Running head: FEED-FORWARD CATEGORIZATION OF BODY EXPRESSIONS

A computational feed-forward model predicts categorization of masked emotional body language for longer, but not for shorter latencies

Bernard M.C. Stienen[1], Konrad Schindler[2], Beatrice de Gelder[1]

[1]Laboratory of Cognitive and Affective Neuroscience, Tilburg University, Tilburg, The Netherlands

[2]Photogrammetry and Remote Sensing, Insititute of Geodesy and Photogrammetry, ETH Zürich, Switzerland

Corresponding author:	Beatrice de Gelder
Address:	Cognitive and affective neuroscience lab
	Tilburg University
	PO Box 90153
	5000 LE Tilburg, The Netherlands
Tel. no.:	+ 31 13 4663203/fax + 31 13 4662370
E-mail addresses:	B.deGelder@uvt.nl

Abstract

Given the presence of massive feedback loops in brain networks, it is difficult to disentangle the contribution of feed-forward and feedback processing to the recognition of visual stimuli, in this case, of emotional body expressions. The aim of the present work is to shed light on how well feed-forward processing explains rapid categorization of this important class of stimuli. By means of parametric masking it may be possible to control the contribution of feedback activity in human participants. A close comparison is presented between human recognition performance and the performance of a computational neural model which exclusively modeled feed-forward processing and was engineered to fulfill the computational requirements of recognition. Results show that the longer the SOA (Stimulus Onset Asynchrony) the closer the performance of the human participants was to the values predicted by the model, with an optimum at an SOA of 100 ms. At short SOA latencies the human performance deteriorated, but the categorization of the emotional expressions was still above baseline. The data suggest that, although theoretically feedback arising from infero-temporal cortex is likely to be blocked when the SOA is 100 ms, human participants still seem to rely on more local visual feedback processing to equal the model's performance.

A computational feed-forward model predicts categorization of masked emotional body language for longer, but not for shorter latencies

Humans are capable of categorizing extremely quickly - and accurately - a wide variety of natural visual stimuli. Recent evidence suggests that this capability may be due to a fast feed-forward processing stream involving brain networks specialized in certain types of stimuli (Fabre-Thorpe, Delorme, Marlot, & Thorpe, 2001). The aim of the present work is to shed some light on how well feed-forward processing explains rapid processing of an important class of stimuli, namely human body postures conveying emotion. To this end we compare a computational model of feed-forward categorization to a behavioral experiment in which the available processing time was carefully limited.

In previous decades a number of research reports have focused on the processing of faces and their expressions in order to explore how we process emotions, and a number of computational models have been offered. More recently researchers have started to investigate the issue of bodily expression recognition. Switching to a new category can potentially provide evidence that human emotion theories may generalize to affective signals other than facial expressions (de Gelder, 2006, 2009). Results from a number of behavioural experiments using independent stimulus sets now allow us to conclude that recognition of emotions is similarly easy for face and body stimuli. Available literature has already firmly established that emotional bodily expressions clearly and rapidly convey the emotional, intentional and mental state of a person (Meeren, van Heijnsbergen, & de Gelder, 2005; Stekelenburg & de Gelder, 2004; Van den Stock et al, 2011) and that full awareness of the visual stimulus or intact striate visual cortex

are not essential (de Gelder, Vroomen, & Weiskrantz, 1999; Stienen, & de Gelder, 2011; Tamietto et al., 2009; Tamietto & de Gelder, 2010).

Schindler, Van Gool and de Gelder (2008) have shown that a computational neural model which modeled exclusively feed-forward processes was capable of categorizing a set of seven different emotional bodily expressions in much the same way as human observers did. However, there was no time limit on the presentation of the bodily expressions in the human categorization task. Given the presence of massive feedback loops in brain networks, it is unclear whether human performance was only based on feedforward processes with no contribution from feedback processes. By controlling the contribution of feedback in human participants a closer comparison between the brain networks and the assumptions of the model is possible.

Masking is one of the most widely used techniques for exploring unconscious processing of visual information in neurologically intact observers, and seems an excellent technique to control the contribution of feedback processes. For example, Esteves and Öhman (1993) found that short duration (e.g. 33 ms) presentations of happy and angry facial expressions, replaced immediately by a neutral face (mask) with a longer duration (e.g. 50 ms), are below the participants' identification threshold.

Lamme and Roelfsema (2000) and Lamme (2006) argue that a visual stimulus activates the visual cortex (striate and extrastriate) between 40 and 80 ms after presentation. Next, the infero-temporal cortex (IT) is feedforward-activated starting from 80 ms. Feedback signals from this area re-enter the visual cortex. Assuming 1 to 3 nodes that separate IT and visual cortex and a maximum firing rate of 100 Hz for cortical neurons (Rennie, Wright, & Robinson, 2000) the signal re-enters the visual cortex between 90-110 ms after the onset of the target. This means that

a mask could interfere with re-entrant signals arising from IT when presented less than 110 ms after presentation. In other words, it is increasingly more likely that feedback is possible from the infero-temporal cortex when the SOA (Stimulus Onset Asynchrony), and thus the processing time for the target, increases.

Neurological evidence indicates that masking selectively disrupts re-entrant signals to V1. For example, Lamme, Zipser and Spekreijse (2002) showed that masking seemed to selectively interrupt the recurrent interactions between V1 and higher visual areas in the macaque monkey brain. Fahrenfort, Scholte and Lamme (2007) found in a human EEG study that when a texture-defined square was masked with an SOA of 16 ms, ERP's typically associated with re-entrant processes were absent. No differences in bilateral occipito-temporal areas were found before 110 milliseconds while more posterior ERP's triggered by seen stimuli started to differ from those triggered by unseen stimuli.

However, the nature of the masking effect still remains a matter of discussion. The masking effect could be a consequence of imprecise temporal resolution starting as early as the retina, but possibly at cortical levels as well. This is called 'integration masking'. Alternatively, the masking effect could arise by interruption of target processing in higher areas involved in object recognition, or in this case, bodily expression recognition (see e.g. review by Enns & Di Lollo, 2000).

In our study we presented participants with masked emotional bodily expressions, using a parametric masking procedure to disentangle the contributions of feedback processing to their categorization performance. Five emotional expressions (including neutral) were presented to the participants while the onset between target and mask (SOA, Stimulus Onset Asynchrony) was

parametrically varied between 33 and 133 ms. The participants were instructed to categorize the emotion and use their intuition whenever they could not clearly see the target stimulus. The same set of stimuli was cross-validated using the neural model designed by Schindler et al (2008) and the outcomes were compared. In addition, the neural model was tested on mixtures (linear combinations) between the targets and the mask, in order to explore how the model performs on degraded images.

It is expected that up to an SOA of 100 ms feedback processes arising from IT would be blocked by the mask. According to theory, full feedback should be possible when the SOA is 133 ms or longer. If human participants can categorize bodily expressions in the absence of information carried by feedback processes, then the model should predict the human performance when SOA latencies are 100 ms or shorter.

Method

Masking Study

Participants

Twenty-two undergraduates of the University of Tilburg participated in exchange for course credits or a monetary reward (12 women, 10 men, M = 21.6 years, SD = 3.2). All participants had normal or corrected-to-normal vision and gave informed consent according to the declaration of Helsinki.

Stimuli and procedure

The same photoset was used as in the previous study by Schindler et al. (2008). However, in the present study only angry, fearful, happy, sad and neutral bodily expressions were used**,** while the expressions surprised and disgusted were left out. The faces were covered with an opaque gray mask. It was decided to use five categories instead of seven for pragmatic reasons. Firstly, we did not want to make the button-pressing too complicated, and secondly we aimed to keep the experiment within reasonable time limits. The reason for our selection of emotions was that "surprise" and especially "disgust" have a clear facial expression but no obvious bodily expression. Neutral bodily postures of 6 actors were used to construct a mask. A picture of a male and a female with an average posture were chosen as the basis. Using Adobe Photoshop 7.0 © these actors were fused together. Arms and legs from the four other identities expressing a neutral emotion were attached to the body at different positions and orientations creating the image of two bodies with more arms and legs than usual (see Figure 1). Average height of the bodies was 8.83 degrees; the average width was 3.41 degrees (distance to the screen was 90 cm). The height of the mask was 10.40 degrees; the width was 6.27 degrees covering the area where the target stimuli were presented completely.

The stimuli were presented on a 17" PC screen with the refresh rate set to 60 Hz. We used Presentation 11.0 to run the experiment. A white cross of 1.22 x 1.22 degrees was used as a fixation mark in the center of the screen. Finally, all stimuli were pasted on a gray background.

Participants were comfortably seated in a chair in a soundproof experimental chamber. A trial started with the white fixation cross on a gray background. The disappearance of this cross signaled the beginning of a trial. After 500 milliseconds the target stimulus appeared for 33 milliseconds. After a variable interval the mask was presented for 50 milliseconds. The SOA latencies were 33, 67, 100 and 133 milliseconds. The actual presentation time was calibrated

with the use of a photodiode and an oscilloscope measuring the latency between onset of the target and the onset of the mask. Moreover a target-only condition and a mask-only condition were included. After the categorization response a fixation cross appeared until the trial time was 3000 milliseconds.

Participants were instructed to categorize the target bodily expressions as angry, fearful, happy, sad or neutral. They responded with two hands using the ring, middle and index finger of the left hand and the index and middle finger of the right hand. The response buttons were labeled with the letter corresponding to the category and a reminder with the full names was situated on a board in front of them underneath the monitor. There were 5 between-subject counterbalance schemes making sure that each label occurred on every position once. They were instructed to be as accurate as possible but that the time for responding is short so they had to respond fast and to use their “gut feeling” if they had not seen the body.

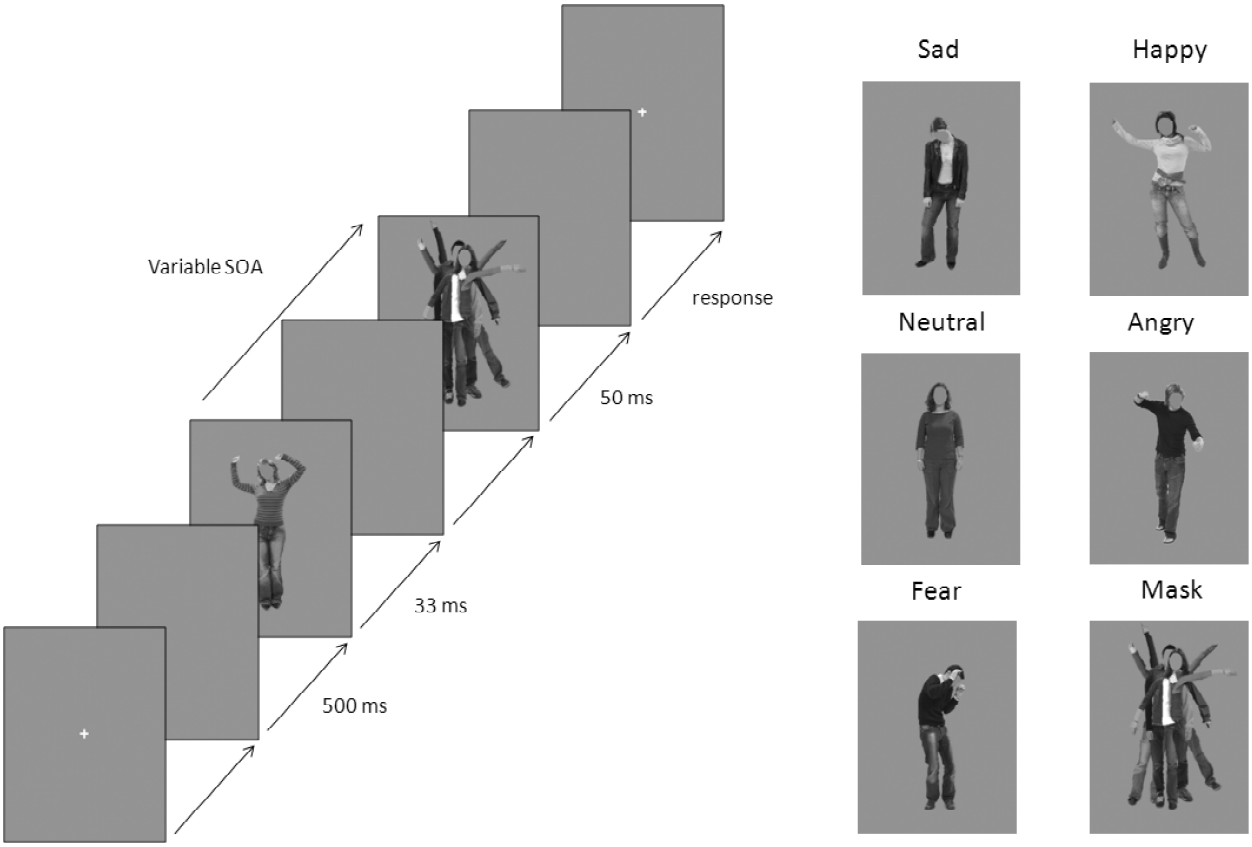

Figure 1. An example trial (Left). A typical example of each stimulus category (right).

Prior to the experimental sessions the participants performed two practice sessions consisting of 60 trials each. Other identities than the ones used in the main experiment served as targets. When the participants did not miss trials and gave notice of a full understanding of the procedures the main experiment was started. One complete run summed up to a total of 1230 trials (41 identities x 5 postures (4 emotions + neutral) x 6 timing conditions (including target-only and mask-only)) which were randomly presented. Every 160 trials there was a break. After the main experiment all targets were presented for 33 milliseconds to validate the stimuli. The instructions remained the same for this session. The experiment lasted 2 hours in total.

Neural model

The computational model has been inspired by the ones of Riesenhuber and Poggio (1999) and Serre, Oliva, and Poggio (2007). It consists of a four-layer feed-forward hierarchy: each processing layer converts the inputs from the previous layer to a set of output features of higher complexity and/or larger receptive field. The input to the bottom layer is the raw image, whereas the output of the top layer is a score for each of the possible categories. A schematic illustration is given in Figure 2. For further details please refer to Schindler, Van Gool and de Gelder (2008). The model was used without modification, thus the only difference to the original work is that in the present study the model categorized only five different expressions (four emotional bodily expressions and one neutral body pose) rather than seven.

To test for the possibility that the processing of the mask interfered with the early stages of processing the bodily expressions, which may be the case if integration masking occurs, we tested the neural model with degraded stimuli, created through pixel-wise linear combinations of the targets and the mask. We created three different stimulus sets by choosing different weight ratios between the target and the mask:

1) 0.8 x target + 0.2 x mask (Mix_1)

2) 0.5 x target + 0.5 x mask (Mix_2)

3) 0.2 x target + 0.8 x mask (Mix_3)

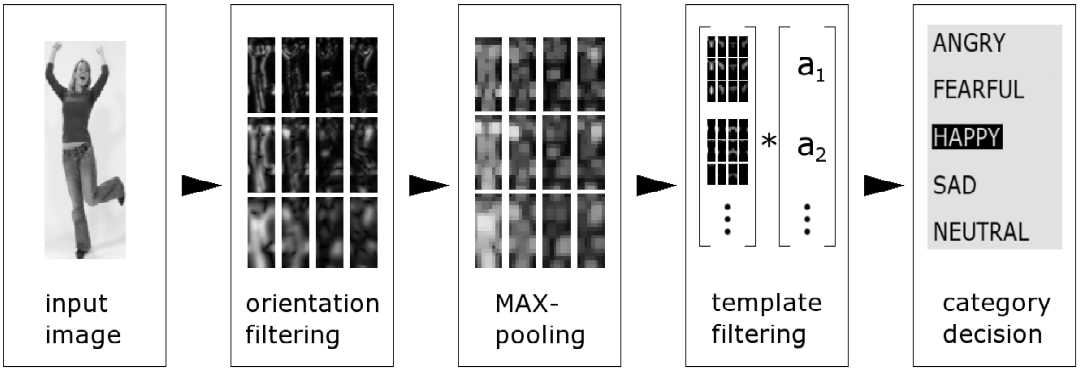

Figure 2. The computational model. From the raw image, local orientations are extracted at multiple scales, pooled over spatial neighborhoods, and compared to learned complex feature templates. The similarities with all complex features are fed into a discriminative (forced-choice) classifier. *Parameters were chosen for illustration purposes and are different from the actual implementation.*

Results

Trials where participants failed to categorize the bodily expression within the duration of the trial were discarded (0.4 percent of all trials, SD = 0.6). One participant was discarded as an outlier in the validation session. While the group was on average 91.3 percent (SD = 4.7) correct in categorizing the body postures, this participant was more than 3 standard deviations below that average. The validation scores for angry, fearful, happy, sad and neutral expressions were 81.8 (SD = 10.6), 94.5 (SD = 6.4), 97.5 (SD = 2.7), 84.6 (SD = 8.0) and 98.3 (SD = 2.3) percent correct respectively.

To calculate Chi-square distances between the observed human performance and the performance of the model we used the basic definition $\chi^2 = \sum ((Fo-Fe)^2/Fe)$ where *Fo* is the

observed correctly categorized stimuli per emotional category and *Fe* is the performance of the neural model per emotional category. The Chi-square distance was computed separately for each participant and the distances were averaged, see Figure 3. When the SOA was 100, 133 milliseconds or when no mask was presented, the model predicted the human performance significantly well (resp. $\chi^2(4, N = 22) = 7.25$, $p > .05$; $\chi^2(4, N = 22) = 4.52$, $p > .05$; $\chi^2(4, N = 22) = 3.49$, $p > .05$).

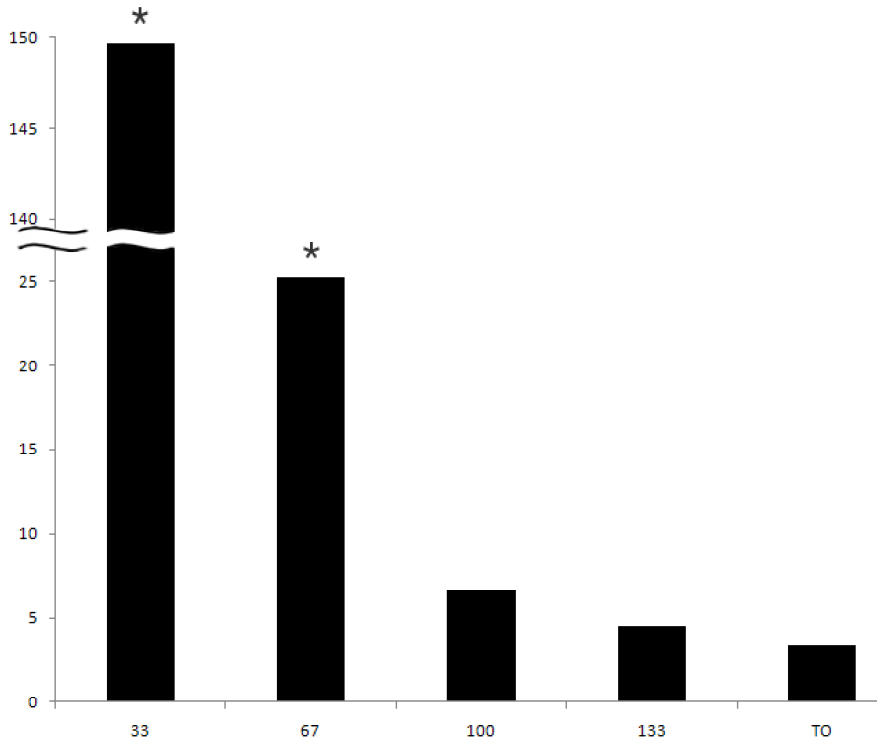

Figure 3. Chi-square distances between neural model performance and human performance per SOA condition. TO = Target-Only.

As shown in Figure 4, while not reaching the performance predicted by the model, participants still performed above baseline for the expressions fearful, happy and sad when the SOA was 33 milliseconds (all $p < .05$), and when the SOA was 67 milliseconds the participants categorized all expressions above baseline ($t(20) = 2.81$, $p < .05$).

To gain further insight which among the higher SOA conditions matched the model best we analyzed the common misclassifications between model and human participants. We counted a stimulus as misclassified when the number of correct answers was more than 1 standard deviation below average per SOA condition or, in the case of the model, below average performance. Since each unique stimulus was only shown once per SOA to the participants**,** the number of correct classifications were indexed on the group level. Next, we indexed how many stimuli were misclassified by both the human participants and the model. Figure 5 shows that the longer the SOA the smaller the number of misclassifications. Interestingly, the total common misclassifications by model and by the humans increases until the SOA is 100 ms and decreases again when the SOA is longer.

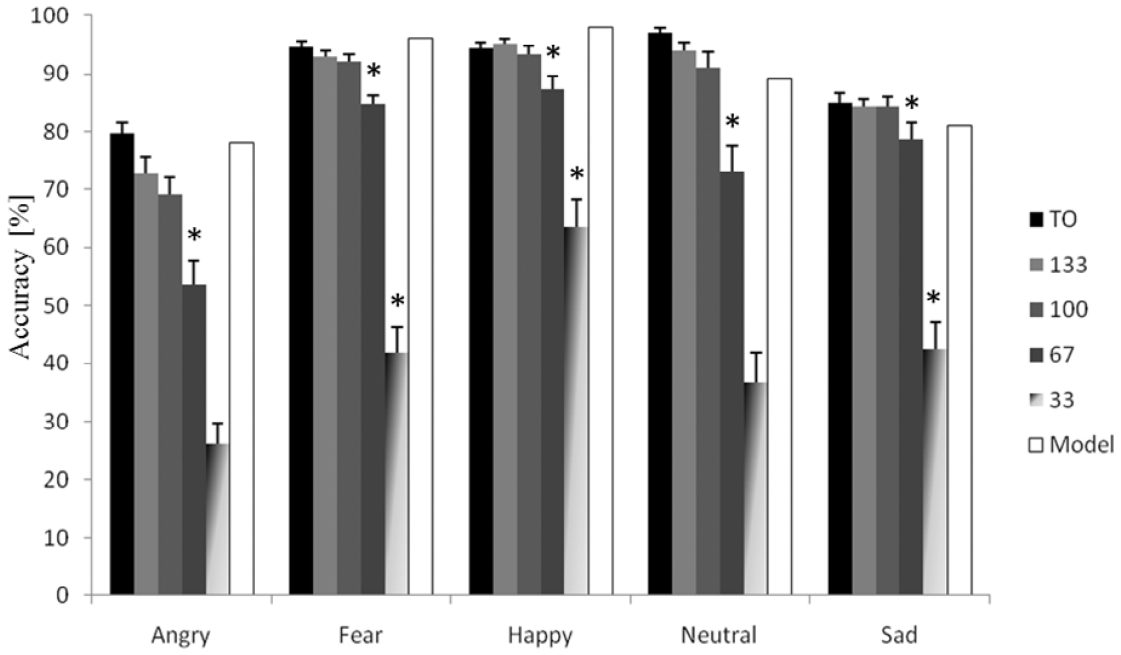

Figure 4. Accuracy rates in percent per emotion category per SOA condition. Stars indicate performance above baseline. Error bars indicate standard error mean. TO = target-only.

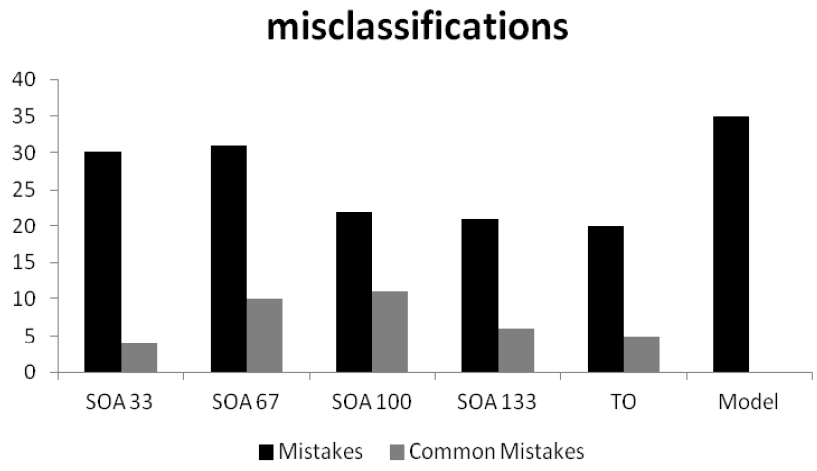

Figure 5. Total number of mistakes by human participants per SOA, number of mistakes made by the computational model, and common mistakes by both the model and human participants per SOA.

The raw confusion matrices were also analyzed. Table 1 shows an overview of the confusions that were observed in the model (respectively, the human participants). As can be seen, the higher the SOA the more the model seems to predict the actual human behavior. In Table 2 the absolute differences between predicted and observed values are shown. Chi-square tests were not performed on these data because not all assumptions were met, e.g. not all cell values were larger than 5. The major differences between the model and the human participants was that the humans mostly confused angry with neutral, while the model confused angry dominantly with sad. When no mask was presented the human participants, contrary to the model, did not confuse neutral with sad.

Model	Angry	Fear	Happy	Neutral	Sad
Angry	78	2	0	0	3
Fear	3	96	2	0	2
Happy	4	0	98	0	0
Neutral	0	0	0	89	13
Sad	15	2	0	11	81

SOA 100 ms	Angry	Fear	Happy	Neutral	Sad
Angry	69	3	2	1	1
Fear	4	92	2	1	2
Happy	11	2	93	3	1
Neutral	10	1	2	91	12
Sad	6	3	1	4	85

SOA 33 ms	Angry	Fear	Happy	Neutral	Sad
Angry	26	12	10	11	7
Fear	8	42	6	7	8
Happy	31	22	64	34	22
Neutral	23	13	13	37	20
Sad	11	12	8	12	43

SOA 133 ms	Angry	Fear	Happy	Neutral	Sad
Angry	73	3	2	1	1
Fear	3	93	2	1	2
Happy	8	2	95	2	1
Neutral	11	1	1	94	11
Sad	6	2	0	3	84

SOA 67 ms	Angry	Fear	Happy	Neutral	Sad
Angry	54	4	4	5	3
Fear	6	85	4	2	3
Happy	18	4	87	12	4
Neutral	14	2	3	73	11
Sad	8	5	2	8	79

Target-Only	Angry	Fear	Happy	Neutral	Sad
Angry	80	2	2	1	1
Fear	3	95	2	0	2
Happy	4	1	94	1	1
Neutral	9	0	1	97	12
Sad	4	2	1	1	85

Table 1. Confusion matrices for the model (*with border*) and the human participants. Columns represent true emotion; rows represent the reported emotion (in percent). The cells are grayscale-coded using the logarithm of the percentages.

	Angry	Fear	Happy	Neutral	Sad
SOA 33 ms					
Angry		9.53	9.54	10.82	4.11
Fear	5.50		3.92	6.87	5.71
Happy	27.50	21.62		33.74	22.32
Neutral	22.70	12.57	13.27		7.27
Sad	3.80	10.33	7.67	0.88	
Total	59.50	54.04	34.39	52.31	39.42
SOA 67 ms					
Angry		2.24	3.64	5.37	0.03
Fear	2.70		2.49	2.10	1.05
Happy	14.41	3.78		11.77	3.98
Neutral	13.78	2.00	2.83		1.77
Sad	6.58	3.19	1.66	3.29	
Total	37.47	11.22	10.62	22.54	6.83
SOA 100 ms					
Angry		0.58	2.01	1.05	2.41
Fear	1.00		0.22	0.94	0.34
Happy	6.57	1.53		3.18	0.82
Neutral	10.22	0.70	2.02		1.26
Sad	8.90	1.07	0.83	7.27	
Total	26.68	3.88	5.07	12.43	4.84

	Angry	Fear	Happy	Neutral	Sad
SOA 133 ms					
Angry		0.71	1.53	0.82	2.06
Fear	0.04		0.35	0.71	0.34
Happy	3.71	1.77		1.76	0.70
Neutral	10.91	0.58	1.19		2.42
Sad	9.25	0.11	0.47	8.31	
Total	23.91	3.17	3.55	11.59	5.52
Target-Only					
Angry		0.11	2.23	0.59	2.17
Fear	0.41		0.24	0.24	0.36
Happy	0.22	0.82		1.18	0.93
Neutral	9.43	0.24	1.18		1.43
Sad	10.98	0.47	0.58	10.06	
Total	21.04	1.65	4.24	12.07	4.89

Table 2. Absolute differences between model and human performance per timing condition. Cells colored black indicate a difference between expected and observed value greater than 2 standard deviations from the average.

Figure 6a shows the averaged Chi-square distances between the results of the model when using the degraded mix_1, mix_2 and mix_3 stimuli and the results of the human participants. Figure 6b shows the actual human performance per emotion per SOA, the original performance of the model**,** and its categorization performance for the mix_2 and mix_3 images. The longer the latency, the more the model deviates from human performance for the mix_3 images. Interestingly, the results for mix 3 angry and sad postures seems to better match those of humans at SOA 33ms, while this is not the case for the remaining categories. However, in all

cases the responses the model returned for the mix_3 images were significantly different from human performance (all $p < .001$).

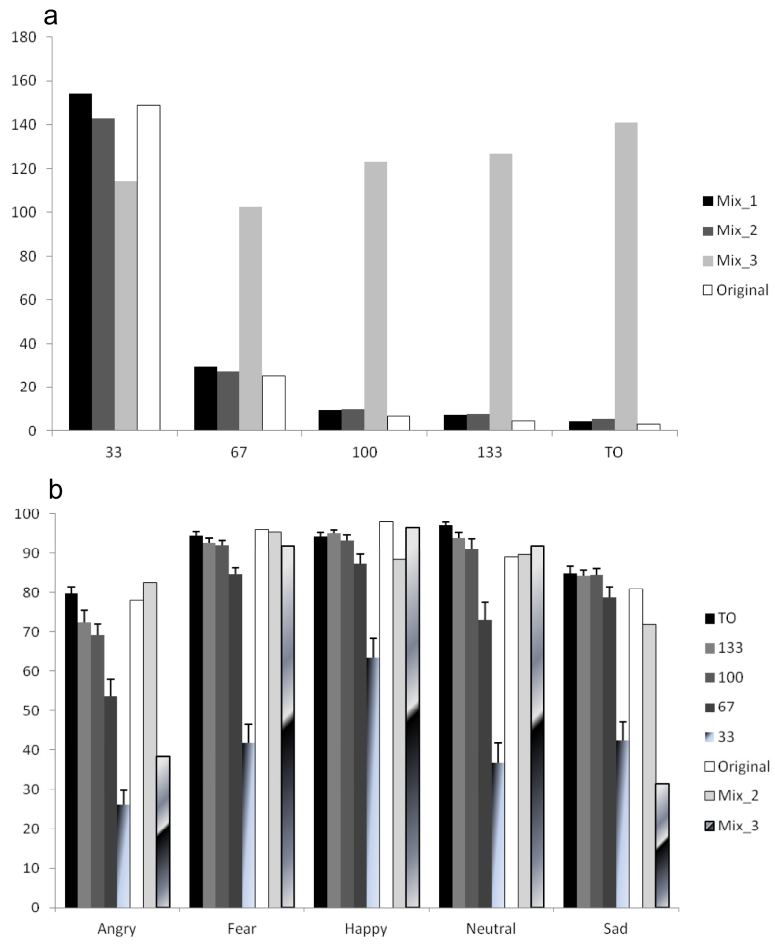

Figure 6. a) Chi-square distances per SOA condition between the human performance and the results of the neural model when classifying the original, mix_1, mix_2 and mix_3 stimuli. b) Accuracy rates in percent per emotion category per SOA latency of the human participants (TO, 133, 100, 67, and 33) and model performance (Original, Mix_2, and Mix_3). Error bars indicate standard error mean. TO = target-only.

Finally, a 5 (emotions) x 5 (SOA latency) multivariate analyses of variance (MANOVA) showed that there was a main effect of emotion ($F(4,16) = 18.49$, $p < .001$) and SOA ($F(4,16) = 28.28$, $p < .001$) on the reaction times. Bonferroni corrected multiple comparisons show that angry bodily expressions are categorized slower than to the other bodily expressions. All SOA conditions differed from each other significantly with the exception when the SOA was 100 and 133 ms. The general trend is that the longer the SOA latency, the shorter the reaction time.

Discussion

We have shown that a feed-forward computational model predicts the human categorization performance for emotional body language strikingly well. The longer the SOA the closer the performance of the human participants matched the performance of the model, with an optimum at an SOA of 100 ms. However, while on short SOA latencies the human categorization performance deteriorated, it was still above baseline. When testing the computational model with targets that had been degraded by mixing them with the mask, its performance also decreased, but was still different from the human participants.

Based on the theoretical framework proposed by Lamme and Roelfsema (2000) one would expect that the performance of the feed-forward neural model equals the performance of the humans at SOA conditions up to 100 ms. Yet human participants are capable to perform the task better than chance when the SOA is low but their performance is much worse than the neural model.

There are four possible explanations for this observation. Firstly, the model works in a context free environment and, other than human participants, is not distracted by the environment**,** for example by the processing of the mask itself. As an alternative, it would be interesting if, as proposed by Lamme (2006), one were able to block re-entrant processing associated with bodily expressions with TMS as being done by Jolij and Lamme (2005) with schematic faces. This method loads the visual system less with distracting visual information. Then one could the compare these results with the performance of a neural model as described here.

Secondly, it may be the case that the target and mask temporally overlap on the retinal level, interfering with the processing of the bodily expressions at an early stage which would fit the view of masking by integration (Enns & Di Lollo, 2000). We showed that although the computational model performs much worse when tested on targets degraded by overlaying the mask, its performance was still different from the one of humans. However, while that experiment gives some insights, there are multiple ways to represent integration between two images on a retinal or cortical level. This multiple solution problem limits the interpretation of our current results. In addition, biases may be present in the computational model, because contrary to the human visual system, the model learns only from stimuli similar to those being tested, while it lacks exposure to the large amount of images human participants are exposed to throughout their lives. In addition, humans categorize bodily expressions viewed in complex contexts.

Thirdly, when the SOA is 67 ms it might just happen to be close to the average required time of the feed-forward mechanism, such that we would be observing a mixture of successful categorizations and random answers.

Fourthly, when the SOA is 100 ms local feedback processing of the target might occur whereas at shorter SOA's, these local feedback processes would be impaired. For example, there could be a distinction between recurrent activation originating from V3 and recurrent activation originating from IT. If we assume that V1 is activated 40 ms after target onset, then V3 (via V2 or not) is activated 50-60 ms after target onset. If we further assume that V3 feeds back directly to V1 then the re-entrant signal arrives there 60-70 ms after target onset. At these latencies the mask is already activating V1 when the SOA is 33 or 67 ms. In conclusion, when the SOA is 100 ms, feedback processes arising from IT are most likely to be disrupted, while shorter SOA's could also interrupt more local feedback from extrastriate areas. This has important implications. For example, could it be that the conscious visual percept is disrupted at an SOA of 33 ms, while at an SOA of 100 ms the human participants are conscious about the visual percept, but nevertheless categorize the bodies automatically?

Pascual-Leone and Walsh (2001) showed that applying TMS to V1 after stimulating V5 in a time window of 5-45 ms led to a decrease in reporting that the TMS induced posphemes moved. In addition, a study of Koivisto, Railo, Revonsuo, Vanni, and Salminen-Vaparanta (2011) showed that recurrent interactions between ventral areas and V1/V2 are necessary for categorization and perception of natural scenes. They found longer response times and degraded quality of subjective perception when applying single pulse TMS in the time window 90-210 ms to V1/V2 and longer response times when applying single pulse TMS to LO after 150 ms and longer. Jolij and Lamme (2005) found that when stimulating V1 110 ms after onset of a display with four smileys, participants had difficulties reporting the location, but not the emotion. It seems that feedback to V1 is necessary for visual awareness. These studies suggest that the

processing of a given visual stimulus around 100 ms in V1 is crucial for conscious perception and to make perceptual decisions, possibly because recurrent activation is required.

Finally, there is the possibility that another less accurate mechanism is aiding the participants to classify the emotions. It is well known that subcortical structures play a role in visual perception. When the SOA was 33 ms, three out of the four emotional body expressions (happy, fearful and sad) were recognized above baseline. This result could be hinting at a subcortico-cortical pathway. When visual signals are prevented from being processed by the cortical mechanisms via the striate cortex, the colliculo-thalamo-amygdala pathway could still process them. This is in line with recent fMRI studies that have observed differential amygdala responses to fear faces as compared to neutral faces when the participants were not aware (Morris et al., 1999; Whalen et al., 1998). However, this study lacks the additional measurement of e.g. subjective awareness to be conclusive on this topic (see e.g. Cheesman & Merikle, 1986).

Caution must be exercised before concluding that the categorization performance when the SOA was 33 ms reflects unconscious processing. While Esteves and Öhman (1993) found that an SOA of 33 ms rendered an emotional face invisible this is not found in this study. Stimulus specific properties in masking studies are known to modulate the sensitivity of the masking effect. For a thorough review see Wiens (2006). It could be that the arms formed a higher contrast against the background when there was no overlapping with the arms of the mask, thus causing the above baseline performance. Further research is needed on this issue.

Our data indicate that the computational model and the human participants confused more or less to the same degree sad bodily expressions with neutral ones. The major difference between the model and the human performance in terms of confusion is the observation that the

model tends to categorize angry as sad, whereas the human participants interpret angry poses as neutral. Some of the actors in the stimulus set expressed anger by a "controlled anger" pose, crossing their arms and tilting the head. The model tends to interpret these deviating poses as being sad, while the human participants interpreted them as being neutral, possibly because they were attentionally biased towards the body and not the head (see Schindler et al. (2008) for more example stimuli). This raises the possibility that the model might not be a sufficiently good proxy for the human recognition process because it lacks an attention mechanism.

The fact that performance does not change a lot when the SOA's are 100 ms or longer deserves special attention. Assuming that the perceptual decision is made in V1, feedback from IT might be blocked by the mask when the SOA is 100 ms. The fact that there are no major performance changes when the processing time of the target increases and feedback from parietal-frontal areas becomes possible suggests that in these kind of tasks participants do not rely on feedback coming from higher areas. The only change was that there were fewer common mistakes between model and humans and that the confusion pattern changed slightly when no mask was presented.

To summarize, the feed-forward neural model predicts human behavior strikingly well although the model slightly outperforms the human participants. According to our study it is likely that emotional bodily expressions can be recognized even when feedback from higher-level areas is blocked, although humans might still rely on some form of local feedback processing (while the model does not).

References

Cheesman, J., & Merikle, P. M. (1986). Distinguishing conscious from unconscious perceptual processes. *Canadian Journal of Psychology, 40*(4), 343-367.

de Gelder, B. (2006). Towards the neurobiology of emotional body language. *Nature Reviews Neuroscience, 7*(3), 242-249.

de Gelder, B. (2009). Why bodies? Twelve reasons for including bodily expressions in affective neuroscience. *Philosophical Transactions of the Royal Society B: Biological Sciences, 364*(1535), 3475-3484.

de Gelder, B., Vroomen, J., Pourtois, G., & Weiskrantz, L. (1999). Non-conscious recognition of affect in the absence of striate cortex. *Neuroreport, 10*(18), 3759-3763.

Enns, J. T., & Di Lollo, V. (2000). What's new in visual masking? *Trends in Cognitive Sciences, 4*(9), 345-352.

Esteves, F., & Ohman, A. (1993). Masking the face: recognition of emotional facial expressions as a function of the parameters of backward masking. *Scandinavian Journal of Psychology, 34*(1), 1-18.

Fabre-Thorpe, M., Delorme, A., Marlot, C., & Thorpe, S. (2001). A limit to the speed of processing in ultra-rapid visual categorization of novel natural scenes. *Journal of Cognitive Neuroscience, 13*(2), 171-180.

Fahrenfort, J. J., Scholte, H. S., & Lamme, V. A. (2007). Masking disrupts reentrant processing in human visual cortex. *Journal of Cognitive Neuroscience, 19*(9), 1488-1497.

Jolij, J., & Lamme, V. A. (2005). Repression of unconscious information by conscious processing: evidence from affective blindsight induced by transcranial magnetic

stimulation. *Proceedings of the National Academy of Sciences of the United States of America, 102*(30), 10747-10751.

Koivisto, M., Railo, H., Revonsuo, A., Vanni, S., & Salminen-Vaparanta, N. (2011). Recurrent processing in V1/V2 contributes to categorization of natural scenes. *Journal of Neuroscience, 31*(7), 2488-2492.

Lamme, V. A. (2006). Zap! Magnetic tricks on conscious and unconscious vision. *Trends in Cognitive Sciences, 10*(5), 193-195.

Lamme, V. A., & Roelfsema, P. R. (2000). The distinct modes of vision offered by feedforward and recurrent processing. *Trends in Neurosciences, 23*(11), 571-579.

Lamme, V. A., Zipser, K., & Spekreijse, H. (2002). Masking interrupts figure-ground signals in V1. *Journal of Cognitive Neuroscience, 14*(7), 1044-1053.

Meeren, H. K., van Heijnsbergen, C. C., & de Gelder, B. (2005). Rapid perceptual integration of facial expression and emotional body language. *Proceedings of the National Academy of Sciences USA, 102*(45), 16518-16523.

Morris, J. S., Ohman, A., & Dolan, R. J. (1999). A subcortical pathway to the right amygdala mediating "unseen" fear. *Proceedings of the National Academy of Sciences of the USA, 96*(4), 1680-1685.

Pascual-Leone, A., & Walsh, V. (2001). Fast backprojections from the motion to the primary visual area necessary for visual awareness. *Science*, 292(5516), 510-512.

Rennie, C. J., Wright, J. J., & Robinson, P. A. (2000). Mechanisms of cortical electrical activity and emergence of gamma rhythm. *Journal of Theoretical Biology*, 205(1), 17-35.

Riesenhuber, M., & Poggio, T. (1999). Hierarchical models of object recognition in cortex. *Nature Neuroscience, 2*(11), 1019-1025.

Schindler, K., Van Gool, L., & de Gelder, B. (2008). Recognizing emotions expressed by body pose: a biologically inspired neural model. *Neural Networks, 21*(9), 1238-1246.

Serre, T., Oliva, A., & Poggio, T. (2007). A feedforward architecture accounts for rapid categorization. *Proceedings of the National Academy of Sciences of the USA, 104*(15), 6424-6429.

Stekelenburg, J. J., & de Gelder, B. (2004). The neural correlates of perceiving human bodies: an ERP study on the body-inversion effect. *Neuroreport, 15*(5), 777-780.

Stienen, B. M. C., & de Gelder, B. (2011). Fear detection and visual awareness in perceiving bodily expressions *Emotion*, *11*(5), 1182-1189.

Tamietto, M., Castelli, L., Vighetti, S., Perozzo, P., Geminiani, G., Weiskrantz, L., et al. (2009). Unseen facial and bodily expressions trigger fast emotional reactions. *Proceedings of the National Academy of Sciences of the USA, 106*(42), 17661-17666.

Tamietto, M., & de Gelder, B. (2010). Neural bases of the non-conscious perception of emotional signals. *Nature Reviews Neuroscience, 11*(10), 697-709.

Van den Stock, J., Tamietto, M., Sorger, B., Pichon, S., Grezes, J., & de Gelder, B. (2011). Sortico-subcortical visual, somatosensory, and motor activations for perceiving dynamic whole-body emotional expressions with and without striate cortex (V1). *Proceedings of the National Academy of Sciences of the USA, 108*(39), 16188-16193.

Whalen, P. J., Rauch, S. L., Etcoff, N. L., McInerney, S. C., Lee, M. B., & Jenike, M. A. (1998). Masked presentations of emotional facial expressions modulate amygdala activity without explicit knowledge. *Journal of Neuroscience*, 18(1), 411-418.

Wiens, S. (2006). Current concerns in visual masking. *Emotion, 6*(4), 675-680.