Vision Research 211 (2023) 108275

Contents lists available at ScienceDirect

Vision Research

journal homepage: www.elsevier.com/locate/visres

What surprises the Mona Lisa? The relative importance of the eyes and eyebrows for detecting surprise in briefly presented face stimuli

Emil Skog [1], C. Stella Qian [1], Anisha Parmar, Andrew J. Schofield *

School of Psychology, College of Health and Life Sciences, Aston University, Birmingham B4 7ET, United Kingdom

ARTICLE INFO

Keywords:
Classification images
Face expression
Emotion
Reverse correlation
Face perception
Noise

ABSTRACT

The classification image (CI) technique has been used to derive templates for judgements of facial emotion and reveal which facial features inform specific emotional judgements. For example, this method has been used to show that detecting an up- or down-turned mouth is a primary strategy for discriminating happy versus sad expressions. We explored the detection of surprise using CIs, expecting widened eyes, raised eyebrows, and open mouths to be dominant features. We briefly presented a photograph of a female face with a neutral expression embedded in random visual noise, which modulated the appearance of the face on a trial-by-trial basis. In separate sessions, we showed this face with or without eyebrows to test the importance of the raised eyebrow element of surprise. Noise samples were aggregated into CIs based on participant responses. Results show that the eye-region was most informative for detecting surprise. Unless attention was specifically directed to the mouth, we found no effects in the mouth region. The eye effect was stronger when the eyebrows were absent, but the eyebrow region was not itself informative and people did not infer eyebrows when they were missing. A follow-up study was conducted in which participants rated the emotional valence of the neutral images combined with their associated CIs. This verified that CIs for 'surprise' convey surprised expressions, while also showing that CIs for 'not surprise' convey disgust. We conclude that the eye-region is important for the detection of surprise.

1. Introduction

Studies of face perception often seek to understand how information about internal mental states is conveyed through facial expressions. This impressive perceptual ability operates with accuracy despite the large diversity found in human faces and the subtlety of emotional expressions. In this study, we explored the detection of surprise as a facial expression using a classification images (CIs) technique called reverse correlation.

Surprise is unusual among the six basic emotions. While the expression of most emotions can be prolonged, surprise is necessarily a fleeting emotion typically expressed only until the source of the surprise is identified and the appropriate emotion takes over (Ekman, 2003). Further, whereas most emotions are clearly positive or negative, surprise can be both Ekman (2003). Similarly, while many emotions have a clear opposite (for example, happy and sad), the opposite of surprise is less well defined. Given that surprises can be both positive and negative, the opposite of surprise might be neutral; indeed, the antonyms of surprise (unimpressed, disinterested, indifferent, etc.) suggest neutrality. Woodworth and Schlosberg (1954) characterise surprise as pleasant and attention orienting and place it opposite disgust which is seen as unpleasant and rejecting. However, Lövheim's (2012) physiological model suggests that surprise is associated with high serotonin, low dopamine and high noradrenaline and might thus be regarded as the opposite of fear (low serotonin, high dopamine, low noradrenaline). In contrast, Ekman (2003) notes that the facial emotions fear and surprise are hard to distinguish. In terms of facial features, the startle response (eyes closed, brows lowered, lips closed and stretched) is the opposite of the surprised expression (eyes wide open, brows raised, mouth open and circular) despite similarities in the causes of these too reactions (Ekman, Friesen, & Simons, 1985). In the light of the above, we wanted to explore the facial features that convey surprise as a fleeting emotion and also explore the opposite emotion of surprise.

The CI method is a useful tool for studying facial emotion as it can be used with short presentation times to reveal critical features for the detection of an emotion and simultaneously reveal the template for the

* Corresponding author.
E-mail address: a.schofield@aston.ac.uk (A.J. Schofield).

[1] ES and CSQ are recognised as joint first authors.

https://doi.org/10.1016/j.visres.2023.108275
Received 14 September 2022; Received in revised form 30 May 2023; Accepted 2 June 2023
Available online 8 July 2023

opposite or anti-emotion. In this method, a typically neutral base stimulus is embedded in visual noise with a new noise sample presented on every trial. Participants are asked to classify each stimulus into one of many categories and the noise samples associated with each response are accumulated. Features in the noise that lead to consistent classifications will be reinforced revealing the internal template for each response class. CIs have thus been used to uncover the diagnostic features that underlie the detection or categorisation of emotional expressions (Gosselin & Schyns, 2003; Jack, Caldara & Schyns, 2012; Kontsevich & Tyler, 2004), gender (Mangini & Biederman, 2004), species (Martin-Malivel et al., 2006), trustworthiness, and dominance (Brinkman, Todorov & Dotsch, 2017; Dotsch & Todorov, 2012).

Kontsevich and Tyler (2004) showed that adding noise to an image of the Mona Lisa painting (da Vinci, c. 1503–1506) can bias her face to appear as if it is expressing an emotion, such as happiness or sadness. The Mona Lisa has a close to neutral expression, but is famous for 'almost smiling', with a facial expression that sits somewhere at the category boundaries of happy/neutral. Kontsevich and Tyler (2004) recorded responses as 'sad' and 'happy' (a binary discrimination task) with an additional confidence level applied to each response. They then constructed CIs from the responses where participants were confident. When all the noise samples that drove confident responses for 'sad' and 'happy' were added back onto the Mona Lisa, the templates altered her facial expression to contain a down- or up-turned mouth, respectively.

Jack et al. (2012) tasked observers with an emotion categorisation task involving the emotions happy, surprise, fear, disgust, anger and sad (a sixfold discrimination task), in race-, gender- and emotion-neutral faces with added noise. Their study further incorporated an examination of cultural differences between Western Caucasian and East Asian participants, finding CIs which contained culturally differing templates for the above emotions. To discriminate surprise, Western Caucasian participants used the eyebrow, mouth, and eye regions whereas East Asian participants used only the eyes. Smith et al. (2005) used a related classification images technique ('Bubbles'; Gosselin & Schyns, 2001) on expressive, non-neutral, face stimuli expressing the above-mentioned emotions plus a neutral expression. They found that different spatial frequency information and different parts of faces carry diagnostic information regarding different emotional expressions. Critically, surprise was primarily classified via the opened mouth, but high-spatial frequency information in the eyes and eyebrows was also involved. Similarly, Blais et al. (2012) used the bubbles technique to show that the mouth region is more informative than the eye region for emotion categorisation and Blais et al. (2017) showed that the eye regions are informative for surprise in dynamic but not static stimuli. Finally, Blais et al. (2017) also measured fixation durations while observers categorised emotions in static and dynamic faces finding that they dwelt longest on the nose region for both kinds of stimuli but dwelt longer on the eyes and mouth for static vs dynamic stimuli.

The discrimination of surprise is known to rely on the mouth, eyes, and eyebrows, (Beaudry et al., 2014; Blais et al., 2012, 2017; Calder et al., 2000; Jack et al., 2012; Smith et al., 2005). The eyebrows and mouth are particularly critical for distinguishing fear from surprise (Roy-Charland et al., 2015), with the eye regions presenting very similar information for these two emotions. It is not clear however if information from the mouth and eyebrows is necessary to merely detect surprise. Observers might detect surprise (or a combination of surprise and fear) from the eyes alone. Nonetheless (Jack et al.'s, 2012 result for East Asian participants notwithstanding) the eyebrows and mouth would seem to be important for judgments of surprise. We thus expected to find features in our CIs clustering around the eye, eyebrow, and mouth regions. Further, we reasoned that the eyebrows might be so important to the detection of surprise that participants would interpolate their presence in an arched configuration even when the eyebrows were absent from the image. This aspect of our study was inspired by Gosselin and Schyns (2003) who explored the detection of an illusory smiling mouth, where no mouth target was presented in a face image. Gosselin and Schyns found that when their observers were asked to detect a smiling mouth, they inferred a mouth-shaped template which correlated more strongly with the shape of a happy mouth than a neutral or angry mouth. However, we did not delete the mouth in any of our conditions.

The above studies used discrimination tasks where the emotion of interest is pitted against the opposite emotion (Kontsevich and Tyler, 2004) or a set of alterative emotions (Jack et al., 2012). As we are interested in surprise and its opposite emotion, which is less well defined, a detection experiment whereby the participant is asked to say how reminiscent of surprise the stimulus is rather than discriminate it from other emotion(s) is more appropriate. Dotsch & Todorov (2012) used a detection-like task to explore trustworthiness and dominance. Here participants were asked to say (for example) which of two identical, neutral face stimuli embedded in different noise samples was most trustworthy, with one noise sample being the negative of the other. The positive and negative noise samples were then aggregated to form positive and negative CIs. While this is a strictly discrimination experiment, the opposite trait was not made explicit – participants were not asked if stimuli were trustworthy versus untrustworthy. Indeed, participants were separately asked to judge trustworthiness and untrustworthiness (likewise submissiveness and dominance). As it happened the negative CI for each trait in a pair closely resembled the positive-CI for the opposite trait. However, to our knowledge, no one has ever studied templates for the opposite of surprise in an experiment where no alternative emotion was specified. We thus adopted a detection paradigm in our study so as to better expose the template for anti-surprise.

The current study thus seeks to extend our knowledge of the perception of surprise through CIs, using an emotion detection task on neutral faces presented for a limited duration of 500 ms. Limiting presentation time provided ecological validity as surprise expressions are fleeting, with brief time-courses (Ekman, 2003). CI templates from 'surprise' responses were kept separate from 'not surprise' response templates. This brings the benefit of allowing exploration of what expression is conveyed by the 'not surprise' templates, with the explicit aim of investigating 'anti-surprise' expressiveness. Furthermore, positive- and negative-response templates may not necessarily be negatives of one another (Eckstein, Shimozaki & Abbey, 2002; Eckstein, Pham & Shimozaki, 2004; Kontsevich & Tyler, 2004; Murray, 2011; Tjan & Nandy, 2006), suggesting that they should not necessarily be combined as is typical in many CI studies (e.g., Ahumada, 1996; Beard & Ahumada, 1998; Gold et al., 2000; Gold, Sekuler & Bennett, 2004; Murray, Bennett & Sekuler, 2005).

In our first experiment, we used the CI method to examine how templates may emerge from classifying surprise in neutral faces modulated by noise. We hypothesized, based on previous findings, that the eyes, eyebrows, and mouth may be critical diagnostic features for the classification of surprise (Beaudry et al., 2014; Calder et al., 2000; Jack et al., 2012; Smith et al., 2005). We further hypothesised that participants would interpolate raised eyebrows into images where none were present. To this end we used three base images. The primary stimulus was a face with a neutral expression sourced from the Karolinska Directed Emotional Faces (KDEF) database (Lundqvist, Flykt & Öhman, 1998). This face was used in a second condition in which the eyebrows were edited out. We also included the Mona Lisa as the third face stimulus, exploring whether this ambiguous face without distinct eyebrows would produce similar results to the KDEF face without eyebrows. Our results suggest that the eye-region is more informative than the eyebrows for the detection of surprise in briefly presented stimuli. This is accentuated when the eyebrows are missing, but the eyebrows themselves are not reliably interpolated when absent.

Our second experiment provided confirmation of the results from the first experiment and extends those results to a larger and more diverse sample of participants. Here we added the resulting 'surprise' CIs from the first experiment back onto the original faces to examine the degree to which they convey the intended emotion and generalise beyond the original participant group. We also examined the emotion conveyed by the CI for 'not surprised' responses from experiment 1, that is, the template for anti-surprise. The results of our second experiment suggest that our 'surprise CIs' indeed promote the perception of surprise and that the opposite of surprise is disgust.

2. Experiment 1

2.1. *Method*

2.1.1. *Stimuli*

A static, 128 × 128 pixel, white noise image subtending 4 × 4 degrees of visual angle was generated on each trial. Three grey-scaled faces (Fig. 1, upper row) were embedded in the noise images such that the noise accounted for 65%, and the faces 35% of the available luminance contrast. This contrast ratio was set during pilot evaluation by the authors, as it allowed for a balance between high visibility of the faces, while allowing the noise to alter the appearance of the faces. See Fig. 1 (lower row) for example stimulus images. Face images were scaled such that the face area (top of chin to forehead) was matched across the images. This area was smaller than the noise samples, subtending 2.233 degrees of visual angle in height. This size is relatively small for a face perception study (Yang, Shafai & Oruc, 2014) but faces were thus presented in parafoveal regions, reducing the need for saccades given our limited presentation time. For example, while Blais et al (2017) recorded on average around 2 fixations for face stimuli presented for 500 ms and 2 fixations are considered sufficient for face recognition (Hsiao & Cottrell, 2008), participants predominantly dwelt on the nose area. Our use of smaller stimuli allowed participants to capture more of the image in a single fixation at the cost of increasing the spatial frequency of all image features. The faces were converted to greyscale and standardized to have the same mean and root-mean-square contrast. The primary face (Fig. 1a, referred to here as 'Neutral Face') was 'Female 11′ from the KDEF database and had a neutral expression. This face was used in the second condition with the eyebrows edited out (Fig. 1b, 'Neutral Face Without Eyebrows'). The third image (Fig. 1c) was a digital image of the Mona Lisa.

2.1.2. *Participants and ethical considerations*

Seven participants (age range: 20–52, mean 30.4, *SD*: 10.1; three females, four males), including the four authors, were recruited. Six participants had lived in Europe throughout their lives: four of these were ethnically White Caucasian, one South Asian, and one Afro-Caribbean. One participant was born and raised in East Asia but had for lived many years in Western countries. Six of the participants had prior experience of participating in behavioural studies. Compensation of £8.33 per hour was awarded to the non-author participants. It took each participant 3 to 3.5 h to complete the experiment. Participant recruitment was based on, and limited by, participants' availability and ability to run the experiment from their own computer (see Procedure and equipment below).

Informed consent was obtained, and participants were assured that their data would be confidential and anonymised. The project was reviewed by Aston University's College of Health and Life Sciences Ethical Review Committee (approval number 856, v3 amended 2020).

2.1.3. *Procedure and equipment*

Participants conducted the experiment in their own homes but with detailed instructions. The experiment was constructed using PsychoPy (Peirce et al., 2019) and all participants had access to a standalone version of this software and ran the experiment on their local computer. We used the full Python implementation of PsychoPy not the online/JavaScript version. Participants noted their demographic information and details about their computer, including gender, age, handedness, computer make, operating system, monitor make, and details about their optical prescription (if applicable). All had normal or corrected to normal vision. Participants measured the height of the active screen

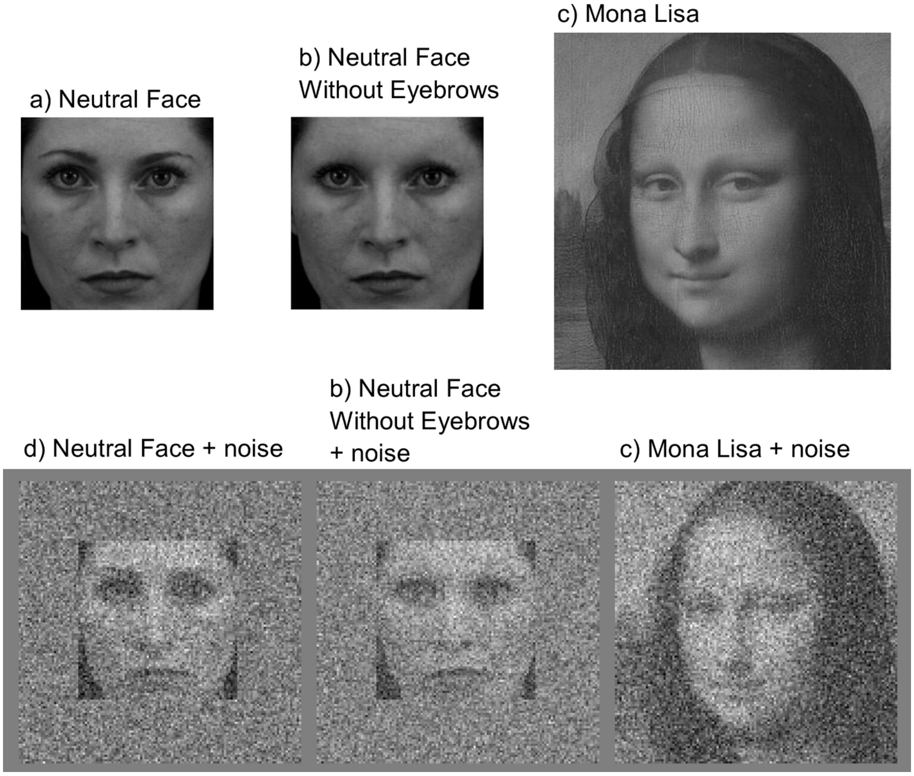
a) Neutral Face
b) Neutral Face Without Eyebrows
c) Mona Lisa
d) Neutral Face + noise
b) Neutral Face Without Eyebrows + noise
c) Mona Lisa + noise

Fig. 1. The three face stimuli used in this experiment, a) Neutral Face, b) Neutral Face Without Eyebrows, c) Mona Lisa. Example stimulus images from the three face conditions, d) Neutral Face, e) Neutral Face Without Eyebrows, f) Mona Lisa. In the experiment, the neutral faces were scaled to match the same vertical extent as the Mona Lisa face.

region of their monitors and based on this were told the viewing distance that they should maintain throughout the experiment. Images were scaled to the intended size by PsychoPy using linear interpolation where the number of pixels in the image no longer matched the number of screen pixels required to render the image at the correct size. As our participants used high resolution monitors this process resulted in an upscaling that will have retained the information in the noise samples. However, this process does introduce some additional variation into the resulting CIs that is not accounted for.

Participants were instructed to interleave the sessions according to a predetermined, counterbalanced, block design structure. Participants were encouraged to balance the sessions between morning and afternoon, across several days. Participants were free to listen to music. Prior to starting, the participants completed a practice run of 100 trials using a different neutral face (also sourced from the KDEF). Participants carried out 2,004 trials for each of the three face conditions across a total of 18 sessions. On each trial, participants were asked to answer the question: 'Did the face express surprise?'. Four responses were afforded: 'Unlikely', 'Less unlikely', 'Less likely', and 'Likely'. These responses were given using keyboard buttons '1′, '2′, '3′, and '4′, respectively. Each trial started with a 500 ms delay period, followed by a 500 ms presentation of the face image and noise mask. Participants then had unlimited time to respond. The next trial started once each response was made.

While measures of interindividual variability are considered desirable (Cone, Brown-Iannuzzi & Dotsch, 2021) the ability of an aggregate CI, as used in our study, to reveal features depends largely on the number of noise samples that are accumulated, averaging across participants, rather than the number of participants. Kontsevich and Tyler (2004) found significant CI features involving small numbers of pixels by presenting 12 observers with 100–120 trials each: a maximum of 1440 noise samples across the entire study. Jack et al. (2012) accumulated around 25,700 samples per CI. Our use of 2004 noise samples per participant in each condition resulted in 14,028 noise samples per full-CI. Thus, we should be able to reliably extract features at a similar level of granularity as Kontsevich and Tyler (2004) and Jack et al. (2012).

2.2. Results

2.2.1. Classification images

The white noise sample used in each trial was categorized based on

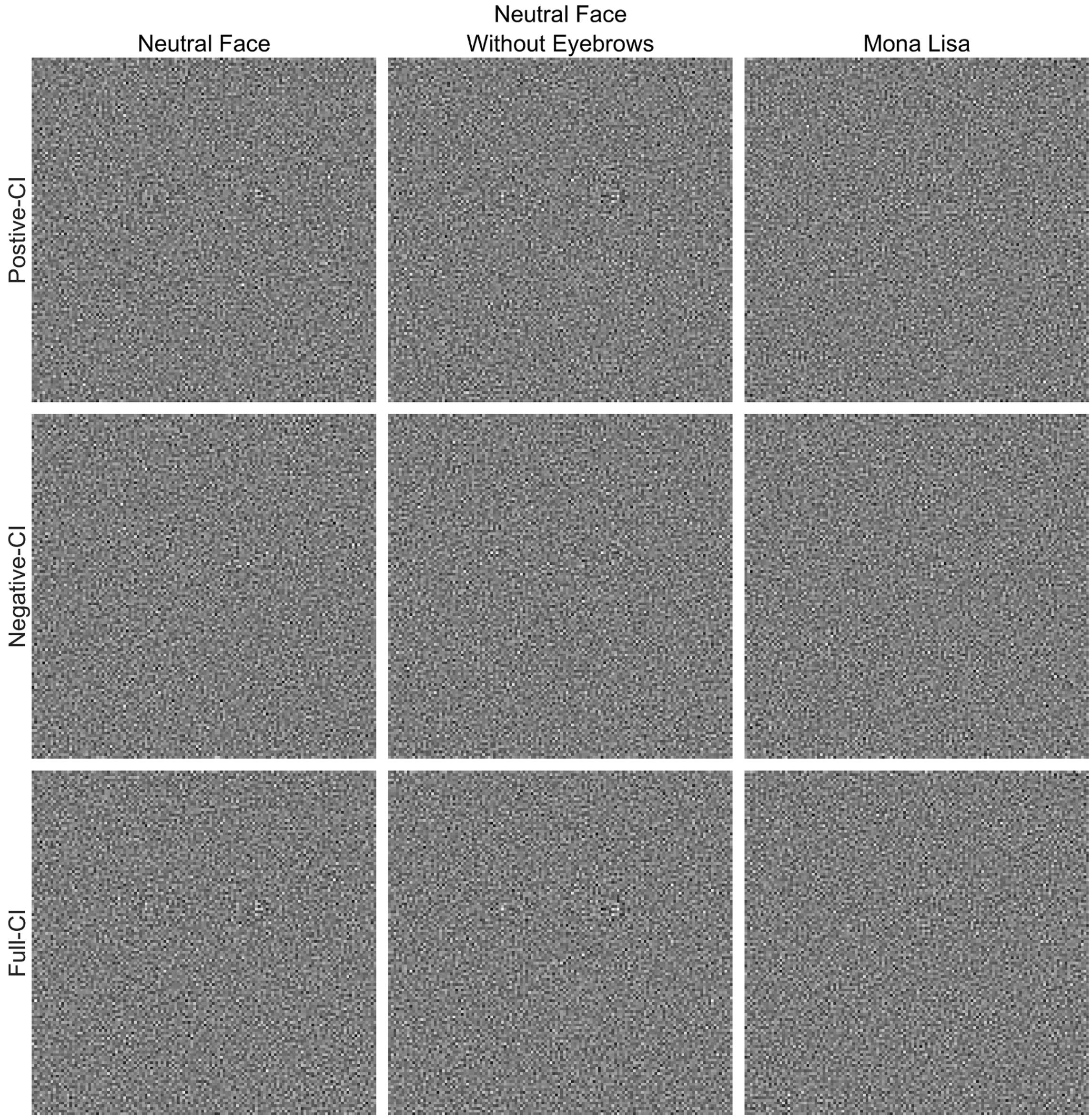

Fig. 2. Positive-CIs, Negative-CIs, and Full-CIs from the three conditions. Left column: Neutral Face. Middle column: Neutral Face Without Eyebrows. Right column: Mona Lisa. Eye-shaped templates emerge in some CIs, where the eyes were located in the underlying faces.

the response made. CIs for positive responses (Positive-CIs) were constructed by summing all noise samples that were associated with positive responses, Negative-CIs were produced by summing noise samples associated with negative responses. We observed that some participants were very hesitant to give confident responses (‘Likely’ or ‘Unlikely’), meaning that their CI containing only confident responses were weak. We therefore collapsed together the positive responses ‘Less likely’ and ‘Likely’ into a single category for all participants, the same was done for the negative responses ‘Less unlikely’ and ‘Unlikely’. Participant seven made no ’likely’ responses to the Mona Lisa image, so they were excluded from this condition. Full-CIs were derived by subtracting the Negative-CIs from the Positive-CIs. Finally, Positive-, Negative-, and Full-CIs were aggregated across participants and are shown in Fig. 2. Aggregated CIs added to their respective original faces are shown in Fig. 3. We observe differences between Positive-CI ‘surprise’ and Negative-CI ‘not surprise’ in how they alter the appearance of the original faces. The alterations are localized primarily in the eye region. An observed effect is an increase in contrast between the sclera and iris in the Positive-CIs, and a masking effect of the sclera in the Negative-CIs.

2.2.2. *Statistical analysis*

The CIs shown in Fig. 2 contained visibly evident structure around the eye regions. To see if this structure represented a significant deviation from random noise, and following Kontsevich and Tyler (2004), we first tested all possible image patches by running a 30 × 30 pixel sampling window across the whole CI with a step size of 2 pixels. Each image patch was tested against a normal distribution and against the distribution for a background region of the CI outside the face area using the generalised Kolmogorov-Smirnov test. Neither method was able to detect significant deviations from the test distribution when correcting for multiple comparisons. Similarly, tests of standard deviation, kurtosis, and skew comparing the same image patches to a background patch were not significant. From this we conclude that the clearly visible structure in the CIs is not revealed in pixel-wise image statistics.

Reasoning that the structure in the CIs is visible to the human visual system we decided to decompose them into orientation and spatial frequency bands similar to the channels that are known to exist in human vision (Blakemore & Campbell, 1969; Campbell & Robson, 1968; Sachs, Nachmias, & Robson, 1971). To this end we passed the CIs though a filter bank (see Fig. 4) comprising quadrature pair Gabor filters at five

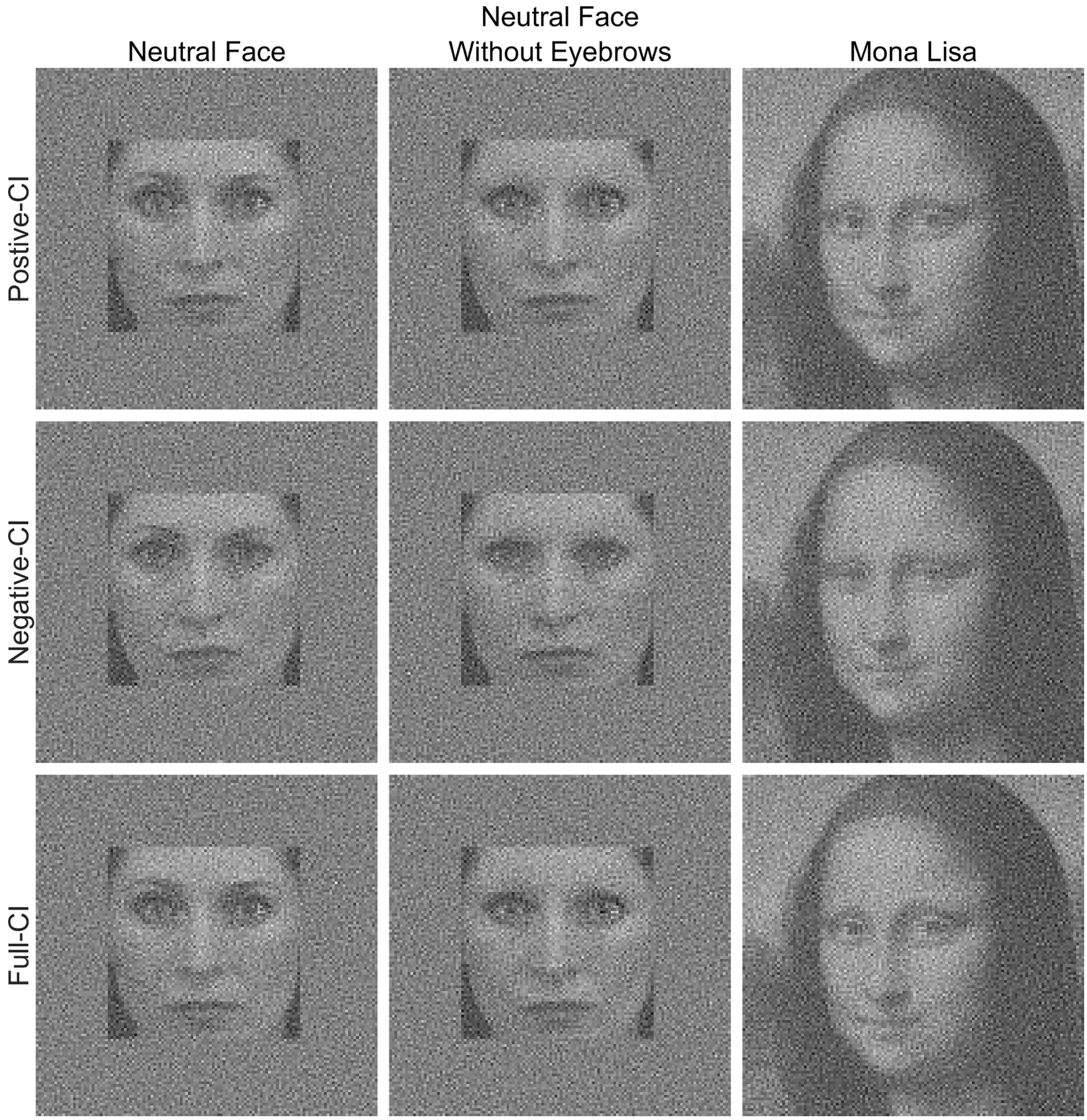

Fig. 3. Positive-CIs, Negative-CIs, and Full-CIs added back to their respective faces. Left column: Neutral Face. Middle column: Neutral Face Without Eyebrows. Right column: Mona Lisa.

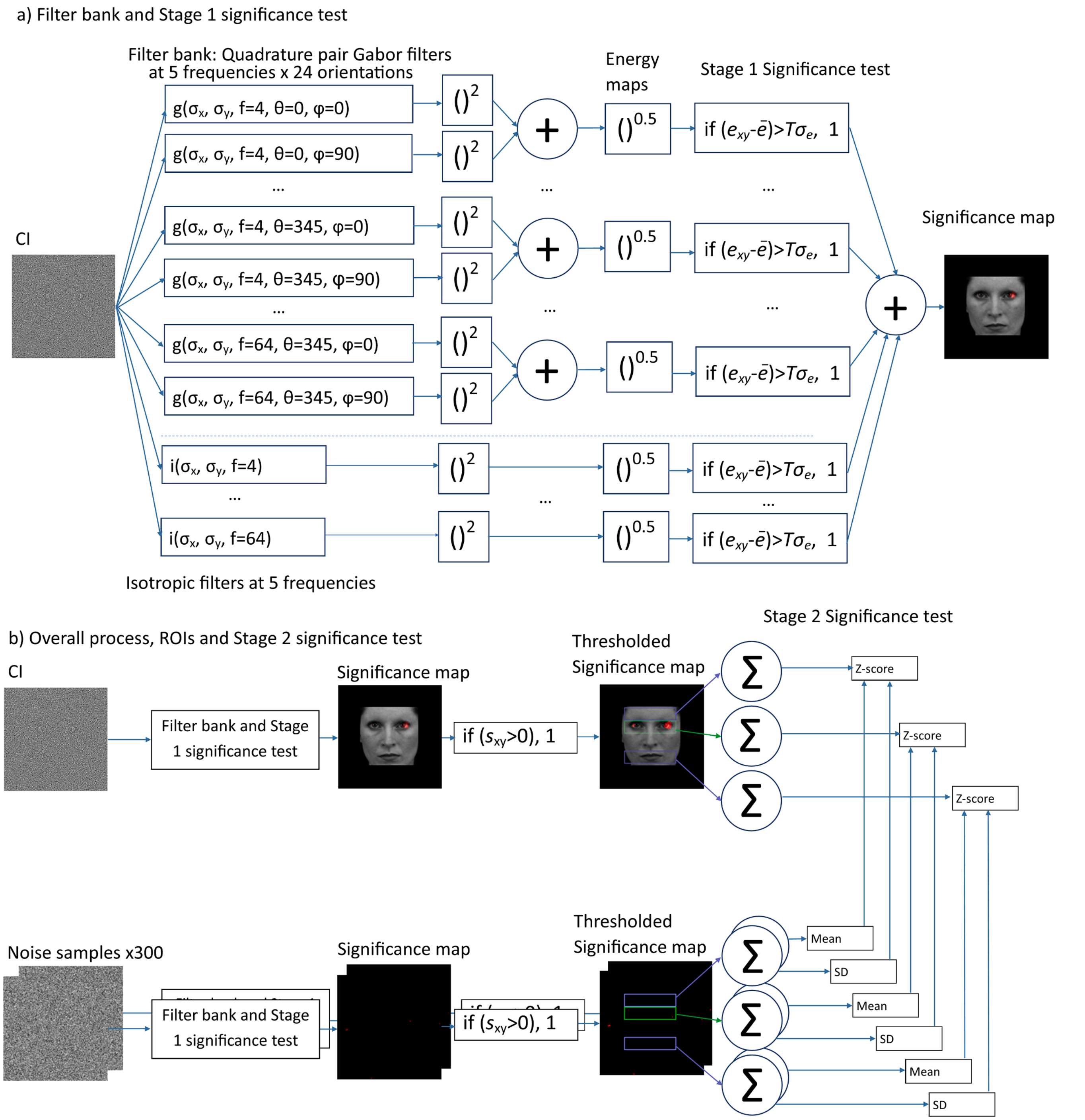
a) Filter bank and Stage 1 significance test
Filter bank: Quadrature pair Gabor filters at 5 frequencies x 24 orientations
Energy maps
Stage 1 Significance test
CI
g(σx, σy, f=4, θ=0, φ=0)
g(σx, σy, f=4, θ=0, φ=90)
g(σx, σy, f=4, θ=345, φ=0)
g(σx, σy, f=4, θ=345, φ=90)
g(σx, σy, f=64, θ=345, φ=0)
g(σx, σy, f=64, θ=345, φ=90)
()2
+
()0.5
if (exy-ē)>Tσe, 1
Significance map
i(σx, σy, f=4)
i(σx, σy, f=64)
Isotropic filters at 5 frequencies
b) Overall process, ROIs and Stage 2 significance test
CI
Filter bank and Stage 1 significance test
Significance map
if (sxy>0), 1
Thresholded Significance map
Stage 2 Significance test
Σ
Z-score
Noise samples x300
Mean
SD

Fig. 4. Schematic of the two-stage analysis method used to test significant deviations in our CIs. (a) Filter bank and Stage 1 significance test. Gabor filters: g(σ,f,θ,φ) = $\exp(-(x'^2/2\sigma_x^2 + y'^2/2\sigma_y^2)).\cos(2\pi f x' - \varphi)$, where $x'=x.\cos(\theta) + y.\sin(\theta)$ and $y'=-x.\sin(\theta) + y.\cos(\theta)$. Isotropic filters (i) where formed as the Fourier domain sum of all the cosine phase (φ = 0) Gabor filters for a given spatial frequency (f). The outputs of each quadrature phase Gabor filter pair are squared, summed and square rooted to derive energy maps with the outputs of the isotropic filters similarly rectified. The Stage 1 significance test then compared each pixel (e_{xy}) in each energy map to the mean thresholding at T = 5.4555 standard deviations above the mean. These were then summed to form significance maps (see Fig. 6). (b) Overall analysis and Stage 2 significance test. Significance maps were thresholded and the number of non-zero pixels counted within regions of interest and compared to the number produced from applying the overall analysis to random noise samples.

spatial frequencies (f = 4,8,16,32 and 64 c/image) and 24 orientations (0—360° in 15° steps - note sine phase filters are not orientation symmetric). The frequency and orientation bandwidths of the filters were 1 octave and 15° respectively which approximate channel bandwidths in human vision (Blakemore & Campbell, 1969; Thomas & Gille, 1979). This post-hoc decomposition of the CIs into orientation and frequency bands has some resonance with Mangini and Biederman's (2004) *a priori* composition of noise images from oriented components at different scales. Outputs for the two filters in each quadrature pair were squared, summed and square-rooted to derive energy maps. We applied 5 additional isotropic filters at the five spatial frequencies above rectifying their outputs to derive energy maps.

In order to test the statistical significance of deviations in our energy maps, we adopted a highly conservative two-stage significance testing procedure. We first tested individual pixels in the filtered energy maps for significant deviations from the background level and then tested the number of such pixels in regions of interest against that which would be expected for random noise samples.

Stage 1: We looked for pixels in each energy map that were more than 5.4555 standard deviations from the map's own mean value to form thresholded energy maps. This z-score threshold corresponds to an alpha of 0.05 Bonferroni corrected for the number of pixels in each filtered image and the number of filters applied – a very conservative measure. These were then summed across the filters (combined thresholded energy maps). This process was repeated for the Positive-, Negative- and Full-CIs for each face image and participant, and for the CIs aggregated across participants.

Stage 2: For the aggregated CIs only, we next counted the number of non-zero pixels in each combined threshold energy map and the total of such pixels in 16 × 64 pixel strips encompassing the eyebrow, eye and mouth regions respectively as shown in the bottom row of Fig. 5. We repeated this process for the aggregated, Positive-, Negative-, and Full-CIs (see Table 1). To assess the significance of our results we also estimated the distribution of above threshold pixels in random noise samples. For each face image and participant, we created 4 random noise images to represent the four response-typed CIs (e.g. the CI for 'likely surprised' responses etc). These were then combined into Positive-CIs, Negative-CIs, and Full-CIs as for our main analysis, pooled across the participants and analysed with the filter bank method outlined above. Counts of significant pixels were taken as in Stage 2 above. This process was repeated 100 times resulting in 300 estimates of significant pixels, 100 per face image. We then estimated the mean number of significant pixels and its associated standard deviation for all 300 images. This provided a baseline against which to test our CIs. Finally, we calculated z-scores for each CI or region thereof, comparing the number of pixels in the appropriate combined thresholded energy map to the means estimated from the random noise samples.

The $z = 5.4555$ threshold applied in Stage 1 above is extremely conservative. We therefore repeated the above analysis with a threshold of 4.5228 this corresponding to an alpha of 0.05 Bonferroni corrected for the number of pixels in the final significance map only. However, this threshold is still conservative and, being pixel based, does not account for smoothness in the filtered images. We therefore also looked for significant clusters using the Stat4Ci algorithm which is based on Random Field Theory (Chauvin et al., 2005) with two cluster forming thresholds. These results are consistent with our pixel-wise approach.

Finally, for consistency and completeness we applied the 2-stage analysis above to unfiltered images testing at Stage 1 to see if individual pixels values were significantly different from the overall mean. This analysis produced no significant pixels at Stage 1 when Bonferroni corrected for the number of pixels in the image. The number of significant pixels without correction was itself significantly higher than random samples in the eye region for the Full CI in all three images, and the mouth region of the Mona Lisa images, but the significance maps lacked obvious structure. Full details for, and the results of, these additional analyses are presented in the supplementary materials.

2.2.3. Significance maps and ROI results

Following the filter-bank and 2 stage significance testing described above, Fig. 5 shows the combined thresholded energy maps for the Full-

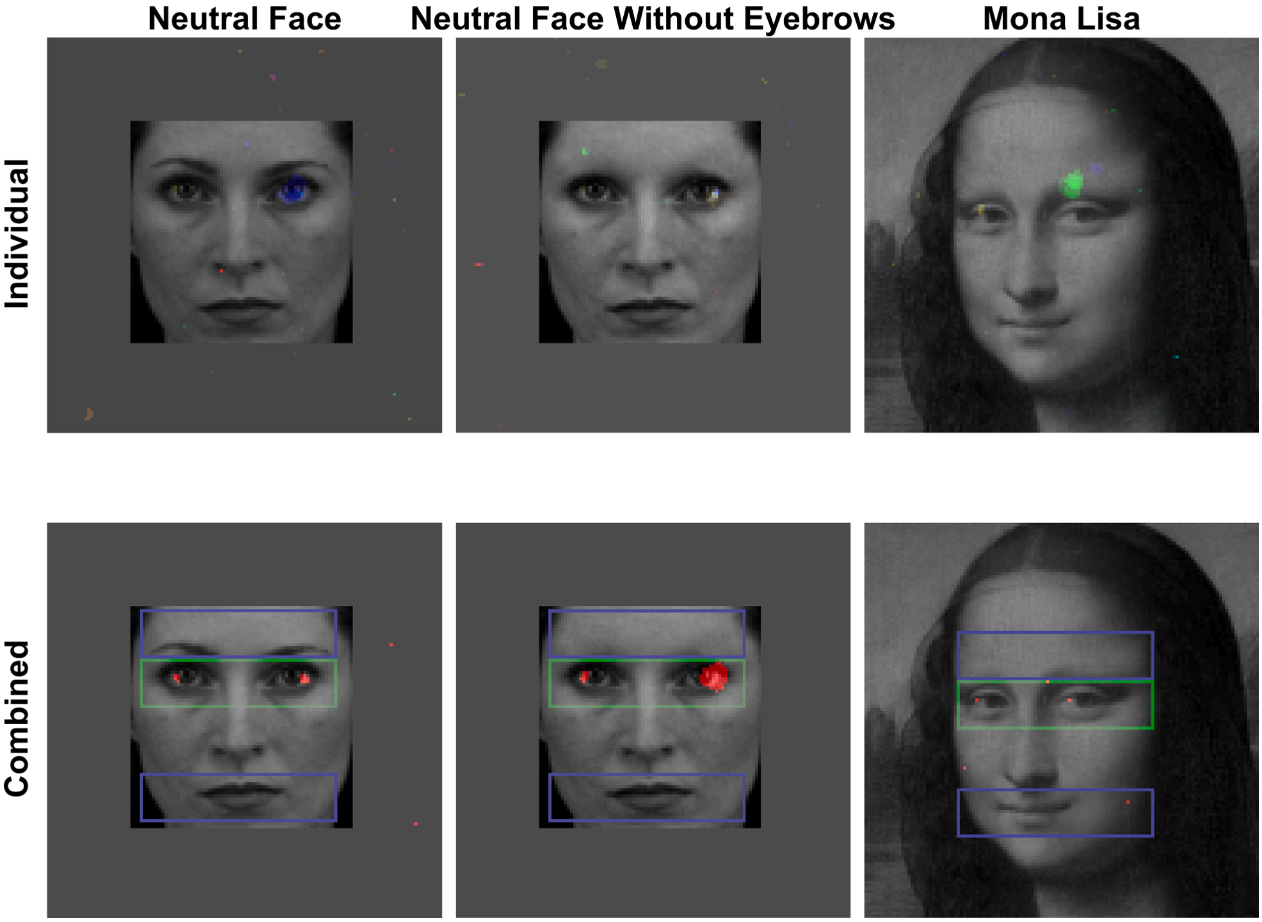

Fig. 5. Top row, significance maps based on the Full-CIs for individuals overlaid onto the face images. Colours (online) represent different participants. Bottom row, Thresholded significance maps for the Full-CI combined across participants. Light-green and dark-blue boxes in the bottom row show regions of interest (ROIs) for the Stage 2 analysis – see text.

Table 1

Statistical analysis of CIs. (a) The number of pixels (#p) above 5.4555 standard deviations from the mean combined across filter bands is shown for the Full-, Positive- and Negative-CIs aggregated across participants. Counts are shown for the whole image, and the eyebrow-, eye- and the mouth-regions as depicted by the blue and green boxes in Fig. 5. Z scores and significance levels (sig) are shown relative to (b) the mean and standard deviation of the number of above-threshold pixels in images derived from normally distributed random noise.

a)	Neutral Face			Neutral Face Without Eyebrows			Mona Lisa			b) Noise	
	#p	z	sig	#p	z	sig	#p	z	sig	mean	sd
Full CI											
Whole image	15	1.38	n.s.	71	11.1	<0.0001	5	−0.36	n.s.	7.08	5.76
Brows	0	−0.22	n.s.	0	−0.22	n.s.	0	−0.22	n.s.	0.52	2.4
Eyes	13	12.04	<0.0001	71	67.3	<0.0001	3	2.51	0.006	0.3633	1.05
Mouth	0	−0.41	n.s.	0	-0.41	n.s.	1	0.58	n.s.	0.4167	1.0
Positive CI											
Whole image	23	2.77	0.003	70	9.19	<0.0001	10	0.51	n.s.		
Brows	0	−0.22	n.s.	0	−0.22	n.s.	2	0.61	n.s.		
Eyes	20	18.71	<0.0001	60	56.82	<0.0001	4	3.46	0.0003		
Mouth	1	0.58	n.s.	0	−0.42	n.s.	0	−0.41	n.s.		
Negative CI											
Whole image	6	−0.19	n.s.	11	0.68	n.s.	0	−1.23	n.s.		
Brows	0	−0.22	n.s.	6	2.28	0.0114	0	−0.22	n.s.		
Eyes	0	−0.35	n.s.	4	3.46	0.0003	0	−0.35	n.s.		
Mouth	0	−0.42	n.s.	0	−0.42	n.s.	0	−0.42	n.s.		

CIs. The top row of Fig. 5 shows individual maps, while the bottom row shows the maps derived from the aggregated Full-CIs shown in Fig. 2. Table 1 shows the results of the Stage 2 analysis showing the number of active pixels in each thresholded significance map or section thereof as well as the mean and standard deviation of the number of such pixels found in maps derived from random noise samples. The number of above threshold pixels in the Full-CI map was significantly higher than that for noise for the Neutral Face Without Eyebrows. For the Positive-CIs the whole image analysis was significant for Neutral Face and Neutral Face Without Eyebrows conditions. When the eye region was compared to a similarly sized section of random noise, significant differences were found for all three images for both the Full-CIs and the Positive-CIs. For comparison, similarly sized regions taken around the mouth and eyebrows were never significantly different from noise. However, individual participants produced significant deviations in the eye-brow region for the Neutral Face Without Eyebrows and Mona Lisa images (top row of Fig. 5). For the Negative-CIs, only the eye and eyebrow regions in the Neutral Face Without Eyebrows produced significantly more above threshold pixels than noise. This weak result is somewhat unexpected given that the eye regions look different when Negative-CIs are overlaid on the original faces (Fig. 3). The eye region produces more significant pixels for the Neutral Face Without Eyebrows than the unedited Neutral Face (71 versus 20 pixels for the Full CI, 60 versus 20 pixels for the Positive CI and 4 versus 0 pixels for the Negative CI). Standardising the difference between the two images against the standard deviation of pixel counts across all regions, images, and CI types we find that for the Full CI the without eyebrows condition produced 3.36 × SD more significant pixels in the eye region than the with eyebrows condition ($z = 3.36, p < 0.0005$). This difference was 2.35 × SD for the Positive CIs ($z = 2.35, p < 0.01$) and not significant for the Negative CIs.

We regard the region of interest analyses as more appropriate for judging all types of CIs than the whole image. The latter includes large areas outside the face, or areas of hair, forehead and cheek that are unlikely to be utilized in judgments of surprise. Inclusion of these areas contaminates the count of above threshold pixels with false positives - diluting any effects observed. Analyses focusing on the eyes, eyebrows, and mouth regions expose the utility of the former, and poor utility of the latter in judgments of surprise.

Lowering the threshold for significance in the combined energy map produced very similar results but with more above threshold pixels as did the cluster-based analyses; see supplementary materials.

2.3. *Control experiment*

Our result showing no significant activity in the mouth region is surprising given the weight of evidence showing that the mouth region is more important for judging emotions, specifically surprise, than the eye region (see introduction but also Blais et al., 2012). We were concerned that our choice of noise sample and image size might have rendered the mouths in our stimuli too obvious and thus incapable of supporting a CI approach. We thus repeated our experiment asking participants to detect surprise in the Neutral Face while concentrating on the mouth region. The same participants were used and other than their increased ages the experimental procedure and analysis was identical to that of the main experiment. Fig. 6 shows the resulting Full-CI and the combined thresholded significance map clearly indicating significant deviations in the mouth region which were also significant in the Stage 2 ROI analysis (number of active pixels in mouth region = 22, z = 21.51, p < 0.0001). We thus conclude that our procedure and stimuli are capable of revealing the mouth area but that for some reason participants did not use this region in our main experiment (see Discussion).

3. Experiment 2

3.1. *Method*

3.1.1. *Materials*

We used 24 images including the Full-CIs, Positive-CIs and Negative-CIs from experiment 1 overlaid onto their respective faces (9 images as in Fig. 3). We also included the 3 neutral faces used in experiment 1 (Fig. 1) with added random noise samples. The remaining face images were sourced from the KDEF database, using the same female actor as was used for the neutral face in experiment 1, now acting 6 different expressions (Fig. 7). We also included edited versions of these with eyebrows removed (Figure S5). Random noise samples were added to these 12 emotionally expressive images. The face stimuli were presented at 40% contrast and the noise and CI stimuli at 25% contrast.

In addition to the images, we tested the participants' emotional intelligence with the Schutte Self-Report Emotional Intelligence Test (SSEIT; Schutte et al., 1998). Participants responded to 33 statement items from 1 strongly disagree to 5 strongly agree. Among these questions, we included two attention check questionnaire items. Specifically, the content of the item was a request to choose a certain option (e.g., strongly disagree and somewhat agree). Images and the SSEIT questionnaire were presented using the Qualtrics Survey platform (https://www.qualtrics.com/).

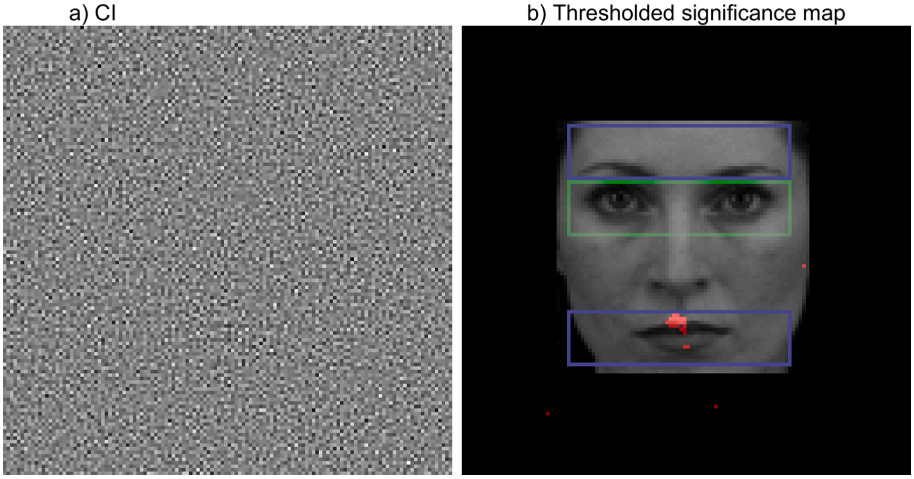

Fig. 6. a) Full CI extracted when participants were instructed to fucus on the mouth region while detecting surprise. b) thresholded significant map resulting from the analysis of (a).

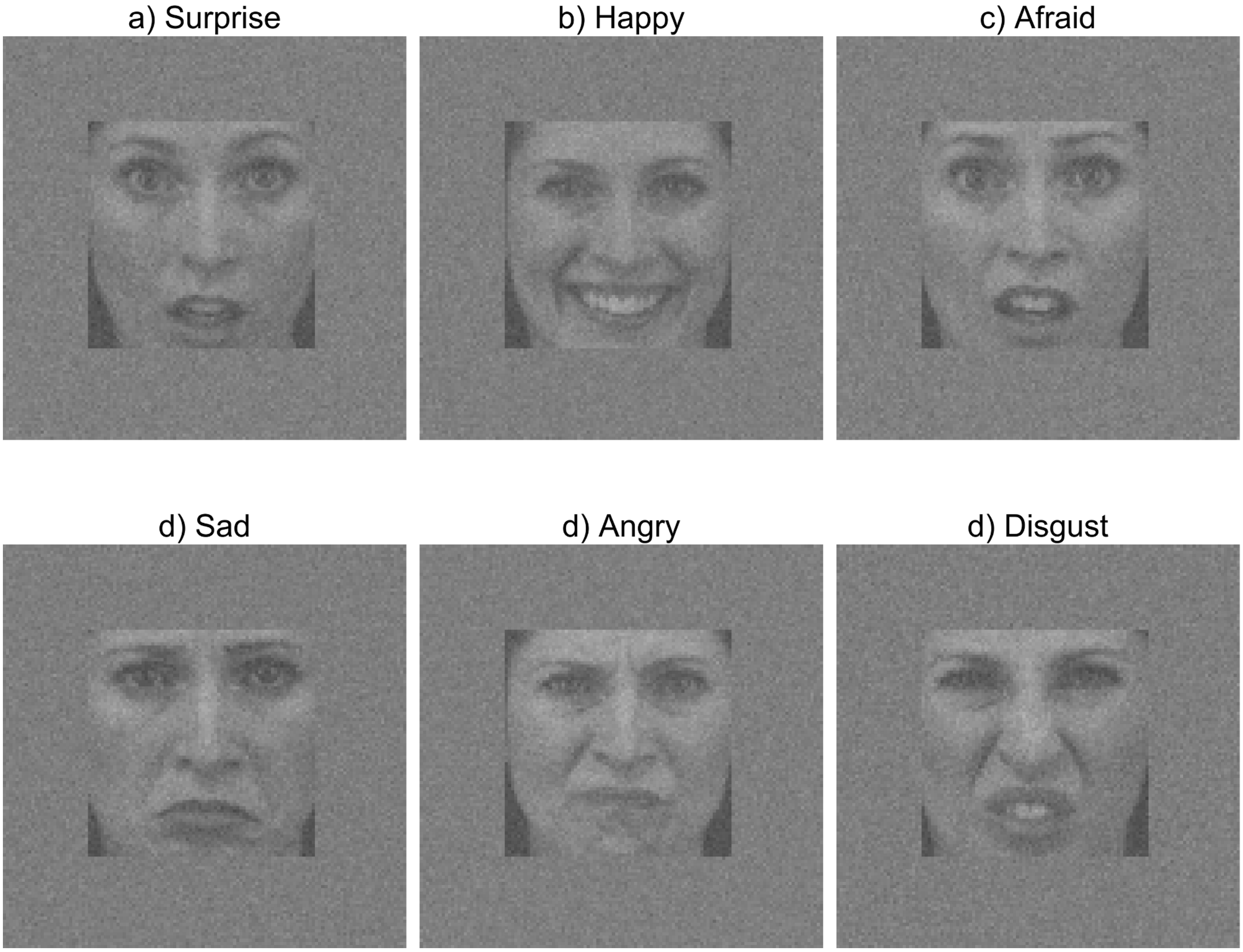

Fig. 7. Emotional face stimuli for Experiment 2. The same actress from the KDEF database with the neutral face in experiment 1 acted different expressions. Here, we added random noise to each face. a) Surprise, b) happy, c) afraid, d) sad, e) angry, and f) disgust.

3.1.2. Procedure

Participants were presented with one image at a time and offered seven emotion options with slide bars (namely: happy, afraid, surprise, sad, angry, disgust, and neutral) presented with the images. They were asked to report the perceived emotion of each image by assigning a score to each possible emotion. The total score for each stimulus had to add up to 100, otherwise, the participant could not continue the survey. Every image was presented in the same size (roughly 4 × 4° for most laptop users). Image size was not scaled by the participant's display size, and we did not record the devices used but only laptop and desktop computers were allowed. The resulting size variation adds noise to our results which was compensated by the large number of participants tested. Thus, we recorded perceived emotion profiles for each image. Participants then completed the SSEIT survey responding on a 5-point Likert scale (1 strongly disagree, 5 strongly agree), with two additional attention checking questions with mandatory answers.

3.1.3. Participants

The participants were recruited from the Prolific (https://www.prolific.co) participation brokering website. We balanced the gender of the sample and required proficiency in English as a prerequisite. All participants who passed our attention check items received £2.75 in compensation. Three out of 104 participants failed the attention check. On average, the survey took our participants 18 min to complete. 48 of the participants identified themselves as female, 49 male, and 4 as non-binary (age range: 19–61, M^{age}: 26.35, *SD*: 8.49). This experiment was evaluated and approved by Aston University's College of Health and Life Sciences Ethical Review committee (approval number 1783). All participants gave informed consent as indicated by responses to a consent from presented via Prolific.

3.1.4. Data analysis

We systematically compared pairs of images to examine whether or not they evoked different emotional response patterns and to elucidate specific differences. First, we conducted a chi-square goodness-of-fit to examine whether each CI (including Full-CI, Positive-CI, and Negative-CI) added to the relevant neutral face image was judged to express a different emotion profile to the same image combined with a random noise sample. To further explore the exact difference between the emotional responses, we conducted repeated measures ANOVAs testing for main effects of emotion category and interactions between emotion category and noise type (face stimulus plus CI versus face stimulus with noise). Post hoc analyses with Tukey correction were also conducted to test differences in the assessment of each emotion between two images. For example, we determined if Neutral Face + Full-CI resulted in higher ratings of surprise than Neutral Face + noise. The statistical results are shown in Tables S1-3 corresponding to Full-CI, Positive-CI, and Negative-CI respectively.

In order to ensure that our participants could correctly judge facial expressions, and that the actor in the KDEF dataset was correctly portraying emotions, responses to images where the KDEF database actor portrayed different emotions were also analysed. The analyses conducted were similar to the comparisons of different CIs. Each facial expression image (surprise, happy, afraid, sad, angry, and disgust) was embedded in random noise and compared to Neutral Face + noise. The statistical results are shown in Tables S4 and S5 corresponding to images with and without eyebrows.

We also tested to see if our Full- and Positive-CIs combined with their original neutral faces resemble the image where the actor portrayed surprise. The results of these tests can be found in Table S6. Finally, the Negative-CIs were compared with the highest rating facial expression (see supplementary materials).

3.2. Results

To summarise the detailed result presented below, when combined with their respective face images all CIs produced emotion response patterns that differed from that produced by the same face combined with random noise. While the pattern of emotional responses recorded did not match that for acted surprised faces, the Positive- and Full-CIs combined with Neutral Face and Neutral Face Without Eyebrows were rated higher for surprise than the corresponding faces plus noise, although these combinations were also rated more highly for 'afraid'. Full- and Positive-CIs added to the Mona Lisa face produced a stronger 'happy' response the same face combined with noise, although the Positive-CI also made her appear surprised. Negative-CIs added to their respective neutral faces produced stronger responses for disgust than the corresponding neutral faces added to random noise. Neutral faces in noise tended to look angry.

Fig. 8 shows the average pattern of emotion type responses for each image. Chi-square goodness of fit indicated that all images with added Full-CIs generated significantly different emotion type responses (Neutral Face: $\chi^2(6) = 75.95$, $p < 0.001$; Neutral Face Without Eyebrows: $\chi^2(6) = 73.67$, $p < 0.001$; Mona Lisa: $\chi^2(6) = 76.46$, $p < 0.001$). Also, regardless of the base image, we found a significant main effect of emotion and an interaction between emotion and mask type (see Table S1 for repeated measures ANOVA results), indicating that the emotion pattern differs between image types. With a focus on surprise, we conducted post hoc analysis with Tukey correction to test if pairs of images differed in the level of surprise. Generally, Full-CIs (generated from neutral faces both with and without eyebrows) overlaid onto the original neutral face images were judged as higher in surprise (Neutral Face: $t(100) = 5.34$, $p < 0.001$; Neutral Face Without Eyebrows: $t(100) = 6.58$, $p < 0.001$) and lower in 'neutral' (Neutral Face: $t(100) = -4.60$, $p < 0.001$; Neutral Face Without Eyebrows: $t(100) = -14.22$, $p < 0.001$) compared to the same face images with random noise overlays. Overlaying the Full-CI onto the Mona Lisa produced more happy ($t(100) = 14.68$; $p < 0.01$) and less neutral ($t(100) = -12.22$, $p < 0.001$) ratings than the random noise counterpart but no difference in surprise ($t(100) = 2.83$, $p = 0.23$). We also found that both versions of the neutral face, combined with their respective Full-CIs, were judged to be higher in 'afraid' than when combined with random noise samples (Neutral Face: $t(100) = 6.14$; $p < 0.001$; Neutral Face Without Eyebrows: $t(100) = 6.41$; $p < 0.001$). The Full-CI added to Neutral Face Without Eyebrows was also judged to be more angry than the same face with random noise ($t(100) = 5.08$, $p < 0.001$).

The pattern of responses to the Positive-CIs plus the two neutral faces were similar to the pattern observed with Full-CIs. Chi-square goodness of fit tests show significant differences between emotion responses for all faces with added Positive-CIs (Neutral Face: $\chi^2(6) = 107.80$, $p < 0.001$; Neutral Face Without Eyebrows: $\chi^2(6) = 93.13$, $p < 0.001$; Mona Lisa: $\chi^2(6) = 50.60$, $p < 0.001$). The main effects of emotion were significant but the interaction between emotion and noise failed to reach significance (see Table S2 repeated measures ANOVA results). Post hoc analyses showed that these Positive-CI images were judged to be higher in surprise (Neutral Face: $t(100) = 9.84$, $p < 0.001$; Neutral Face Without Eyebrows: $t(100) = 9.99$, $p < 0.001$; Mona Lisa: $t(100) = 3.72$, $p = 0.02$) and lower in neutral (Neutral Face: $t(100) = -4.48$, $p = 0.002$; Neutral Face Without Eyebrows: $t(100) = -13.97$, $p < 0.001$; Mona Lisa: $t(100) = -9.53$, $p < 0.001$) compared to their corresponding random noise counterparts. Ratings for afraid were higher for the Positive-CIs for both versions of the neutral face as compared to their corresponding neutral image plus random noise (Neutral Face: $t(100) = 4.36$, $p = 0.002$; Neutral Face Without Eyebrows: $t(100) = 5.55$, $p < 0.001$). Neutral Face + Positive-CIs produced lower ratings for anger than its random noise counterpart ($t(100) = -3.87$, $p = 0.01$). The Mona Lisa + Positive-CI was rated happier ($t(100) = 8.56$, $p < 0.001$) than Mona Lisa + random noise.

The Negative-CIs were analysed with the same method as the other CIs. Regardless of base image, there was a significant effect in chi-square goodness of fit, and a significant emotion main effect, but the interaction between emotion and image type was only significant for the two neutral face images (see Table S3 for chi-square goodness of fit and repeated measures ANOVA results). The interaction between emotion and noise type failed to reach significance ($F(2.85,284.96) = 37.3$, $p = 0.07$) for the Mona Lisa image. Post hoc analyses showed that, when added to their respective base images, all Negative-CIs were rated higher in disgust (Neutral Face: $t(100) = 4.50$, $p = 0.001$; Neutral Face Without Eyebrows: $t(100) = 5.78$, $p < 0.001$; Mona Lisa: $t(100) = 4.89$, $p < 0.001$) and lower in neutral expression (Neutral Face: $t(100) = -7.27$, $p < 0.001$; Neutral Face Without Eyebrows: $t(100) = -18.31$, $p < 0.001$; Mona Lisa: $t(100) = -8.48$, $p < 0.001$), compared to their respective faces

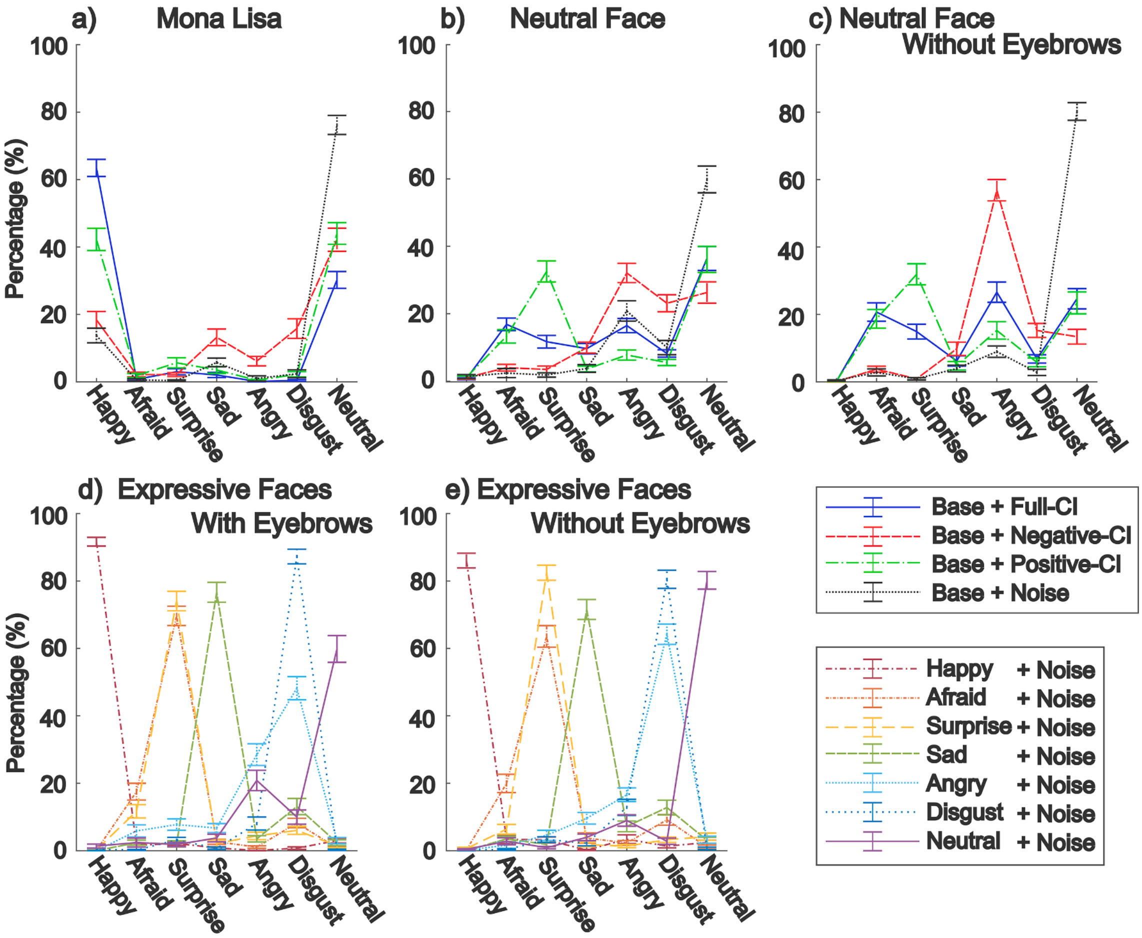

Fig. 8. Response pattern of all items in Experiment 2. (a-c) Response pattern of all CIs and neutral image with noise displayed together for each base image. (d-e) Response pattern of all emotional images. Different colours / line styles represent the acted emotion of the image. Each plot is participants' averaged response to different images. Here, noise refers to the addition of a random noise sample. Error bars denote the standard error of each response.

with random noise. Additionally, Neutral Face Without Eyebrows + Negative-CI was rated higher in anger than this face with a random noise sample ($t(100) = 13.04$, $p < 0.001$).

The images portraying acted emotions (as in Fig. 7 and Figure S1) were compared to their neutral emotion counterpart. With the exception of 'anger', all the images in the database (both with and without eyebrows) were judged higher in their intended emotion than the corresponding neutral image overlaid with random noise (With Eyebrows: surprise: $t(100) = 22.30$, $p < 0.001$; happy: $t(100) = 62.75$, $p < 0.001$; afraid: $t(100) = 5.11$, $p < 0.001$; sad: $t(100) = 21.44$, $p < 0.001$; angry: $t(100) = 1.77$, $p = 0.89$; disgust: $t(100) = 25.80$, $p < 0.001$; Without Eyebrows: surprise: $t(100) = 34.85$, $p < 0.001$; happy: $t(100) = 39.51$, $p < 0.001$; afraid: $t(100) = 6.02$, $p < 0.001$; sad: $t(100) = 20.48$, $p < 0.001$; angry: $t(100) = 2.86$, $p = 0.21$; disgust: $t(100) = 25.79$, $p < 0.001$; see Table S4 and S5 for detailed statistics). However, this does not mean that the intended emotion was rated the highest for every image. For example, the afraid image was judged to be higher in surprise than the neutral image, both with and without eyebrows. Among the 12 comparisons between an emotionally valent image and their neutral counterparts, 7 pairs showed that the neutral image was rated higher for anger than the acted image. It is likely that the neutral image with noise overlay appears somewhat angry which may be why the angry image failed to reach significance in judgements of anger.

We tested to see if our CIs combined with their respective neutral faces resemble the acted surprise image. We examined both images with and without eyebrows, with the added Full- and Positive-CI. Both the chi-square goodness of fit and repeated ANOVA showed that the CIs were significantly different from the acted surprise images (see Table S6 for chi-square goodness of fit and repeated measures ANOVA results). All four comparisons showed higher ratings for surprise for the acted surprise images than our CIs (Neutral Face + Full-CI: $t(100) = -17.39$, $p < 0.001$; Neutral Face Without Eyebrows + Full-CI: $t(100) = -20.61$, $p < 0.001$; Neutral Face + Positive-CI: $t(100) = -9.29$, $p < 0.001$; Neutral Face Without Eyebrows + Positive-CI: $t(100) = -14.67$, $p < 0.001$), and higher neutral response in CIs than the acted surprise images (Neutral Face + Full-CI: $t(100) = 9.72$, $p < 0.001$; Neutral Face Without Eyebrows + Full-CI: $t(100) = 6.44$, $p < 0.001$; Neutral Face + Positive-CI: $t(100) = 8.86$, $p < 0.001$; Neutral Face Without Eyebrows + Positive-CI: $t(100) = 6.04$, $p < 0.001$). Further, both Full- and Positive-CIs combined with Neutral Face Without Eyebrows were rated higher for afraid (Full-CI: $t(100) = 5.20$, $p < 0.001$; Positive-CI: $t(100) = 4.48$, $p = 0.002$) and anger (Full-CI: $t(100) = 8.15$, $p < 0.001$; Positive-CI: $t(100) = 5.28$, $p < 0.001$) than the corresponding acted surprise images. Neutral Face + Full-CI showed higher ratings for sad ($t(100) = 3.90$, $p = 0.012$) and angry ($t(100) = 5.45$, $p < 0.001$) than the corresponding acted surprise image.

Negative CIs combined with the base faces were rated high for disgust. We therefore compared this combination with the acted disgust image added to noise. The results of this analysis (Table S7) showed that the response profile for the negative CIs differed to that of acted disgust.

The results of emotional intelligence were also analysed. Our participants scored 124.37 (standard deviation: 15.79) on average. A Kolmogorov-Smirnov test was used to test for normality of the distribution of emotional intelligence scores ($D(101) = 0.0793$, $p = 0.53$), where we accepted the null hypothesis that the participants' emotional intelligence scores conform to a normal distribution. The emotional intelligence score did not show significant correlation with participants' accuracy in judging the acted emotion ($r(99) = 0.0128$, $p = 0.90$).

4. Discussion

Experiment 1 shows that our participants relied on the eye-region for detecting surprised expressions in briefly presented stimuli. For 'surprise' responses, Positive-CIs contain a template which enhances the contrast of the eye region, and for 'not surprise' responses, Negative-CIs masks the eyes. The eyebrows were not significantly activated in the CI templates for any face. Removal of the eyebrows from the neutral face further localized sampling to the eye-region, suggesting an effect of eyebrows on the detection of surprise. Experiment 2 shows that the neutral face with and without eyebrows combined with CIs for surprise indeed convey surprise to naïve observers more so than the corresponding faces with random noise, although this is less strong than their response to a face acting surprised. Adding surprise CIs to the Mona Lisa image produces the impression of happiness.

While CIs for detecting surprise suggest that the eyes are the primary diagnostic feature which carry the most information for the detection of surprise, participants did not rely on a roughly defined eye-region, but rather a detailed integration of fine-grained information in the eye that drove them to perceive surprised expressions. The Positive-CIs increase the contrast of the eye features, specifically between the iris and sclera, making the eyes appear wide open in an expression of surprise. Conversely, the Negative-CIs add a dark luminance mask to the sclera reducing its contrast with the iris, almost as if the eyes were partially closed. Additionally, the absence of eyebrows did not, generally, encourage the participants to add that facial feature, instead, participants focused more on the eye region in conditions where the eyebrows were absent. We observed no significant contribution from the mouth region in any stimulus condition.

We did not observe the presence of 'illusory eyebrows' in the CIs for the Neutral Face Without Eyebrows image. Gosselin and Schyns (2003) found that their participants interpolated a missing mouth feature, but their task involved direct detection of the mouth, while our task does not directly require the participants to inspect the eyebrows to detect surprise. However, we did observe stronger eye templates in the no-eyebrows condition. This may suggest that reliance on eye information increases in the absence of eyebrow information. It is possible that our participants were relying on sampling information from the eyebrows to some degree when they were present, thus reducing the amount of sampling from the eyes, but not strongly enough to activate the eyebrow regions in most CI templates. Further, the eyebrow region was significantly activated in the Negative CI for the Neutral Face Without Eyebrows.

Our results are at odds with previous studies (Blais et al., 2012, 2017; Smith et al., 2005) showing that the mouth region is important for classifying emotions in general and surprise in particular, and with Jack et al. (2012) who found that Western observers use the mouth and eyebrow regions to judge surprise. There are a number of reasons that might explain this apparent discrepancy. (i) Unlike some of the previous studies we restricted presentation time to focus on observers' 'first-impressions'. We did this because surprise is a somewhat fleeting emotion that tends to not to be expressed over prolonged periods (Ekman, 2003). Thus, our result may apply only to the initial assessment of surprise, however, Blais et al. (2012) used the same presentation time and found the mouth to be important so this conclusion seems unlikely. (ii) Our experiment involved the detection of surprise versus an unspecified alternative emotion whereas the previous studies required a discrimination between multiple emotions. It may be that the mouth and eyebrow regions are informative for discriminating between emotions that have similar content around the eyes but are less relevant for detection. The discrimination of fear and surprise in particular is known to rely on the eyebrows and mouth (Roy-Charland et al., 2015). (iii) Jack et al. (2012) found cultural differences between Western and East Asian participants with the latter group using predominantly eye information. Our observers were ethnically diverse but were all either born in a Western country or have lived in one for a long time. Culture is malleable and it may be that Western observers have shifted away from using the mouth and eyebrow regions in the intervening decade – perhaps exacerbated by the prevalence of mask wearing at the time of our study. However, we think such an extreme change as would be required to produce our results is unlikely, and mask wearing would not account for a shift away from eyebrows. (iv) Our white noise samples had a flat spatial frequency spectrum and as such had constant masking power at all frequencies, but face images do not have a flat spectrum and may contain more energy at the lower spatial frequencies associated with large features such as the mouth and eyebrows. This difference may have caused the mouth to be more visible and thus less ambiguous than the eyes. Further, previous studies using the bubbles method used emotionally valent images with expressive mouths based on acted emotions (Blais et al., 2012). In common with other reverse correlation studies, we used emotionally neutral stimuli with closed mouths, and observers may have regarded this as uninformative for judging the presence of surprise. However, our control condition suggests that participants could use information in the mouth region when directed to do so. (v) Finally, we manipulated the eyebrow region but did not include conditions where other features, particularly the mouth, were removed. The absence of eyebrows in early sessions may have cued participants to focus on the eye / eyebrow region rather than the mouth. While this could explain the lack of activity in the mouth region it does not explain why participants used the eye region in preference to the eyebrows. On balance we feel that task demands such as detection versus discrimination and our removal of the eyebrows are likely to explain the apparent difference between our findings and those in the literature, but more work would be needed to distinguish between these two factors. In particular, the eyebrows and mouth should be deleted in counterbalanced sessions to avoid any misdirection of attention away from the mouth. Thus, while we are confident that participants make more use of the eye region than the eyebrows in our task, further experiments are needed to confirm the relative importance of the eye and mouth regions. Finally, our results may be specific to the small group of people tested and a study in which larger numbers of participants each contributed fewer judgement would confirm our results and allow greater generalisation.

Experiment 2 shows that naïve participants see significant emotional valence in the CI templates from experiment 1 when these are overlaid onto the original neutral face images. The Full-CI templates and the Positive-CI templates appear surprised in the case of the neutral face, with and without eyebrows. For the Mona Lisa, the Full-CI appeared happy, but the Positive-CI appeared both happy and surprised. These results suggest that the CI templates from experiment 1 contain the same emotional valence that the participants were asked to detect. That the Mona Lisa face was seen as happy rather than surprised, is likely because her face is not truly neutral, but rather slightly smiling. Thus, surprised eyes from the CIs, combined with a slight smile in the base image may give the impression of happiness. Similarly, the enhanced ratings for afraid in the two neutral faces may be related to the fact that the neutral face itself portrayed some sense of anger. Surprise faces tinged with anger may appear somewhat afraid.

When Kontsevich and Tyler (2004) examined their negative templates for happy, they clearly got a template for sadness; happy and sad expressions are in clear opposition. Here, we found that the Negative-CIs for surprise tend to produce disgust when added to neutral faces. Although the neutral face's baseline was regarded as slightly angry, while the Mona Lisa's baseline was seen as happy, the disgust emotion in the Negative-CI templates was found consistently in all face stimuli. We speculate (in contrast to Ekman, 2003) that surprise typically follows the consumption of pleasant information, while disgust is more likely to be related to aversion or regret following the consumption of unpleasant information. In that regard, these two emotions could be in opposition with each other. Our finding was consistent with two models which theorised surprise and disgust as opposite (Lövheim, 2012; Woodworth and Schlosberg, 1954; Young et al., 1997). For example, while Lövheim's (2012) model may place surprise and fear in opposition it also highlights that disgust and surprise respectively reflect low and high level of noradrenaline while serotonin is high and dopamine is low for both emotions. Further, behavioural results show that manipulations that hinder the perception of disgust amplify the perception of surprise (Daudelin-Peltier et al., 2017). The KDEF dataset generally demonstrated high validity. All the acted emotions were judged high on the attribute that they were supposed to represent. However, the angry image was not rated as significantly more angry than the neutral face, due to the neutral face being rated as angry by our participants. Further, the afraid image was often confused as being surprised instead of afraid, probably due to the correlated nature of the two emotions. It is possible that this feature affected our results as noted above and future research might define a more consistently neutral baseline.

The current study highlights the methodological potential in the CI method by showing that CIs can reveal sampling of diagnostic features for surprise (and its opposite 'negative-surprise' template) as a facial expression. We find that our participants rely on a detailed, high spatial frequency, information sampling strategy around the eyes to drive surprise detection. This depends on structural information that is not revealed in low-level image statics. We also find that CIs generated through surprise detection convey surprise to other naïve participants, indicating that CIs can capture and convey important elements of face perception. The current study thus furthers our understanding of how humans sample information about each other's mental states from faces.

CRediT authorship contribution statement

Emil Skog: Conceptualization, Methodology, Software, Investigation, Resources, Writing – original draft, Writing – review & editing, Visualization. **C. Stella Qian:** Conceptualization, Methodology, Validation, Formal analysis, Investigation, Resources, Writing – original draft, Writing – review & editing, Visualization. **Anisha Parmar:** Methodology, Resources, Investigation. **Andrew J. Schofield:** Conceptualization, Methodology, Validation, Formal analysis, Investigation, Writing – original draft, Writing – review & editing, Visualization, Supervision.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The data supporting this publication can be accessed at: https://doi.org/10.17036/researchdata.aston.ac.uk.00000596

Acknowledgements

ES was supported by funding from Aston University (UK) and Ordnance Survey Ltd. CSQ was supported by the Engineering and Physical Sciences Research Council; grant number EP/S016260/1. We particularly thank the non-author participants in experiment 1 and the control experiment for their contribution. We thank the two reviewers for their insightful comments that improved this paper.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.visres.2023.108275.

References

Ahumada, A. J., Jr. (1996). Perceptual classification images from Vernier acuity masked by noise. *Perception, 25*(1_suppl), 2. https://doi.org/10.1068/v96l0501

Blais, C., Roy, C., Fiset, D., Arguin, M., & Gosselin, F. (2012). The eyes are not the window to basic emotions. *Neuropsychologia, 50*(12), 2830–2838.

Blais, C., Fiset, D., Roy, C., Saumure Régimbald, C., & Gosselin, F. (2017). Eye fixation patterns for categorizing static and dynamic facial expressions. *Emotion, 17*(7), 1107.

Beard, B. L., & Ahumada, A. J., Jr. (1998). Technique to extract relevant image features for visual tasks. *Human Vision and Electronic Imaging III, 3299*, 79–85. https://doi.org/10.1117/12.320099

Beaudry, O., Roy-Charland, A., Perron, M., Cormier, I., & Tapp, R. (2014). Featural processing in recognition of emotional facial expressions. *Cognition and Emotion, 28*(3), 416–432. https://doi.org/10.1080/02699931.2013.833500

Blakemore, C., & Campbell, F. W. (1969). On the existence of neurones in the human visual system selectively sensitive to the orientation and size of retinal images. *The Journal of physiology, 203*(1), 237–260.

Brinkman, L., Todorov, A., & Dotsch, R. (2017). Visualising mental representations: A primer on noise-based reverse correlation in social psychology. *European Review of Social Psychology, 28*(1), 333–361. https://doi.org/10.1080/10463283.2017.1381469

Calder, A. J., Keane, J., Young, A. W., & Dean, M. (2000). Configural information in facial expression perception. *Journal of Experimental Psychology: Human Perception and Performance, 26*(2), 527–551. https://doi.org/10.1037/0096-1523.26

Campbell, F. W., & Robson, J. G. (1968). Application of Fourier analysis to the visibility of gratings. *The Journal of physiology, 197*(3), 551.

Chauvin, A., Worsley, K. J., Schyns, P. G., Arguin, M., & Gosselin, F. (2005). Accurate statistical tests for smooth classification images. *Journal of vision, 5*(9), 1.

Cone, J., Brown-Iannuzzi, J. L., Lei, R., & Dotsch, R. (2021). Type I error is inflated in the two-phase reverse correlation procedure. *Social Psychological and Personality Science, 12*(5), 760–768.

da Vinci, L. (c. 1503-1506). Mona Lisa. [Oil on Poplar]. Louvre, Paris, France.

Daudelin-Peltier, C., Forget, H., Blais, C., Deschênes, A., & Fiset, D. (2017). The effect of acute social stress on the recognition of facial expression of emotions. *Scientific Reports, 7*(1), 1–13.

Dotsch, R., & Todorov, A. (2012). Reverse Correlating Social Face Perception. *Social Psychological and Personality Science, 3*(5), 562–571. https://doi.org/10.1177/1948550611430272

Eckstein, M. P., Pham, B. T., & Shimozaki, S. S. (2004). The footprints of visual attention during search with 100% valid and 100% invalid cues. *Vision Research, 44*(12), 1193–1207. https://doi.org/10.1016/j.visres.2003.10.026

Eckstein, M. P., Shimozaki, S. S., & Abbey, C. K. (2002). The footprints of visual attention in the Posner cueing paradigm revealed by classification images. *Journal of Vision, 2*(1), 25–45. https://doi.org/10.1167/2.1.3

Ekman, P. (2003). *Emotions revealed* (2nd ed.). New York: Times Books.

Ekman, P., Friesen, W. V., & Simons, R. C. (1985). Is the startle reaction an emotion? *Journal of Personality and Social Psychology, 49*(5), 1416–1426.

Gold, J. M., Murray, R. F., Bennett, P. J., & Sekuler, A. B. (2000). Deriving behavioural receptive fields for visually completed contours. *Current Biology, 10*(11), 663–666. https://doi.org/10.1016/S0960-9822(00)00523-6

Gold, J. M., Sekuler, A. B., & Bennett, P. J. (2004). Characterizing perceptual learning with external noise. *Cognitive Science, 28*(2), 167–207. https://doi.org/10.1016/j.cogsci.2003.10.005

Gosselin, F., & Schyns, P. G. (2003). Superstitious perceptions reveal properties of internal representations. *Psychological Science, 14*(5), 505–509. https://doi.org/10.1111/1467-9280.03452

Gosselin, F., & Schyns, P. G. (2001). Bubbles: A technique to reveal the use of information in recognition tasks. *Vision research, 41*(17), 2261–2271.

Hsiao, J. H. W., & Cottrell, G. (2008). Two fixations suffice in face recognition. *Psychological science, 19*(10), 998–1006.

Jack, R. E., Caldara, R., & Schyns, P. G. (2012). Internal representations reveal cultural diversity in expectations of facial expressions of emotion. *Journal of Experimental Psychology: General, 141*(1), 19–25. https://doi.org/10.1037/a0023463

Kontsevich, L. L., & Tyler, C. W. (2004). What makes Mona Lisa smile? *Vision Research, 44*(13), 1493–1498. https://doi.org/10.1016/j.visres.2003.11.027

Lövheim, H. (2012). A new three-dimensional model for emotions and monoamine neurotransmitters. *Medical hypotheses, 78*(2), 341–348.

Lundqvist, D., Flykt, A., & Öhman, A. (1998). *Karolinska Directed Emotional Faces (KDEF)* [Database record]. *APA PsycTests*. https://doi.org/10.1037/t27732-000

Mangini, M. C., & Biederman, I. (2004). Making the ineffable explicit: Estimating the information employed for face classifications. *Cognitive Science, 28*(2), 209–226. https://doi.org/10.1016/j.cogsci.2003.11.004

Martin-Malivel, J., Mangini, M. C., Fagot, J., & Biederman, I. (2006). Do humans and baboons use the same information when categorizing human and baboon faces? *Psychological Science, 17*(7), 599–607. https://doi.org/10.1111/j.1467-9280.2006.01751.x

Murray, R. F. (2011). Classification images: A review. *Journal of vision, 11*(5), 1–25. https://doi.org/10.1167/11.5.2

Murray, R. F., Bennett, P. J., & Sekuler, A. B. (2005). Classification images predict absolute efficiency. *Journal of Vision, 5*(2), 139–149. https://doi.org/10.1167/5.2.5

Peirce, J., Gray, J. R., Simpson, S., MacAskill, M., Höchenberger, R., Sogo, H., … Lindeløv, J. K. (2019). PsychoPy2: Experiments in behavior made easy. *Behavior Research Methods, 51*(1), 195–203. https://doi.org/10.3758/s13428-018-01193-y

Roy-Charland, A., Perron, M., Young, C., Boulard, J., & Chamberland, J. A. (2015). The confusion of fear and surprise: A developmental study of the perceptual-attentional limitation hypothesis using eye movements. *The Journal of Genetic Psychology, 176*(5), 281–298.

Sachs, M. B., Nachmias, J., & Robson, J. G. (1971). Spatial-Frequency Channels in Human Vision. *JOSA, 61*, 1176–1186.

Schutte, N. S., Malouff, J. M., Hall, L. E., Haggerty, D. J., Cooper, J. T., Golden, C. J., & Dornheim, L. (1998). Development and validation of a measure of emotional intelligence. *Personality and Individual Differences, 25*(2), 167–177. https://doi.org/10.1016/S0191-8869(98)00001-4

Smith, M. L., Cottrell, G. W., Gosselin, F., & Schyns, P. G. (2005). Transmitting and decoding facial expressions. *Psychological Science, 16*(3), 184–189. https://doi.org/10.1111/j.0956-7976.2005.00801.x

Thomas, J. P., & Gille, J. (1979). Bandwidths of orientation channels in human vision. *JOSA, 69*(5), 652–660.

Tjan, B. S., & Nandy, A. S. (2006). Classification images with uncertainty. *Journal of Vision, 6*(4), 387–413. https://doi.org/10.1167/6.4.8

Woodworth, R. S., & Schlosberg, H. (1954). *Experimental psychology* (Revised edition). New York: Henry Holt.

Yang, N., Shafai, F., & Oruc, I. (2014). Size determines whether specialized expert processes are engaged for recognition of faces. *Journal of vision, 14*(8), 17.

Young, A. W., Rowland, D., Calder, A. J., Etcoff, N. L., Seth, A., & Perrett, D. I. (1997). Facial expression megamix: Tests of dimensional and category accounts of emotion recognition. *Cognition, 63*(3), 271–313.