Running head: ROBUSTNESS OF EMOTIONAL EXPRESSION RECOGNITION UNDER LOW VISIBILITY

Robustness of emotional expression recognition under low visibility

DongWon Oh¹, Daniel N. Osherson¹, & Alexander Todorov¹

¹Department of Psychology, Princeton University

Please address correspondence to:

DongWon Oh
Department of Psychology
Princeton University
Princeton, NJ 08540
dongwono@princeton.edu

Manuscript word count: 5506
Abstract

Human beings have the capacity to recognize facial emotions even under conditions of imperfect visual access to the face. We tested for the robustness of this process using different kinds of visual noise simulating occlusion, peripheral or unattended vision, and distant vision. We prepared videos of dynamic faces exhibiting one of six basic emotions. The videos were then degraded with either partial occlusion, low contrast, or low pass spatial frequency filtering, to various degrees. Participants were able to recognize the intended emotion in videos under severe visual noise of all types, showing the most robust recognition of surprised expressions. Additionally, the pattern of confusability of one emotion for another was sustained across different kinds and levels of distortion, suggesting the same perceptual mechanism at work under various levels of noise.

Keywords: emotions; faces; dynamic expressions; expression similarity
Introduction

Human beings are adept at recognizing facial emotions though they do not always have full access to relevant visual information when interpreting such emotional gestures. For example, facial expressions might be observed despite occluding objects, in peripheral vision, or at a distance. In this paper, we investigate the impact of such visual degradation on the perception of emotion. For this purpose, we compared the recognizability of emotions as a function of various types and levels of visual noise.

Previous studies have shown that emotion recognition from facial gestures remains robust even with various types of incomplete visual information. For instance, studies had people view videos of moving light points on a face making an emotional expression, where facial features were obscured in the dark and only the motion of expressions was visible (Bassili, 1978, 1979). The observers showed high accuracy in emotion recognition. People were also able to correctly detect dynamic emotions from computer-generated synthetic expressions and their degraded variants, such as expressions without head motions or facial textures (Kätsyri & Sams, 2008; Wallraven, Breidt, Cunningham, & Bülthoff, 2008). Similarly, people were able to identify emotions, especially anger, after viewing only a small part of a dynamic body image (e.g., face, arm) or their degraded forms (e.g., silhouette, dots; Visch, Goudbeek, & Mortillaro, 2014). However, in the previous studies, the levels of noise superimposed on dynamic gestures are hard to quantify. Studies which quantitatively varied the level of noise have mostly been limited to static images (e.g., F. W. Smith & Schyns, 2009). In contrast, the present study focuses on the impact of quantitatively varying noise levels on the recognition of emotion from dynamic faces, in order to better understand the relation between the visibility and the recognizability of emotional expressions.

Different kinds of visual noise can result in qualitatively different effects on facial emotion recognition, across emotions as well as across levels of noise. To address this possibility, we chose types of visual noise that 1) simulate forms of noise that we experience in our life in various forms and 2) have distinctive visual effects on emotion stimuli. Specifically, we created dynamic stimuli that exhibited three types of noise, each of which has unique features: partial occlusion, low contrast, and low pass spatial
frequency filtering. The three types of noise can be considered simulations of occlusion, peripheral or unattended vision, and distant vision, respectively (Figure 1, see Stimuli for detail).

We used dynamic images as emotion stimuli because emotional expressions in real life are dynamic, and more easily recognized than their static counterparts (Ambadar, Schooler, & Cohn, 2005; Cunningham & Wallraven, 2009; Ehrlich, Schiano, & Sheridan, 2000; Kätsyri & Sams, 2008; Wehrle, Susanne, Susanne, & Klaus, 2000; for review, see Krumhuber, Kappas, & Manstead, 2013). The relative ease of recognizing emotions from dynamic rather than static faces cannot be explained by the mere presence of multiple images or motion per se (Ambadar et al., 2005; Cunningham & Wallraven, 2009). Rather, the relative advantage of dynamic faces depends on the perception of change over time from the baseline facial expression (Ambadar et al., 2005) that allows observers to distinguish transitory from permanent facial features (Calder, Burton, Miller, Young, & Akamatsu, 2001; Calder & Young, 2005). This aspect of real-life emotional gestures can be captured only by dynamic stimuli.

We are interested not only in the accuracy of recognition of different emotions as a function of types and levels of visual noise but also in the confusability of emotions. Previous studies have revealed systematic biases in confusion of one emotional expression for another. Specifically, people often mistake anger for disgust and fear for surprise, and vice versa (Bassili, 1978, 1979; Dailey, Cottrell, Padgett, & Adolphs, 2002; Palermo & Coltheart, 2004; Susskind, Littlewort, Bartlett, Movellan, & Anderson, 2007; Tottenham, Tanaka, Leon, McCurry, Nurse, Hare, et al., 2009). This confusion among emotion categories is attributable to the overlapping facial musculatures in the generation of different expressions (C. A. Smith & Scott, 1997). Interestingly, biologically plausible computer algorithms trained to discriminate static basic emotional expressions show confusion errors similar to human errors (Dailey et al., 2002; Susskind et al., 2007). Findings based on degraded dynamic emotional stimuli revealed a similar pattern of confusion as static stimuli (e.g., Bassili, 1978, 1979).

Despite the consistent evidence for patterns of confusability, it is unclear if and how these patterns of confusion will change when the emotional signals are subjected to severe noise. In particular, noise might produce different patterns of emotion-confusion at different levels of intensity. Alternatively,
despite the decrease in overall recognition accuracy, the pattern of confusion might be sustained until observers reach chance performance.

In sum, by applying qualitatively different types of visual noise at quantitatively different levels, the current study investigates how visibility of dynamic emotional stimuli affects their recognition accuracy and confusion. In other words, the objective of the present research is to examine how the emotion recognition process deteriorates under different kinds of visual noise.

**Experiment 1: Partial Occlusion**

An object occluding the object of interest is one of the most common obstacles in visual perception. Partial visual obstruction also serves as one of the monocular cues for depth. Unlike low contrast (Experiment 2) or low spatial frequency filtering (Experiment 3), occlusion has all-or-nothing effects on the visibility of each part of an image, and its intensity is not even across an image. Here we imposed opaque grey bars on videos to see the observer’s ability to reconstruct facial emotional signals through the occlusion of varying intensity.

**Methods**

*Participants.*

1) Actors

To generate a set of emotional expression stimuli, twenty-five Princeton University undergraduate students and Princeton community members were recruited as actors (19 women; age range: 19-50 years).

2) Observers

Fifty Princeton University undergraduate students and Princeton community members participated as observers for either course credit or monetary compensation (30 women, \( M_{\text{age}} = 20.88 \) years)

*Stimuli.*
Each actor sat alone in front of a Panasonic Lumix DMC-FZ200 camera mounted on a tripod. Actors videotaped themselves while making six basic facial expressions: angry, disgusted, fearful, happy, sad, and surprised. For each expression, they were given a prompt describing a relevant hypothetical situation. These prompts were used in previous research (Ekman & Friesen, 1971; de Gelder & van den Stock, 2011; Hadjikhani & de Gelder, 2003). For example, for the angry expression the prompt read, "You are in a quarrel and you threaten the opponent to fight back. Make an angry facial expression." While videotaping, the actors were instructed by the experimenter to "start a recording with a blank, "neutral" expression, staring at the camera lens, and go up to the peak expression, and then come back to the blank expression, and then end the recording." When making the neutral expression, they were asked to relax their facial muscles and not to make any particular facial or bodily gestures. Actors recorded each of the seven expressions 3 to 5 times. Every video contained the entire face as well as part of the neck and shoulders. The actors were asked to hold their head and posture as steady as possible while filming. They initiated and terminated the recording of each video by pressing the record button on a wired controller, which they held throughout the session in a manner that allowed them a comfortable posture. No feedback was provided concerning the videos. The entire recording process took less than 30 minutes for each actor.

We screened the 3 to 5 videos per emotion using the following criteria: 1) the video started with a neutral facial expression, 2) the intensity of emotion expressed in the video increased and then decreased over time, and 3) the video ended with a neutral facial expression. When there were multiple videos fitting all these criteria, we chose the earliest recording. That way, each actor contributed a single video for each emotion in the final stimulus set.

To manipulate visibility of the videos, we superimposed sixteen vertical and sixteen horizontal gray bars on the videos in Experiment 1. Observers saw only half a bar at the horizontal and vertical edges of the videos, respectively. The bars were positioned so that all the distances between neighboring bars were the same (Figure 1A). The width of the grids was varied to produce 5 levels of visible area in the videos (74.49, 52.64, 34.45, 19.92, and 9.06%).
Procedure and Analyses.

Seated in a dark room, observers watched the emotion videos and guessed the portrayed emotion. Each trial consisted of a fixation point presented at the center of the computer screen (0.5 s), an emotion video (2 - 20 s, median: 5 s), a perceptual mask (0.2 s), and a judgment (self-paced). Observers selected what emotion was expressed in the video out of six response options: angry, disgusted, fearful, happy, sad, and surprised.

Each observer viewed all the videos and was exposed to each emotional expression of each actor only once. Videos with different visibility levels were presented with equal frequency to each observer. The visibility levels applied to the videos were counterbalanced across observers. The experiment consisted of 150 trials in total (25 actors × 6 emotions) and took less than 30 minutes.

We used an unbiased hit rate ($H_U$) as the index of emotion recognition performance (Wagner, 1993). The index is the joint probability that “a stimulus category is correctly identified given that it is presented at all and that a response is correctly used given that it is used at all.” It has many desirable characteristics as an accuracy measure for separate-category responses. Unlike a hit rate, an unbiased hit rate takes into account the observer’s bias to use specific categories more often than others. Also, unlike $d'$, the signal detection theory measure (Swets, 1964), it provides separate measures of performance for different response categories without additional assumptions or calculations. In addition, its calculation and interpretation is more straightforward than that of $d'$.

To compute unbiased hit rates, for each observer we made a confusion matrix at each visibility level. Then further analyses were performed to see how emotion and visibility affected recognition performance, to compare the robustness of recognition processing across emotion, and to see if and to what extent the same mechanism was engaged in recognizing emotion as emotion becomes less visible.

Results and Discussion

As shown in Figure 2, with the smallest amount of occlusion, the observers showed high accuracy on average, which validated the distinctiveness of each emotion category in the original videos. The
observers showed above-chance level accuracy in all emotion × visibility conditions except for fearful expressions (Bonferroni correction; all t’s > 3.978, all p’s < .008). The recognition performance deteriorated as the occlusion noise increased which was shown by the fact that a two-way 6 (emotions) × 5 (visibility levels) repeated-measures ANOVA on Hit’s found a significant effect of occlusion ($F(4, 196) = 79.675, p < .001, \eta^2 = .874$) and emotion ($F(5, 245) = 171.290, p < .001, \eta^2 = .950$). There was no significant effect of the interaction between the two variables ($F(20, 980) = 1.741, p = .082$). Post-hoc $t$-tests showed that happy expressions were better recognized than all the other expressions; angry and disgusted expressions were better recognized than fearful, sad, and surprised expressions. Fearful expressions were more poorly recognized than all the other expressions (Bonferroni correction; all t’s > 3.427, all p’s < .003).

To compare the robustness of recognition processing across emotions, we first conducted regression analysis for each participant predicting their Hit’s from the visibility of the emotion. Second, we conducted a one-way repeated-measures ANOVA on the regression coefficients. A smaller absolute coefficient indicates more robust recognition for one emotional expression versus another, given that performance declines with decreases in visibility. Across emotions, there was a significant difference in the robustness of recognition (Figure 3). This was reflected in the results of the repeated-measures ANOVA ($F(5, 245) = 4.568, p < .001$). Post-hoc $t$-tests found that recognition of surprised expressions was more robust than recognition of angry or sad expressions (all t’s > 4.0, all p’s < .003).

Next, to see if the pattern of emotion confusion was sustained across visibility levels, we conducted correlational analyses across the confusion matrices of the five visibility levels. The correlational analyses between the confusion matrix pairs showed that there was high consistency in how observers make correct and incorrect response categorization judgments across varying levels of visibility (Figure 3; all p’s < .001).

**Experiment 2: Low Contrast**
Lower contrast is another common visual noise in life. For instance, a stimulus looks dimmer (i.e., lower in contrast) on the peripheral vision than on the center (Daitch & Green, 1969). Even within a relatively central visual field, an object may be perceived as lower in contrast if another visual area is covertly attended (Carrasco, 2006; Pestilli & Carrasco, 2005). Contrast strength defines the magnitude of the difference between visual areas, so lowering contrast weakens visual distinctiveness of the configuration of inner facial features as well as the contour of the face and body images. Also, infection or inflammation in the eye or the adjacent structure, which can be caused by macular degeneration or cataracts, lead to dim vision. With low contrast, we simulated these visual effects.

**Methods**

*Participants.*

1) Actors

The same actors' videos that were used Experiment 1 were used to create the stimuli.

2) Observers

Fifty Princeton University undergraduate students and Princeton community members participated as observers for either course credit or monetary compensation (28 women, $M_{age} = 20.06$ years)

*Stimuli.*

We applied 5 levels of low contrast to the emotion videos (Figure 1B; .135, .069, .035, .018, and .009, where 1 indicated the contrast of the original videos). The contrast was manipulated in MATLAB (Mathworks Inc.) with the alpha compositing function of Psychophysical Toolbox 3 where the video was combined with the gray background to create the partial transparency (Brainard, 1997).

*Procedure and Analyses.*

The procedure and analyses were the same as in Experiment 1.
The recognition performance deteriorated as the video contrast decreased (Figure 5). The observers showed above-chance level accuracy in all emotion × visibility conditions, except for disgusted, fearful, and sad expressions under the most severe level of noise (Bonferroni correction; all $t$'s > 3.547, all $p$'s < .008). The repeated-measures ANOVA against $H_U$'s found a significant effect of contrast ($F(4, 196) = 330.057, p < .001, \eta^2 = .966$), emotion ($F(5, 245) = 75.575, p < .001, \eta^2 = .894$), and the interaction between contrast and emotion ($F(20, 980) = 5.169, p < .001, \eta^2 = .775$). Post-hoc $t$-tests showed that happy expressions were better recognized than all the other expressions; angry expressions were better recognized than all the other expressions but happy ones; and disgusted expressions were better recognized than surprised expressions (Bonferroni correction; all $t$'s > 3.427, all $p$'s < .003).

Across emotions, there was a significant difference in the robustness against the contrast noise (Figure 6). This was shown in the results of the repeated-measures ANOVA against the regression coefficients from the model of $H_U$ and visibility ($F(5, 245) = 6.953, p < .001$). Post-hoc $t$-tests found that surprised expressions showed a smaller mean regression coefficient than those of angry, happy, and sad, happy expressions. Also, the recognition of disgusted expressions showed smaller regression coefficients than those of angry expressions (all $t$'s > 3.627, all $p$'s < .003).

The correlational analyses between the confusion matrix pairs showed that there was high consistency in the way people correctly and incorrectly recognize emotional expressions across visibility (Figure 7; all $p$'s < .001). This shows the pattern of emotion confusion was sustained across the levels of noise.

**Experiment 3: Low Pass Filtering**

Low pass filtering is another ecologically plausible and common noise in the process of emotion recognition. It also has different effects on observers’ perception of images than the other two types of noise used in the current study. A low pass filter removes fine-grained information from a face image, which is necessary for identity recognition, blurring the image. The selective perception of low spatial
frequency information of an image is similar to the perception of distant information without any other type of noise (F. W. Smith & Schyns, 2009) or the perception under poor visual acuity.

Methods

Participants.

1) Actors

The same actors' videos that were used Experiment 1 was used to create the stimuli.

2) Observers

Fifty Princeton University undergraduate students and Princeton community members participated as observers for either course credit or monetary compensation (32 women, \( M_{\text{age}} = 20.42 \) years).

Stimuli.

We applied 5 low pass spatial frequency filters with varying cutoffs to the emotion videos (Figure 1C; 0.213, 0.532, 0.851, 1.595, and 2.552 deg/cycle). Specifically, the visual information in each frame from each video that was higher than the cutoff spatial frequency was eliminated with fast Fourier transformation, on/off type low pass filtering, and inverse fast Fourier transformation. The cutoff levels were determined in such a way that the mean accuracy of performance averaged across emotion on each visibility level matched that of Experiment 2, based on a pilot experiment with low pass filtering (\( N=10 \)).

Procedure and Analyses.

The procedure and analyses were identical with Experiment 1 and 2 except that the observers were asked to place their head on a chin rest throughout the experiment. This was in order to make sure that the observers were exposed to the same spatial frequency information from each video. The distance between the monitor and their eyes was approximately 50 cm.

Results and Discussion
The recognition performance deteriorated as the video became blurry (Figure 8), although the observers showed above-chance level accuracy in all emotion × visibility conditions, except for fearful and disgusted expressions (Bonferroni correction; all t’s > 2.774, all p’s < .008). The repeated-measures ANOVA against $H_U$’s found a significant effect of low pass cutoff ($F(4, 196) = 149.859, p < .001, \eta^2 = .929$), emotion ($F(5, 245) = 78.904, p < .001, \eta^2 = .898$), and the interaction between low pass cutoff and emotion ($F(20, 980) = 4.082, p < .001, \eta^2 = .731$). Post-hoc t-tests showed that happy expressions were better recognized than any other expressions; angry expressions were better recognized than all the other expressions but happy ones; and fearful expressions were more poorly recognized than disgusted, sad, and surprised expressions (Bonferroni correction; all t’s > 3.280, all p’s < .003).

Across emotions, there was a significant difference in the robustness against the low pass noise (Figure 9). This was shown in the results of the repeated-measures ANOVA against the regression coefficients from the model of $H_U$ and visibility ($F(5, 245) = 7.833, p < .001$). Post-hoc t-tests found that surprised expressions showed a smaller mean regression coefficient than those of angry, disgusted, happy, and sad expressions (all t’s > 4.385, all p’s < .003).

The correlational analyses between the confusion matrix pairs showed that there was high consistency in the way people correctly and incorrectly recognize emotional expressions across visibility (Figure 10; all p’s < .001). This shows the pattern of confusion across emotions was largely identical across levels of visibility.

**General Discussion**

We examined the robustness of recognition of emotional expressions by quantitatively degrading dynamic emotional stimuli using three qualitatively different types of visual noise simulating occlusion, peripheral or unattended vision, and distant vision. In Experiment 1, with opaque grey bars we occluded part of images, which sometimes hid crucial inner facial features for the recognition of a particular expression (e.g., the mouth for happiness). Lower contrast noise in Experiment 2 blurred the contour of the facial features and that of the face, simulating peripheral vision or lower level of visual attention.
(Experiment 2). Low pass filtering in Experiment 3 blocked fine-grained information of the images from the viewers, which simulated emotion recognition at a distance (Experiment 3). Despite the differences between the visual effects of the three types of noise, the pattern of robustness was consistent across the experiments.

To begin with, observers showed remarkably high level of accuracy given some of the noise levels applied to the images (Figure 1). For instance, they recognized emotions better than chance, even under the most severe level of noise, except for fearful expressions in Experiments 1-3, disgusted in Experiment 2-3, and sad expressions in Experiment 2. All three types of visual noise deteriorated observers’ accuracy in judging emotion, to varying degrees depending on the emotion intended in the videos. The recognition of surprise was the most robust (i.e., the performance deteriorated more slowly as the visibility of the videos diminished, compared to the performance for the other emotions). This seems to result from the conspicuous emotion cues embedded in the headshot videos of surprise (i.e., sudden backwards movement of head and shoulders). The other emotions gave rise to subtler cues than surprise. Also, under all three types of low visibility, the recognition of sad and angry expressions was always less robust (i.e., rapidly deteriorated as the noise became stronger) than other expressions. This suggests that at least the types of noise used in the current investigation (i.e., evenly distributed visual occlusion, lower contrast, and low pass spatial filtering) can rapidly remove the emotion cues from angry or sad expressions, more so than from the other expressions, when the noise becomes stronger.

We found consistency in the patterns of emotion recognition across noise manipulations. This was perhaps because the three noise types used here largely spared low spatial frequency information, despite the crucial differences across them. This explanation is consistent with Schyns and Oliva (1999), where the observers heavily relied on low spatial frequency information (< 8 cycles/image) when categorizing emotional expressions of faces. However, a distinctive pattern of robustness across noise types stood out in the recognition of one emotion in the current data. Specifically, in Experiment 3 (low pass filtering), disgusted expressions were significantly more fragile to the noise (i.e., the performance deteriorated more rapidly as the stimuli degraded), compared to the results from Experiment 1 (partial
occlusion) and 2 (low contrast), suggesting that removal of high spatial frequency information more effectively interfered with the perception of disgust from dynamic facial expressions than occlusion or low pass noise did. This is consistent with previous findings that high spatial frequency information of inner facial features (e.g., wrinkles on/near the nose) is diagnostic for the recognition of disgust (M. L. Smith, Cottrell, Gosselin, & Schyns, 2005). This was the type of information that became obscure under low pass filtering, unlike under the other two noise types, which explains the larger fragility of disgust recognition in Experiment 3 than in Experiment 1 and 2.

As for the confusability of emotions, in all three experiments, the observers showed similar types of correct and incorrect emotion judgments across the visibility levels. In other words, the general decrease in the emotion recognition accuracy did not affect the pattern of observers' misjudgment in different visibility levels regardless of the type of visual noise added. This suggests a similar mechanism that differentiates emotional expressions under various levels of noise.

To sum, we superimposed three types of noise with ecological value on dynamic emotional expressions. Across the experiments, the recognition of emotional expressions degraded in a similar pattern under various types of low visibility. Specifically, across all types of noise, the recognition of surprised expressions was the most robust, and the recognition of angry expressions was the least robust. However, low pass filtering prevented the recognition of disgust more severely than the other types of noise, which suggests different types of diagnostic information within each emotional expression. The results also showed a consistent pattern of emotion confusion across various levels of visibility, regardless of the noise type. This implies an identical emotion differentiating mechanism at work under various visibility levels. By and large, as far as the robustness of emotion recognition is concerned, it seems the commonality trumps the minor exceptions.
References


Footnotes

1. No effects were discovered in reaction times to the stimuli. We expected this as 1) we did not ask the observers to respond as quickly as possible and 2) the response key number (1 - 6) of emotion categories randomly changed over trials, which potentially increased the inter-trial variance of response times for each emotion.
Figures

A. Occlusion

B. Low contrast

C. Lowpass filtering

Figure 1: Three types of noise manipulations used in Experiments 1-3.
Figure 2: The emotion recognition performance in Experiment 1 as a function of emotion and visibility. The dotted line denotes chance level performance.
Figure 3: Mean slopes of the linear model of recognition in Experiment 1 as a function of emotion.

The smaller the value is, the more robust the recognition of the emotion is (**: $p < .003$).
Figure 4: Correlations between confusion matrices in Experiment 1 (all $p$’s < .003). The rows and columns indicate the visibility levels of stimuli, i.e., the occlusion proportions. Each smaller matrix is the confusion matrix at each occlusion proportion. The rows and columns of each small matrix indicate the actual emotion categories and the emotion category responses, respectively, i.e., angry, disgusted, fearful, happy, sad, and surprised.
Figure 5: The emotion recognition performance in Experiment 2 as a function of emotion and visibility. The dotted line denotes the chance level performance.
Figure 6: Mean slopes of the linear model of recognition in Experiment 2 as a function of emotion.

The smaller the value is, the more robust the recognition of the emotion is (**: p < .003).
Figure 7: Correlation among confusion matrices in different noise levels in Experiment 2 (all p’s < .003). The rows and columns indicate the visibility levels of stimuli, i.e., the contrast levels. Each smaller matrix is the confusion matrix at each contrast level. The rows and columns of each small matrix indicate the actual emotion categories and the emotion category responses, respectively, i.e., angry, disgusted, fearful, happy, sad, and surprised.
Figure 8: The emotion recognition performance in Experiment 3 as a function of emotion and visibility. The dotted line denotes the chance level performance.
Figure 9: Mean absolute values of the slopes of the linear model of recognition in Experiment 3 as a function of emotion. The smaller the value is, the more robust the recognition of the emotion is (**: p < .003).
Figure 10: Correlations among confusion matrices in Experiment 3 (all $p$'s < .003). The rows and columns indicate the visibility levels of stimuli, i.e., the cutoff levels of the low pass spatial frequency filters (deg/cycle). Each smaller matrix is the confusion matrix at each cutoff level. The rows and columns of each small matrix indicate the actual emotion categories and the emotion category responses, respectively, i.e., angry, disgusted, fearful, happy, sad, and surprised.