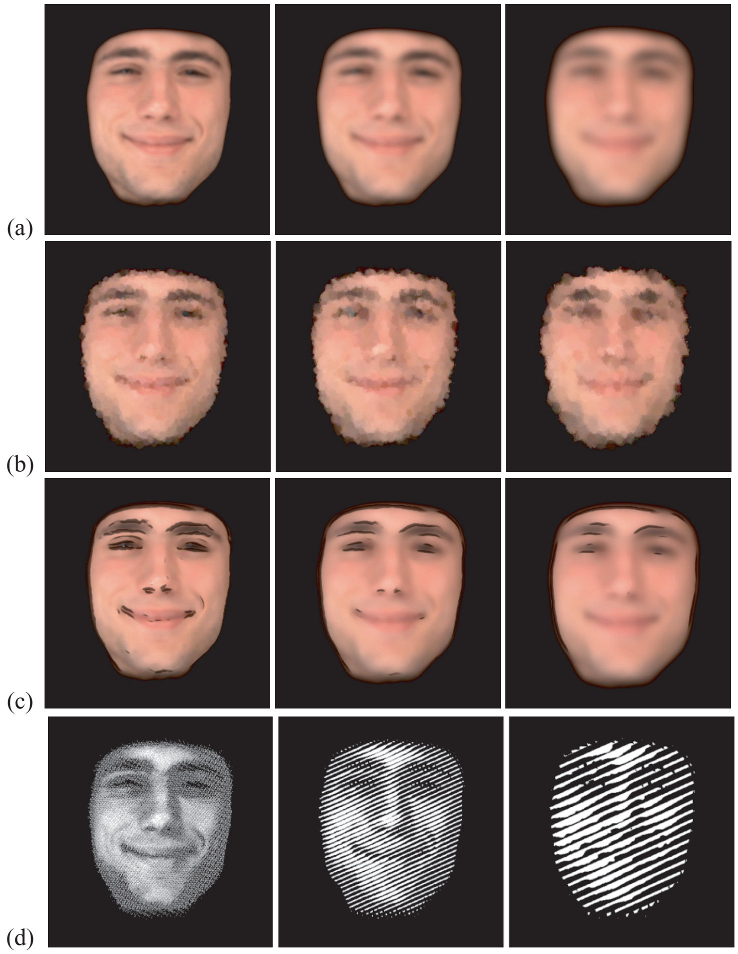


Fig. 2. Stylization techniques used in this paper for a happy expression taken from real-world video footage and all resolution levels. (a) Standard avatar, (b) brush-stroke, (c) cartoon, and (d) illustrative stylization.

input pixels and the typically rather large sampling point distance result in limited frame coherence or motion continuity for animated image sequences.

3.2.2 *Cartoonlike Stylization.* In the cartoonlike stylization technique (see Figures 1c and 2c) used here, each input image is processed so that the resulting image consists of mostly uniformly colored areas enclosed by black silhouette lines. The algorithm, which was described in Fischer et al. [2006b], is designed as a postprocessing filter in a real-time rendering pipeline. The implementation of the algorithm uses vertex and fragment shaders, which run on the programmable graphics processing units (GPUs) of recent graphics cards.

The stylization filter consists of two steps. In the first step, a simplified color image is computed from the input image. The basis of this computation is a nonlinear filter, which is inspired by bilateral filtering, as described in Tomasi and Manduchi [1998]. The nonlinear filter performs a repeated, photometric weighting of the pixels, taking into account only their chrominance components. The repetition of the filter operation is necessary in order to achieve a sufficiently good color simplification. The second stage of the image-stylization filter is an edge detection step, based on the simplified color image. Thus, the silhouette lines are located between similarly colored regions in the image, which is an approximation of a cartoonlike rendering style. Finally, the simplified color image is combined with the edge-detection results. The responses of the edge-detection filter are drawn over the output image as black lines. A specific weight function is used for computing a transparency for the detected edge pixels, which produces a smooth blending over the color image. As an increasing number of filtering iterations in the first step results in highly simplified, blurred image as well as less distinct edges, this parameter was chosen to create different resolution levels for this technique.

3.2.2.1 *Characteristics.* The cartoonlike stylization stresses high contrast edges in the image and preserves the dominant color in larger image regions. It does, however, remove small details, as well as low-contrast details, as an effect of the nonlinear filter.

3.2.3 *Illustrative Stylization.* The illustrative stylization used here (see Figures 1d and 2d) generates output images, which reproduce the brightness of the input image with black-and-white hatching. Moreover, high-contrast edges are rendered as black lines. This algorithm is based on aspects of the illustrative-rendering method described in Fischer et al. [2005b].

In order to creat the hatching, a procedural half-toning technique similar to the one described by Strothotte and Schlechtweg [2002] is applied.

Parameters of this algorithm are the orientation of the main hatching direction, the minimum intensity required for the addition of perpendicular cross-hatching strokes, as well as the size of the hatching pattern in pixels. As with the brush-stroke technique, the size of the pattern was used to determine the resolution levels. In addition to the black-and-white representation of the input image, silhouette lines are added to the stylized output. A Sobel edge-detection filter delivers the location of high-contrast edges in the image. These locations are then overlaid as black lines over the output image. As can be seen in Figure 1d, these lines contribute to the final image mainly in high-contrast regions, such as the eyes.

3.2.3.1 *Characteristics.* The illustrative stylization emphasizes intensities in the image. These are computed as the Y component of the YUV color-space representation of each input pixel. Moreover, high-contrast edges are stressed by the edge-detection step used in this stylization method. The illustrative stylization removes all color information from the image, rendering a purely black-and-white representation of the input image. Small details in the image are not preserved, depending on the selected size of the half-toning pattern.

3.3 Stylization for Animated Versus Real-World Sequences

Even though we tried to equalize the low-level image properties of the images, stylization of animated sequences (Figure 1) and of real-world sequences (Figure 2) does produce different visual impressions. Overall, it seems that the avatar has a much more homogeneous “look” to it. One of the reasons for this lies in the smooth face texture used for the avatar, which results in sharper contrasts in the face compared to the video sequences—this can be seen most clearly when comparing the first resolution level in Figures 1d and 2d, which show the illustrative stylization. Another reason for this homogeneous look lies in the lighting and material properties used for rendering the animated sequences: the 3D geometry of the avatar was lit by an omnidirectional light coming from slightly above and a standard Blinn shader was used to simulate diffuse and specular properties of the skin. The video sequences, on the other hand, were recorded under much more flat lighting conditions resulting in less contrast differences across the face. Another difference between the two stimulus classes lies in the temporal domain: the avatar—even though designed to capture all the relevant facial deformations—is still limited by its underlying morph shapes. A “real” face might make many more micro-motions between a neutral face and a peak expressive face that are not captured by a linear morph between the endpoint. This difference in high-frequency motion content results in a richer, more complex motion pattern for real-world video sequences and, therefore, also results in different stylization effects on motion compared to the animated sequences.

Even though these differences make a direct comparison between animated and real-world sequences more difficult, here we were interested in the relative pattern of effects: given such “standard” stimulus material (“clean,” computer-animated sequences, on the one hand, compared to “noisier,” real-world video sequences, on the other) what would be the effects of stylization on each stimulus class and would those effects be dependent on the chosen evaluation measure?

4. EXPERIMENTS

This section describes the experiments that were conducted to investigate the perceptual impact of the different stylization techniques. Experiment 1 + 2 were done with the animated sequences and compared different evaluation measures. Experiment 3 was a replication of Experiment 2 with the real-world sequences allowing us to examine domain-specific effects of the different stylization techniques

- **Direct preference (Experiment 1):** By showing two stylized sequences side-by-side and asking “which sequence captures the essence of the expression better?” we allow participants to compare and contrast two stylization techniques at the same time. One of the advantages of this method is that although the question asks for a very subjective evaluation, participants are forced to chose one sequence, which allows for a clean analysis of the data.
- **Recognizability (Experiments 2 + 3):** It is, of course, crucial that the different techniques support the recognition of the facial expressions. In particular, stylization should neither decrease recognition accuracy nor increase recognition time compared to the non stylized version. In contrast to the two previous criteria, recognition accuracy constitutes an objective, quantifiable criterion.
- **Intensity and sincerity (Experiments 2 + 3):** These two criteria constitute higher-level characteristics of facial expressions. The ratings can, for example, be of interest if the goal is not only to create recognizable, but also convincing, facial expressions. This will be important in areas such as virtual sales and kiosk applications.
- **Introspective questionnaire (Experiments 1 + 2 + 3):** The questionnaire asks participants to rank the three stylization techniques. The techniques are ranked three times: Once according to aesthetic principles, once according to effectiveness for rendering facial expressions, and once according to subjective preference. All three criteria are evaluated by introspection.

4.1 Experiment 1—Direct Comparison

In the first experiment, participants directly compared two sequences in terms of their effectiveness. In addition, participants had to fill out a standard introspective questionnaire.

4.1.1 *Stimulus Sequences.* In this experiment, we used seven expression sequences from the avatar: confusion, disgust, fear, happy, sad, surprise, and thinking. Each sequence was contrast normalized in order to provide a consistent input to each of the three stylization techniques. We chose default parameters for rendering each technique that were derived from their standard use [Fischer et al. 2006b, 2005a, 2005b]. Three different resolution levels were obtained by (a) setting the diameter of the element size for the brush-stroke algorithm to 2, 8, and 16 pixels, (b) treating the texture map in the cartoon algorithm by blurring the image 2, 8, and 16 times, and (c) setting the basic element size for the illustrative stylization to 2, 8, and 16 pixels. Finally, we created three different resolution levels for the avatar by blurring it with an equivalent blurring filter as for the cartoon stylization (see Figure 1a–d for example images).

4.1.2 *Design.* The two video sequences were presented side-by-side at a resolution of 1024 × 768 pixels on a CRT monitor (each sequence was shown at 512 × 512 pixels). Participants viewed these sequences using a chin rest at a distance of 0.5 m (each face on the monitor subtended a visual angle of 11.4°). A single trial of the experiment consisted of the video sequences being shown repeatedly in the center of the screen. A 200-ms blank screen was inserted between repetitions of the video clip. In each trial, participants had to indicate by a timed button press whether they thought the left or the right sequence captured the essence of the expression better (no name or description of expression itself was given). The experiment compared all methods and resolution levels within each expression, leaving out same–same comparisons—the total number of trials was thus (7 expressions) [(12 combinations of method·resolution level) (12) - 12 same–same comparisons)/2] = 462 trials.

After the experiment, we showed participants high-quality printouts of the different techniques and resolution levels that were taken from the peak frame of the happy expression. We then asked them to fill out a questionnaire in which they had to rank these 12 images according to three different criteria. The first criterion asked how artistic participants thought the different stylization techniques to be. The second criterion asked which of the techniques was the most effective in rendering facial expressions—the same question as in the direct comparison task. Finally, we asked participants to rank the techniques according to which one they liked best.

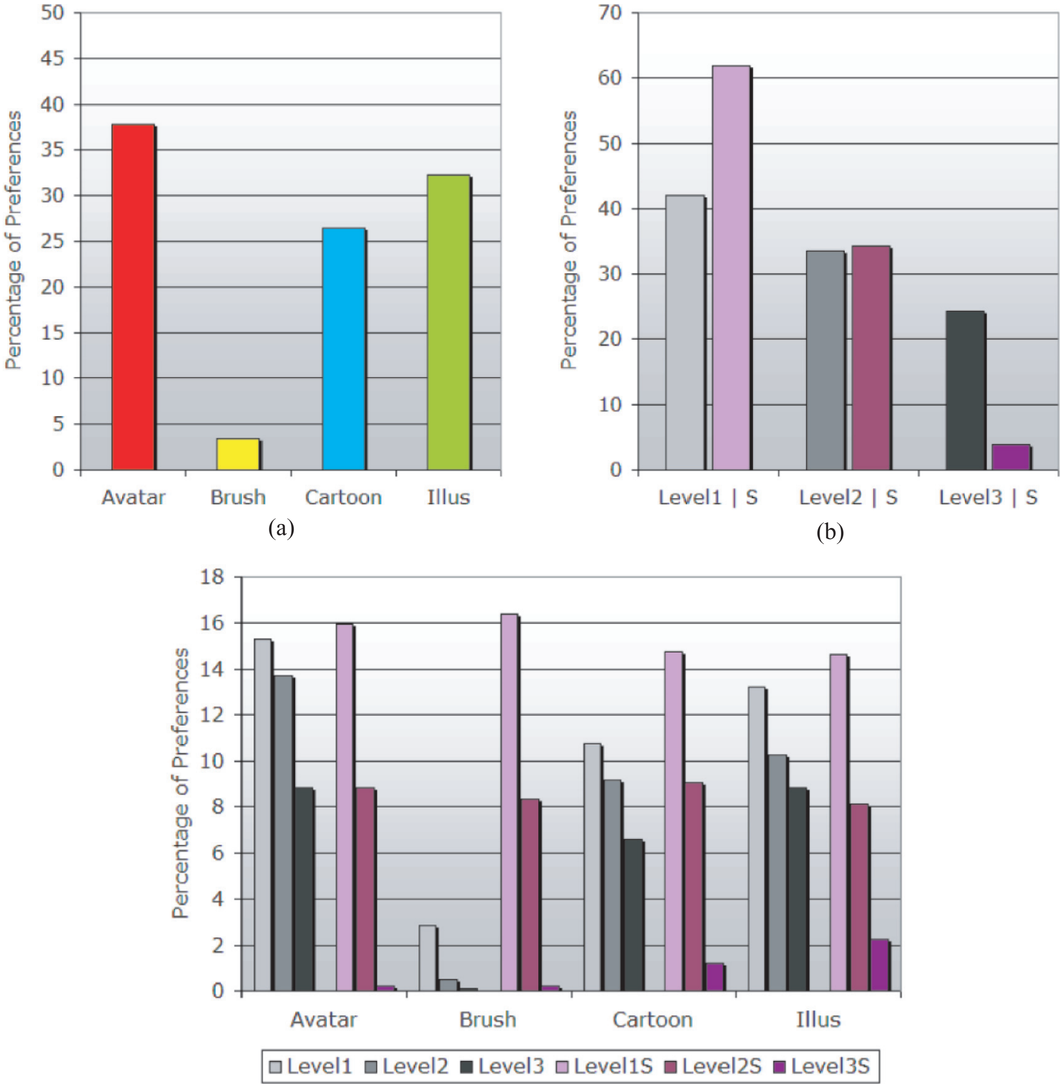
4.1.3 *Results and Discussion.* The direct preference data from ten participants were analyzed as frequency histograms using χ^2 tests with the factors “stylization technique”, “resolution level,” and “expression” for between technique comparisons. The analysis of the trials in which both image sequences used the same stylization technique was done separately in order to look for effects of “resolution level” and “expression” within each technique.

4.1.3.1 *Preference for between Technique Comparisons.* We found highly significant main effects for technique ($p < 0.001, df = 3, \chi^2 = 1036.43$), resolution level ($p < 0.001, df = 2, \chi^2 = 178.76$) as well as an interaction between level and method ($p < 0.001, df = 6, \chi^2 = 101.84$). These effects are plotted in Figures 3.

As can be seen in Figure 3a, when faced with two different stylization techniques, participants most often choose the original avatar animation, followed by the illustrative and cartoon techniques. The brush-stroke method got chosen a mere 4% of the time. Figure 3b shows that there is a clear trend in preferences as a function of resolution level—the most detailed level is preferred over the medium level, which, in turn, is preferred over the least detailed level with a 10% drop in preference for each step. Figure 3c shows that this effect depends on the stylization technique; for the avatar and cartoon conditions, there is virtually no difference between the first and the second level, but a large drop for the third. For both illustrative and brush-stroke techniques, the second and third levels are chosen equally often; both levels are chosen significantly less than the first, most detailed level.

Participants clearly thought that the avatar captured the essence of the expressions best. Of the three stylization techniques, the illustrative style seems to be preferred and the brush-stroke seems to be not seen as not very effective. For all sequences, the highest level of detail is preferred—a result that might be expected as the least detailed levels contain only severely reduced visual information that masks much of the facial motion.

4.1.3.2 *Preference for within Technique Comparisons.* The analysis revealed a highly significant main effect for resolution level ($p < 0.001, df = 2, \chi^2 = 422.12$) as well as an interaction between level



(a)

(b)

(c)

Fig. 3. Experiment 1. Preference results for (a) stylization techniques, (b) resolution level, and (c) both techniques and levels. In (b) and (c), the grey-shaded bars represent the between-technique trials, while the colored bars represent the within-technique trials (labeled with a suffix “S”).

and method ($p < 0.001, df = 6, \chi^2 = 25.67$). These effects are plotted in Figure 3b and c using colored bars.

Figure 3b shows that for within technique comparisons, the first level is preferred 60% of the time, whereas the second level drops sharply to 35% and the third, least detailed, level is rarely chosen (t test: all $p < 0.01$). Again, this pattern depends on the stylization technique—for both illustrative and cartoon stylization the least detailed level was occasionally chosen, whereas for both avatar and brush-stroke stylization it was almost never chosen. Interestingly, the preference of the highest level of detail is much more pronounced for within than for between technique comparisons.

4.1.3.3 *Response Times.* Participants tended to respond more slowly in trials where at least one of the two image sequences was rendered with brush-stroke stylization than in other trials (2.5 versus 2.1s, $p < 0.05$). There were no other statistically significant effects for reaction time. Overall, however, the general lack of a reaction-time effect, as well as the absolute numbers jointly suggest that the decision was not dependent on the stylization technique and that it was not particularly hard for participants to reach a decision.

4.1.4 *Questionnaires.* The questionnaires of Experiments 1 + 2 yielded very similar results, which will be discussed in Section 4.2.3.

4.1.5 *Summary.* When directly comparing two image sequences, participants clearly felt that the original avatar animation captured the essence of the expressions best. Among the stylization techniques, the illustrative method was preferred. In addition, the most detailed resolution was judged as more effective than lower resolution levels, with the exact pattern of decrease depending on the stylization technique used. This indicates that some techniques are less prone to degradation than others—a result that is confirmed by the pattern of preferences found for within-technique comparisons. In summary, the results from this experiment seems to indicate that stylization would actually *harm* the effective depiction of facial expressions.

4.2 Experiment 2—Recognizability

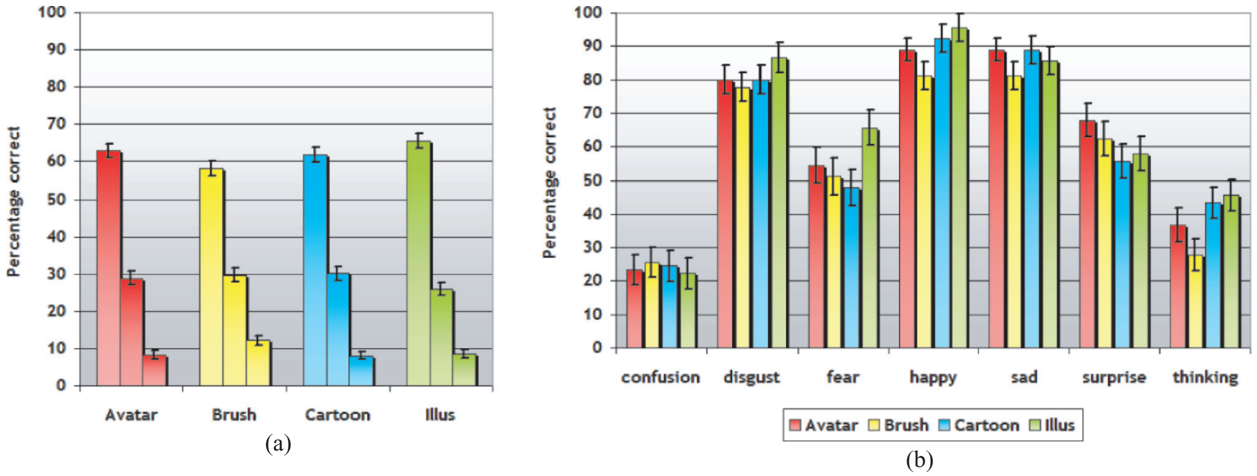
The second experiment provides a different and complementary view to the same issue examined in the first experiment. Here, in the context of a specific task, several perceptual measures (including recognition, reaction time, and the perception of intensity and sincerity) are investigated.

4.2.1 *Design.* The setup and design of this experiment followed closely that of Wallraven et al. [2005]. The first task was to *identify* the expression by selecting the name of the expression from a list displayed on the side of the screen. The list of choices included all seven expressions as well as "none-of-the-above" (an eight-alternative nonforced-choice task (see Cunningham et al. [2003] for a detailed discussion of this paradigm). The second task was to rate the *intensity* of the expressions on a scale from 1 (not intense) to 7 (very intense). In the third task, participants were to rate the *sincerity* of the expressions, with a rating of 1 indicating that the actor was clearly pretending and a value of 7 indicating that the actor really meant the underlying emotion. Participants were explicitly instructed to anchor the scales at a value of 4 (normal intensity and sincerity) and to try and use the whole range of the scale during the experiment. The experiment used three repetitions of each sequence, yielding a total of (7 expressions) (4 stylization techniques) (3 resolution levels) (3 repetitions) = 252 trials. After the experiment, we asked participants to fill out the same questionnaire as in Experiment 1.

4.2.2 *Results and Discussion.* Data were collected from ten participants who had not taken part in the previous experiment. The results were analyzed using standard analysis of variance (ANOVA) methods, which analyze statistical significances for each factor (expression, stylization technique, resolution level) for the different measures (recognition, reaction time, intensity, and sincerity).

4.2.2.1 *Recognition.* The ANOVA found main effects of expression ($F(6, 54) = 12.201$, $p < 0.001$) stylization method ($F(3, 27) = 3.27$, $p < 0.001$) as well as interactions between expression and method ($F(18, 162) = 1.555$, $p = 0.05$) and expression and resolution level ($F(12, 108) = 1.998$, $p < 0.05$).

As was expected - and in accordance with the pattern of results obtained in Cunningham et al. [2005] and Wallraven et al. [2005], we found that some expressions were more easily recognized than others. In particular, "thinking" and "confusion" are hard to recognize and are often confused with each other—these expressions also cause the overall "low" level of performance of 60%. More interesting is the



(a)

(b)

Fig. 4. Experiment 2. Recognition results broken down by (a) stylization technique (showing correct, incorrect and "none-of-the-above" answers as three separate bars), and (b) stylization technique and expression.

main effect of stylization technique (Figure 4a), here we found that the brush-stroke stylization was significantly worse than the remaining three rendering methods (t test: all $p < 0.05$). In addition, the illustrative stylization was slightly better (t test: $p = 0.05$, $p = 0.07$, marginally significant) than both avatar and cartoon conditions. Interestingly, a closer analysis of the incorrect and "none-of-the-above" responses for all techniques showed that the illustrative stylization had significantly less incorrect responses than the other techniques (t test: all $p < 0.05$), whereas the brush stroke stylization had a significantly higher amount of "none-of-the-above" responses than the other three techniques (t test: all $p < 0.05$). These results can be summarized by ranking the four different techniques according to their *discriminatory performance*: in this case, the illustrative stylization technique supports the most discriminative recognition performance, followed by the original avatar and the cartoon stylization at the same rank, followed by the brush-stroke stylization.

As suggested by the significant interaction between expression and method (see also Figure 4b), which stylization produces the best recognition performance depends on what the expression is. The illustrative stylization produced superior performance for some expressions ("fear," to a lesser degree, also "disgust"). For "surprise," the original avatar animation is most easily recognizable. For "thinking," both cartoon and illustrative stylization provide increased performance, whereas the "sad," "confused," and "happy" expressions show no clear trend favoring one single technique. This pattern of results suggests that different techniques emphasize different types of information that is relevant for the recognition of certain expressions.

Finally, an examination of the interaction between expression and resolution level shows that across all techniques "thinking" is recognized significantly better in the most detailed level, whereas ”fear” is recognized much better in the lowest resolution. For the remaining expressions, no clear dependence on resolution level could be found. The result for "thinking" is because of the fact that this expression requires a close analysis of the eye motion in order to be reliably recognized [Cunningham et al. 2005]. This eye motion, however, is masked by the coarser, blurred visual information in the lower resolution levels. In contrast, "fear" is recognized much better in the lowest resolution level. This expression is driven mainly by the large amount of rigid head motion [Wallraven et al. 2005], which is more visible at the lowest level.

4.2.2.2 *Response Times.* Overall, response times in this experiment showed no significant effects. Restricting the analysis to just the correctly answered trials, we found a small, but significant, increase in response times for the brush stylization method (2.5 as opposed to 2.2 s for the other three methods). This small increase most probably mirrors the impaired recognition performance observed for this method. In general, however, stylization incurred no additional cost in processing time. In other words, we did not find a speed-accuracy tradeoff as in Fischer et al. [2006a]. In addition, no effect of resolution level on reaction times was found, which provides additional support for the data gathered in Wallraven et al. [2005], who also found no effect of resolution level on response times.

4.2.2.3 *Intensity.* For intensity ratings,[3] the ANOVA found main effects of expression ($F(6, 54) = 7.882, p < 0.001$), stylization technique ($F(3, 27) = 7.02, p = 0.001$), resolution level ($F(2, 18) = 9.192, p < 0.01$) as well as interactions of expression and method ($F(18, 162) = 2.285, p < 0.01$) and stylization technique and resolution level ($F(6, 54) = 2.596, p < 0.05$).

Similar to the results in [Wallraven et al. 2005], we found a large main effect of expressions—emotional expressions, such as "disgust" and "fear" were rated as more intense. We again found a main effect of stylization technique (Figure 5a). In this case, brush stylization was rated as much less intense than the other three methods (t tests, all $p < 0.05$). One reason for this is that the brush-stroke pattern masks both rigid and nonrigid head motion, which are highly correlated with ratings of perceived intensity [Wallraven et al. 2005]. Analysis of the interaction between expression and method revealed, in particular, that "happy," "sad," "surprise," and "thinking" were rated as much less intense for the brush-stroke technique than the remaining three expressions. Finally, the main effect and interaction for resolution level showed a large decrease in intensity for the avatar and the brush-stroke technique at the lowest resolution level, whereas this decrease was less pronounced for the cartoon technique, and virtually absent for the illustrative technique. It, thus, seems that the illustrative technique provides a very stable impression of intensity, even at low resolutions.

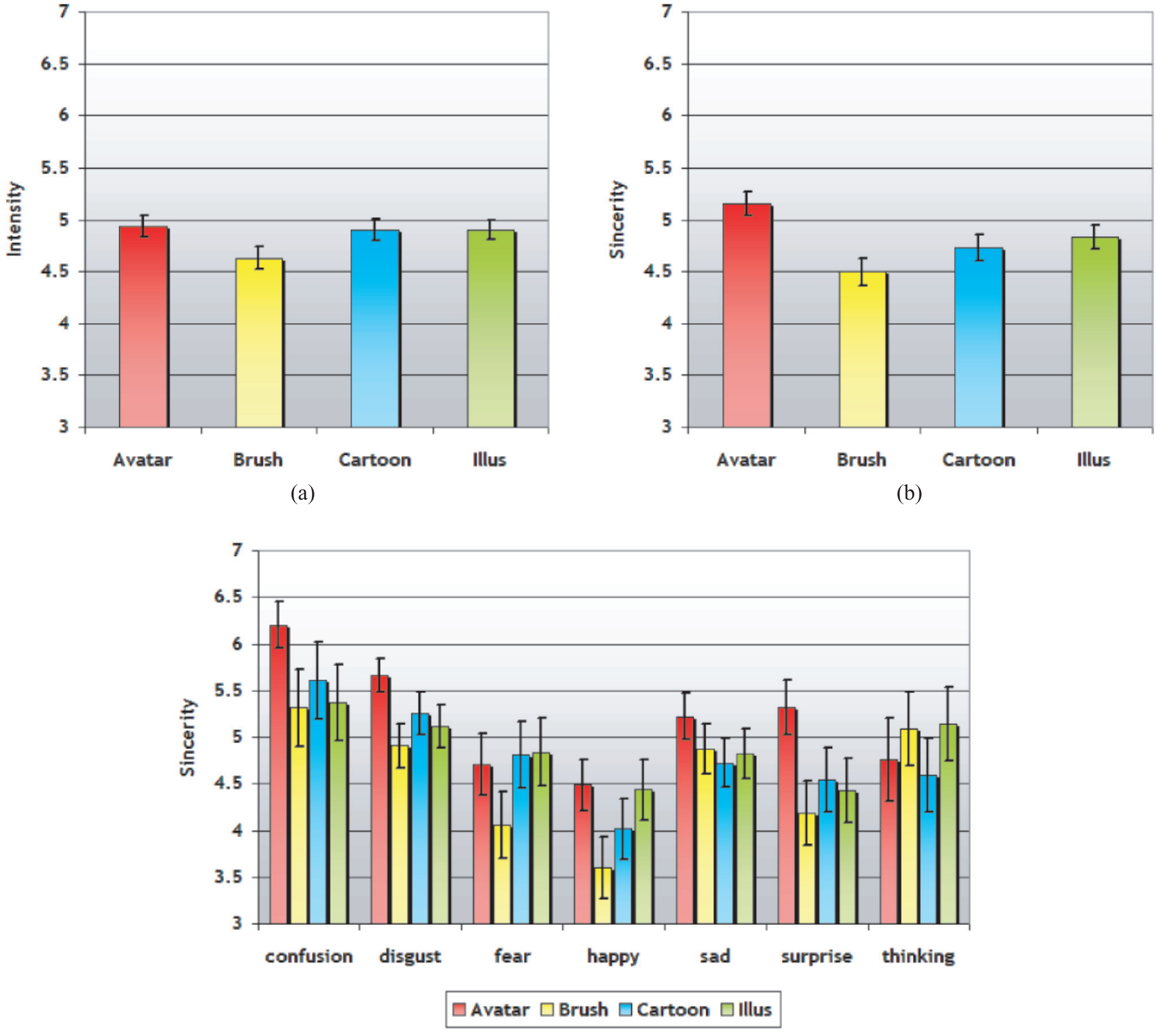
4.2.2.4 *Sincerity.* We found main effects of expression ($F(6, 54) = 3.602, p < 0.01$), stylization technique ($F(3, 27) = 2.92, p = 0.05$), resolution level ($F(2, 18) = 6.736, p < 0.01$) as well as an interaction of expression and method ($F(18, 162) = 2.090, p < 0.01$).

Both main effects of expressions and resolution level provide additional support for the data found in Wallraven et al. [2005]. Most importantly, lower resolutions provide less sincere expressions across all techniques (Figure 5b). The interaction of expression and method is shown in Figure 5c. In particular, "confusion," "disgust," and "surprise" are rated as most sincere in the original avatar animation, whereas for "fear" and "happy," the brush-stroke technique is rated as the least sincere of all techniques.

4.2.3 *Questionnaires.* Both sets of questionnaires in Experiments 1 + 2 provided similar trends, suggesting that the difference in task did *not* influence the introspective rankings of the different techniques. We, therefore, pooled the answers to the questionnaires for the final analysis. Analysis of this data was done by determining for each of the possible 12 ranks the winning technique (a combination of rendering technique and a specific resolution level). These results are summarized in Table I (note that double entries can occur using this analysis).

4.2.3.1 *Aesthetic Preference.* The clear winner in terms of aesthetic preference is the original avatar animation. Of the stylization techniques, the illustrative stylization was judged as most aesthetic, followed by cartoon and brush stylization. One of the reasons why we did not find a clearer preference for one of the stylized techniques is probably that participants regarded all of techniques as equal, rather than judging them as one nonstylized and three stylized versions.

[3] Both intensity and sincerity ratings were analyzed only for correct answers.



(a)

(b)

(c)

Fig. 5. Experiment 2. (a) Intensity ratings and (b) sincerity ratings broken down by stylization technique; (c) sincerity ratings broken down by technique and expression

4.2.3.2 *Effectiveness Preference.* For effectiveness, the avatar animation was ranked highest, followed closely by the illustrative and cartoon techniques, whereas the brush-stroke method was judged as least effective. This pattern mirrors more the one found in Experiment 1 rather closely, although the degree to which the techniques are separated in terms of their preference was much more pronounced in Experiment 1.

4.2.3.3 *Subjective Preference.* For subjective preference, the ranking ordering changes. Here, the illustrative style clearly wins, followed by the avatar and cartoon renderings. As with the previous

Table I. Questionnaires[a]

Rank	Aesthetic	Effectiveness	Subjective
1	Ava1	Ava1	Ill1
2	Ava2	Ill1	Ava1
3	Ill1	Ava2,Car1	Ill1
4	Ava1	Ava2	Ill2
5	Ill2	Car2	Ill3
6	Car2	Ill2	Ava2,Car2
7	Ava2,Car2,Ill3	Ill2	Ava3
8	Car1	Car3	Ava2,Car3,Ill2
9	Car3	Ava3,Bru1	Ava3
10	Bru1	Bru1	Bru1
11	Bru2	Bru2	Bru2
12	Bru3	Bru3	Bru2

[a] Results for aesthetic, effectiveness, and subjective preference judgments. Abbreviations indicate stylization technique and resolution level, respectively.

measures, the brush-stroke technique comes in last. It is interesting that this measure clearly differs from the aesthetic preference—at least for subjective ratings. It seems that stylization is preferred much more than the original animation.

Finally, for all measures, the questionnaire responses show a clear preference ordering of the resolution levels from high to middle to low—a pattern that was seen throughout this study. Overall, the pattern seems to correspond quite well to the one found in Experiment 1. The main disadvantage of the questionnaire analysis, however, as can be seen from Table I, is that it allows only for a rather coarse interpretation of the results. For a more in-depth analysis, other approaches, such as done in the two main experiments, are required.

4.2.4 *Summary.* While we found an effect of stylization across all measures, the most important result of this experiment is that the pattern of expression recognition did not match the subjective judgments measured in Experiment 1. In Experiment 1, participants thought that the avatar captured the essence of the expressions best. In Experiment 2, however, illustrative stylization resulted in the highest level of discriminative recognition, demonstrating a small, but significant, advantage of abstraction for recognition performance. We found no effect of response times, which shows that abstraction of information induces no time penalty. Analysis of the sincerity ratings revealed that stylization might have an adverse effect after all. Here, the original avatar animation had the highest degree of sincerity. Finally, results of the questionnaires showed that participants' introspective responses were rather similar to Experiment 1, even though the question of "effectiveness" should be much more correlated with the recognition results in this experiment.

4.3 Experiment 3—Real-World Video Sequences

In the third experiment, we were interested whether the application domain would make a difference in performance. More specifically, the experiment replicated Experiment 2 using the *real-world sequences* described earlier. Again, the task employed several perceptual measures (including recognition, reaction time, and the perception of intensity and sincerity) in order to examine the effect of stylization across technique, expression, and stylization level.

4.3.1 *Design.* The setup and design of this experiment were identical to Experiment 2. The first task was to *identify* the expression by selecting the name of the expression from a list displayed on the side of the screen. As this experiment is based on the video data used in Cunningham et al. [2003], the list of expressions included agreement, confusion, disagreement, disgustedness, don't know, happiness, sadness, a pleasant surprise, and a thoughtful expression. The list of choices included all nine expressions as well as "none-of-the-above" (a ten-alternative nonforced-choice task). Similar to the previous experiment, we also asked participants to rate the *intensity* and the *sincerity* of the expressions. The experiment used 3 repetitions of each sequence, yielding a total of (9 expressions) (4 stylization techniques) (3 resolution levels) (3 repetitions) = 324 trials. After the experiment, participants filled out the same questionnaire as in the previous two experiments—this time using static peak frames from the video sequences.

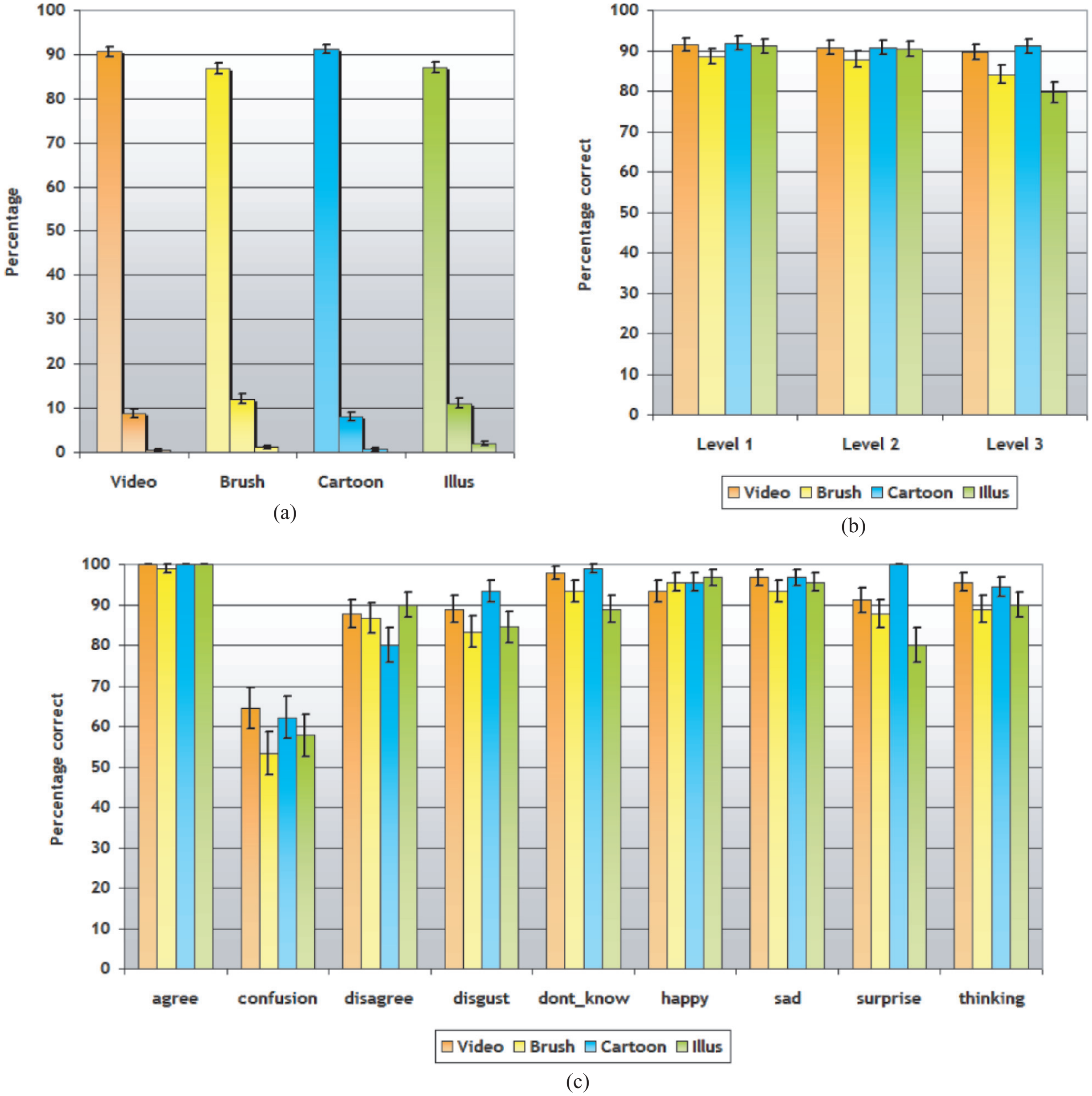
4.3.2 *Results and Discussion.* Data for this experiment were collected from ten participants who had not taken part in the previous experiments. Again, ANOVA methods were used to investigate statistical effects of the factors expression, stylization technique, and resolution level for the different measures (recognition, reaction time, intensity, and sincerity). In the following discussion, we mainly focus on effects involving the stylization method and the resolution level as these are the main factors of interest in the context of this paper.

4.3.2.1 *Recognition.* The ANOVA found main effects of expression ($F(8, 72) = 6.353$, $p < 0.001$), stylization method ($F(3, 27) = 4.963$, $p < 0.01$), and resolution level ($F(2, 18) = 13.808$, $p < 0.001$) as well as interactions between expression and method ($F(24, 216) = 1.649$, $p < 0.05$), expression and resolution level ($F(16, 144) = 2.834$, $p < 0.001$), and method and resolution level ($F(6, 54) = 2.821$, $p < 0.05$).

As Figure 6a shows, overall recognition performance was higher for the six overlapping expressions (confusion, disgust, happy, sad, surprise, and thinking) in the case of real-world sequences than in the case of animated sequences. The general level of performance for real-world data, in addition, is in line with previous results on these video sequences [Cunningham et al. 2003]. This shows, first of all, that animated sequences still need improvements until they capture the same degree of recognizability as real-world data (see also Wallraven et al. [2005]. Turning to the main effect of stylization method, we found that both the original video sequences as well as the cartoon stylization were equally well recognized, whereas there was a significant drop in performance for the brush and illustrative methods; this pattern is also mirrored in the distribution of incorrect and "none-of-the-above" answers (Figure 6a). As Figure 6b shows, this effect is dependent on a given expression: for "surprise," for example, the cartoon stylization provides almost perfect recognition performance compared to the other methods, whereas for "agree," "happy," and "sad" performance is the same across methods. Finally, the interaction of method and resolution level plotted in Figure 6c reveals that the differentiation in performance is mainly because of the coarsest resolution level: performance for both the original and the cartoonized versions stays the same throughout all resolution levels, whereas illustrative and especially brush-stroke stylization seem to provide far less adequate cues for robust recognition at the coarsest level.

4.3.2.2 *Response Times.* The analysis revealed main effects of expression ($F(8, 72) = 4.343$, $p < 0.001$), stylization method ($F(3, 27) = 3.460$, $p < 0.05$), and resolution level ($F(2, 18) = 7.621$, $p < 0.01$) as well as interactions between expression and method ($F(24, 216) = 1.936$, $p < 0.01$), expression and resolution level ($F(16, 144) = 2.186$, $p < 0.01$), and method and resolution level ($F(6, 54) = 3.797$, $p < 0.01$).

First of all, average response times were a little higher than for the animated sequences (2.8 versus 2.4 s), this increase in response times, however, could be because of the slightly longer average sequence lengths for the video sequences (1.45 average length compared to 1.2 s for the animated sequences). More interestingly, we found several significant effects of stylization method and resolution level on the response times, which is in contrast to the previous experiment. For the main effect of stylization



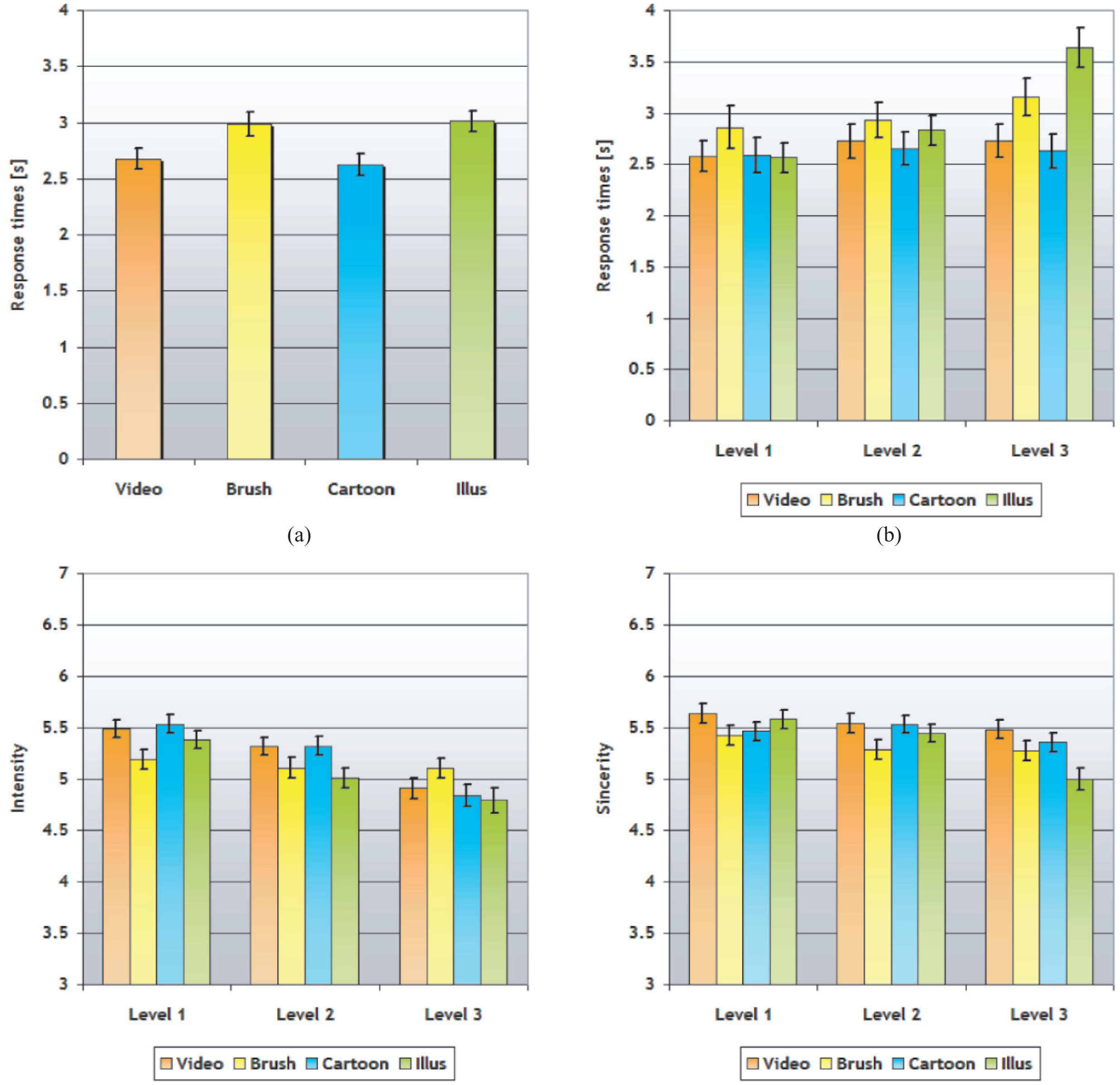
(a)

(b)

(c)

Fig. 6. Experiment 3. Recognition accuracy for real-world video sequences broken down by (a) stylization method (showing correct, incorrect and "none-of-the-above" answers as three separate bars), (b) resolution level and method, and (c) expression and method.

method that is shown in Figure 7a, we found that both the original sequences as well as the cartoonized sequences were recognized equally fast, whereas the illustrative and brush stylization were recognized significantly slower. The interaction between method and resolution level (see Figure 7b) shows that this increase is mainly because the coarsest resolution level for which the latter two methods result



(a) (b)

(c) (d)

Fig. 7. Experiment 3. Response times for real-world video sequences broken down by (a) stylization method, (b) method and resolution level, (c) intensity ratings, and (d) sincerity ratings broken down by stylization method and resolution level.

in significantly slower response times. This parallels the pattern observed for the recognition accuracy and shows that strong stylization not only impacts recognition, but also response times for those two methods. A closer look at the interaction between expression and method shows that response times for the cartoonized versions of both “sad” and “surprise” are significantly faster in comparison to the other methods.

4.3.2.3 *Intensity.* The ANOVA analysis found main effects of expression ($F(8, 72) = 11.553$, $p < 0.001$) and resolution level ($F(2, 18) = 18.434$, $p < 0.001$), as well as interactions between expression and method ($F(24, 216) = 3.936$, $p < 0.001$), expression and resolution level ($F(16, 144) = 3.554$, $p < 0.001$), and method and resolution level ($F(6, 54) = 2.914$, $p < 0.05$).

A post-hoc test on stylization method, in addition, revealed that the illustrative stylization method was judged as significantly less intense than both the original and the cartoonized versions. This result is in contrast to Experiment 2, in which we found that brush-stroke stylization was consistently judged as less intense. Paralleling the results found in Experiment 2, however, perceived intensity is dependent on the resolution level (see Figure 7c)—higher abstraction levels result in less intensity. Interestingly, this pattern is not present for the brush stylization method (see also Figure 7c): across all resolution levels, perceived intensity stays constant for this method.

4.3.2.4 *Sincerity.* The analysis revealed main effects of expression ($F(8, 72) = 5.485$, $p < 0.001$), stylization method ($F(3, 27) = 5.603$, $p < 0.01$), and resolution level ($F(2, 18) = 7.530$, $p < 0.01$) as well as interactions between expression and resolution level ($F(16, 144) = 2.238$, $p < 0.01$).

As Figure 7d shows, the original sequences were judged as being more sincere than both brush-stroke and illustrative stylization; this difference was not significant compared to the cartoon stylization. The general trend, however, mirrors the results found in the previous experiment: stylized versions of the original sequences are judged as being less sincere than the original. Again, this effect depends on the resolution level: coarser abstraction results in less sincerity.

4.3.3 *Summary.* The results on the real-world video sequences parallel the ones found for animated sequences: stylization has an effect across all measures. The exact pattern of results, however, changes: here, the cartoonized sequences are on par with the original sequences whereas the other two stylization techniques result in a drop in performance for the highest abstraction levels. Overall, however, it seems that the adverse effects on recognition accuracy and response times that we observed for the brush-stroke and illustrative methods are mainly because of the coarsest resolution level. The fact that we could not see any increase in performance for the cartoonized sequences (similarly to the increase for the illustrative stylization in the previous experiment) might be because of a "ceiling" effect—recognition performance was already fairly high to begin with, which might have masked any consistent positive effects of stylization that were present, for example, in the case of the "surprise" expression. Finally, the effects of stylization on intensity and sincerity paralleled those found in the previous experiment. The results stress the need for designers of human-machine interface agents to pay close attention to the abstraction level, as higher abstraction (or coarser resolution) results in a decrease in both perceived intensity and sincerity compared to the original sequences.

4.3.4 *Questionnaires.* Analysis of the experimental questionnaire data was again done by determining for each of the possible 12 ranks the winning technique (a combination of rendering technique and a specific resolution level). These results are summarized in Table II (again, note that double entries can occur using this analysis).

4.3.4.1 *Aesthetic Preference.* This data is the hardest to interpret as preference for a particular stylization technique varies across participants. The *least* preferred class of stimuli, however, is given by the original and blurred video sequences. This consistent choice represents a much clearer response favoring stylized techniques over the original sequences than in the previous experiment, for which the "original" avatar animation still scored high aesthetic rankings. One of the reasons for this could be—as mentioned earlier—that participants did not judge the avatar animation as being "real-world" data, but rather that they saw it as another artistic rendering style. Again, this underscores the need for the animation to become more realistic both in terms of task-dependent and introspective measures.

Table II. Questionnaires[a]

Rank	Aesthetic	Effectiveness	Subjective
1	Car1,Ill3	Vid1	Vid1
2	Bru3,Ill3	Vid1,Vid2	Vid2
3	Bru2,Ill1	Vid2	Vid3
4	Bru1	Bru1,Car1	Car1
5	Bru2,Car3	Ill1	Car2
6	Bru2,Bru3	Bru1	Ill1,Bru2
7	Car2,Car3	Bru2	Bru2
8	Ill1	Vid3	Bru1,Car3
9	Ill2	Car2	Bru2
10	Vid3	Car3	Bru3
11	Vid2	Bru2	Ill2
12	Vid1	Ill3	Ill3

[a] Results for aesthetic, effectiveness, and subjective preference judgments for the video sequences. Abbreviations indicate stylization technique and resolution level, respectively.

4.3.4.2 *Effectiveness Preference.* Similarly to the previous experiments, participants clearly judge the original images to provide the most effective recognizability. The original video sequences are followed by the brush-stroke, cartoon, and illustrative stylization methods. Again, this introspective judgment does not agree with the data from the recognition experiment: here, both the original sequences and the cartoonized sequences resulted in equally good performance.

4.3.4.3 *Subjective Preference.* As in the previous questionnaire data, participants here prefer the original stimuli over the stylized material. Following, this there is a clear ordering of preference, in which cartoon is preferred over brush which, in turn, is preferred over the illustrative stylization technique.

With the exception of the aesthetic preference, all questionnaire responses again show a very clear preference ordering of the resolution levels from high to middle to low, confirming the pattern observed throughout this study. Overall, we could observe a very similar pattern of responses for stylized video sequences as for animated sequences—perhaps most importantly, we have found additional evidence that there are crucial differences between what participants think to be an effective rendering method for recognition and what we can actually observe in a direct test of recognizability.

5. CONCLUSION AND OUTLOOK

In this paper, we presented a series of evaluations of three different stylization techniques in the context of facial expressions and found effects of stylization on almost all measures. The first experiment investigated the question of effectiveness in a direct comparison task for animated sequences. The results indicated that stylization would potentially reduce effectiveness compared to the original avatar animation. A similar pattern of results was found for the introspective evaluation of effectiveness using questionnaires. In the second experiment, we collected several task-specific measures that were centered on recognizability as well as perceived intensity and sincerity. The results did not correlate with the introspective effectiveness measures—the illustrative stylization provided the most discriminatory performance. This pattern was confirmed in the third experiment in which we investigated the effects of stylization on task-dependent and introspective measures using real-world sequences.

The most obvious explanation for the difference between the recognition results and the introspective evaluation and direct comparison of "effectiveness" is that they do not measure the same thing. This explanation is, in part, contradicted by the fact that almost all participants mentioned during debriefing that recognizability was one the central criterion they used in determining "effectiveness." Our study thus has shown the need for a more encompassing and, if possible, task-dependent evaluation that allows to build a more complete picture of the perceptual impact of stylization methods.

Experiment 3 served as a replication of Experiment 2 with real-world sequences. Here, we also found an influence of stylization on all measures—ranging from differences in recognition accuracy to perceived sincerity of expressions. The impact of the stylization methods on the different measures, however, was different from Experiment 2: for recognition, for example, we found that cartoonized versions were equally well recognized as the original sequences, whereas both brush-stroke and illustrative stylization techniques resulted in a drop in recognition accuracy, as well as an increase in response times. The fact that we could not find an increase in discrimination for any stylization might be due to the relatively high recognition performance ("ceiling effect"). As discussed earlier, one might suspect that the reason for the differences in performance between the real-world sequences and the animated sequences in general lies in the different levels of high-frequency detail that are contained both in the static texture, as well as in the dynamics of the motion of facial features: in both cases, the avatar has much more homogeneous characteristics. This issue needs to be investigated in future studies, in which we will manipulate those high-frequency details in the video sequences in order to examine their influence on the different measures. Another issue that one can derive from the relatively poor performance of the brush technique is the importance of dynamic information: the static stroke pattern results in dynamic noise—a flickering that is especially noticeable around the eyes. Taken together with the results of earlier studies, which highlight the importance of correct facial dynamics for recognition [Wallraven et al. 2005], our results show that in order to support real-world performance, one needs to make sure that the temporal qualities of facial expressions are preserved in animated sequences.

In terms of practical applications, the results of our experiments can, however, be already summarized as preliminary guidelines for effective rendering: On the one hand, if the goal is to convey an *animated* facial expression most effectively, choosing a stylized rendering method (such as illustrative rendering) might help—apart from offering other dimensions, such as, aesthetics and sparse representation. On the other hand, if the goal is to provide subjective certainty about the conveyed expression, one needs to resort to a "realistic" rendering method. In addition, one should, in general, avoid high degrees of abstraction for stylization (this is true both for real-world data and for animated sequences) as these adversely influence the perceived intensity and sincerity.

The systematic investigation of the visual information that drives the exact pattern of results observed in the stylized sequences (especially the interaction effects) will need to be done in future studies—nevertheless, we might speculate, for example, that for the illustrative technique in the case of animated sequences, one of the reasons for its comparatively good performance lies in the emphasis of shape: hatching in connection with silhouette lines that highlight small details of the face could give extra cues for more reliable recognition. In addition, previous studies have shown that the loss of color does not impact recognition of identity [Yip and Sinha 2002]; our study has shown the same for expression recognition for the illustrative technique both for animated sequences and to a lesser degree (for the first two resolution levels) also for real-world sequences.

The relatively small, yet significant, effects of stylization found in this study are in contrast to the studies by Gooch and Willemsen [2002] and Winnemöller et al. [2006] who found clearer advantages of stylization on naming. One of the crucial differences between their and our study is the task: here, we were interested in recognition of facial expressions as opposed to recognition of identities. Given that the stylization process results in sometimes rather dramatic changes to the image, the fact that we found a significant drop (that is, >10%) in performance only for the coarsest levels of abstraction, testifies to the inherent robustness with which we process facial expressions. It would be interesting to repeat the experiments done here with an identity recognition task in order to test whether the difference between introspective and task-dependent measures would also show up for this new task.

In summary, our study has evaluated the effectiveness of three different stylization techniques across multiple perceptual and introspective dimensions. Our results have provided further insight into the robustness of expression recognition as well as demonstrated critical differences of evaluation methodology.

REFERENCES

ADOLPHS, R. 2002. Recognizing emotions from facial expressions: Psychological and neurological mechanisms. *Behavioral and Cognitive Neuroscience Reviews 1,* 1, 21–61.

AGRAWALA, M. AND STOLTE, C. 2001. Rendering effective route maps: Improving usability through generalization. In *Proc. of ACM SIGGRAPH*. ACM Press, New York. 241–249.

BREIDT, M., WALLRAVEN, C., CUNNINGHAM, D. W., AND BÜLTHOFF, H. H. 2003. Facial animation based on 3d scans and motion capture. *SIGGRAPH '03 Sketches & Applications*.

BULL, P. 2001. State of the art: Nonverbal communication. *The Psychologist 14*, 644–647.

CUNNINGHAM, D. W., BREIDT, M., KLEINER, M., WALLRAVEN, C., AND BÜLTHOFF, H. H. 2003. How believable are real faces: Towards a perceptual basis for conversational animation. In *Computer Animation and Social Agents 2003*. 23–39.

CUNNINGHAM, D., NUSSECK, M., WALLRAVEN, C., AND BÜLTHOFF, H. 2004. The role of image size in the recognition of conversational facial expressions. *Computer Animation & Virtual Worlds 15,* 3–4 (07), 305–310.

CUNNINGHAM, D., KLEINER, M., WALLRAVEN, C., AND BÜLTHOFF, H. 2005. Manipulating video sequences to determine the components of conversational facial expressions. *ACM Transactions on Applied Perception 2,* 3 (07), 251–269.

DECARLO, D. AND SANTELLA, A. 2002. Stylization and abstraction of photographs. In *Proc. of ACM SIGGRAPH*. 769–776.

EKMAN, P. 1972. *Universal and Cultural Differences in Facial Expressions of Emotion*. University of Nebraska Press, Lincoln, NB. 207–283.

FERWERDA, J. 2003. Three varieties of realism in computer graphics. In *Proc. of SPIE Human Vision and Electronic Imaging*. 290–297.

FISCHER, J., BARTZ, D., AND STRASSR, W. 2005a. Artistic reality: Fast brush stroke stylization for augmented reality. In *Proc. of ACM Symposium on Virtual Reality Software and Technology (VRST)*. 155–158.

FISCHER, J., BARTZ, D., AND STRASSR, W. 2005b. Illustrative display of hidden iso-surface structures. In *Proc. of IEEE Visualization*. 663–670.

FISCHER, J., CUNNINGHAM, D., BARTZ, D., WALLRAVEN, C., BÜLTHOFF, H., AND STRASSR, W. 2006a. Measuring the discernability of virtual objects in conventional and stylized augmented reality. In *Eurographics Symposium on Virtual Environments (EGVE)*.

FISCHER, J., EICHLER, M., BARTZ, D., AND STRASSR, W. 2006b. Model-based hybrid tracking for medical augmented reality. In *Eurographics Symposium on Virtual Environments (EGVE)*.

FREUDENBERG, B., MASUCH, M., AND STROTHOTTE, T. 2002. Real-time halftoning: A primitive for non-photorealistic shading. In *Proc. of Eurographics Workshop on Rendering*. 227–231.

GOOCH, A. A. AND WILLEMSEN, P. 2002. Evaluating space perception in npr immersive environments. In *NPAR '02: Proceedings of the 2nd International Symposium on Non-Photorealistic Animation and Rendering*. ACM Press, New York. 105–110.

GOOCH, B., REINHARD, E., AND GOOCH, A. 2004. Human facial illustrations: Creation and Psychophysical Evaluation. *ACM Trans. Graph. 23,* 1, 27–44.

HAEBERLI, P. 1990. Paint by numbers: Abstract image representations. In *Proc. of ACM SIGGRAPH*. 207–214.

KLEINER, M., WALLRAVEN, C., AND BÜLTHOFF, H. 2004. The mpi videolab - a system for high quality synchronous recording of video and audio from multiple viewpoints. Tech. Rep. 123. May.

LITWINOWICZ, P. 1997. Processing images and video for an impressionist effect. In *Proc. of ACM SIGGRAPH*. 407–414.

SANTELLA, A. AND DECARLO, D. 2004. Visual interest and npr: An evaluation and manifesto. In *NPAR '04: Proceedings of the 3rd international symposium on Non-Photorealistic Animation and Rendering*. ACM Press, New York. 71–150.

STOKES, W. A., FERWERDA, J. A., WALTER, B., AND GREENBERG, D. P. 2004. Perceptual illumination components: A new approach to efficient, high quality global illumination rendering. *ACM Trans. Graph. 23,* 3, 742–749.

STROTHOTTE, T. AND SCHLECHTWEG, S. 2002. *Non-Photorealistic Computer Graphics - Modelling, Rendering, and Animation*. Morgan Kaufmann Publi., San Francisco, CA.

TOMASI, C. AND MANDUCHI, R. 1998. Bilateral filtering for gray and color images. In *Proc. of IEEE International Conference on Computer Vision (ICCV)*. 839–846.

WALLRAVEN, C., CUNNINGHAM, D., BREIDT, M., AND BÜLTHOFF, H. 2004. View dependence of complex versus simple facial motions. H. H. Blthoff and H. Rushmeier, Eds. *Proceedings of the First Symposium on Applied Perception in Graphics and Visualization*, 181.

WALLRAVEN, C., BREIDT, M., CUNNINGHAM, D. W., AND BÜLTHOFF, H. H. 2005. Psychophysical evaluation of animated facial expressions. In *Proc. of APGV '05*. ACM Press, New York, NY, USA, 17–24.

WINNEMÖLLER, H., OLSEN, S. C., AND GOOCH, B. 2006. Real-time video abstraction. In *SIGGRAPH '06: ACM SIGGRAPH 2006 Papers*. ACM Press, New York. 1221–1226.

YIP, A. W. AND SINHA, P. 2002. Contribution of color to face recognition. *Perception 31,* 8, 995–1003.

Received December 2006; revised June 2007; accepted June 2007