Computer Science and Artificial Intelligence Laboratory

Technical Report

MIT-CSAIL-TR-2013-004
CBCL-311

March 18, 2013

# Faces as a "Model Category" for Visual Object Recognition

Cheston Tan and Tomaso Poggio

massachusetts institute of technology, cambridge, ma 02139 usa — www.csail.mit.edu

# Faces as a “Model Category” for Visual Object Recognition

**Authors:** Cheston Tan[1,2,3], Tomaso Poggio[1,2]

**Affiliations:**

[1] McGovern Institute for Brain Research, Cambridge, MA 02139, USA.

[2] Department of Brain and Cognitive Sciences, MIT, Cambridge, MA 02139, USA.

[3] Institute for Infocomm Research, Singapore 138632, Singapore.

**Correspondence:** C. T. (cheston@alum.mit.edu) or T. P. (tp@ai.mit.edu).

**Visual recognition is an important ability that is central to many everyday tasks such as reading, navigation and social interaction, and is therefore actively studied in neuroscience, cognitive psychology and artificial intelligence. There exist thousands of object categories[1], all of which pose similar challenges to biological and artificial visual systems: accurate recognition under varying location, scale, view angle, illumination and clutter. In many areas of science, important discoveries have been made using “model organisms” such as fruit flies, mice and macaques. For the thousands of object categories, the important and well-studied category of faces could potentially serve as a “model category” upon which efforts are focused, and from which fundamental insights are drawn. However, it has been hotly debated whether faces are processed by the brain in a manner fundamentally different from other categories[2-6]. Here we show that “neural tuning size” – a single parameter in a computational model of object processing – is able to account for important face-**

**specific phenomena. Thus, surprisingly, "face-like" processing is explainable by physiological mechanisms that differ only quantitatively from "object-like" processing. Our computational proof-of-principle provides specific neural tuning properties that correspond to the so-far qualitative and controversial notion of "holistic" face processing. Overall, faces may be a viable model category. Since faces are highly amenable to complementary experimental techniques like functional MRI[7], electrophysiology[8], electroencephalography[9] and transcranial magnetic stimulation[10], this further raises the odds that the algorithms and neural circuits underlying visual recognition may first be solved for faces[11]. With faces serving as a model category, the great scientific challenge of understanding and reverse-engineering general visual recognition can be greatly accelerated.**

Building upon the family of simple, biologically-plausible visual recognition models[12-16], we found that a single parameter determines whether processing is "face-like" or "object-like", as gauged by two important face-specific behavioural phenomena. The first, the Composite Face Effect[17] (CFE), is the phenomenon whereby two identical top halves are sometimes incorrectly perceived as different when paired with different bottom halves (Fig. 1a). This effect is ostensibly due to the top and bottom halves of each composite being perceived "holistically" (together as a whole) when aligned, despite instructions to ignore the bottom halves. Perception is more accurate when the halves are misaligned (Fig. 1b). Crucially, this effect occurs only for faces, and is therefore commonly taken as evidence that face and object processing are qualitatively different[6,18-20]. Is this necessarily so?

We probed the minimal conditions required to produce – or abolish – such holistic face processing, and found that a vital factor is the size of the template that specifies the tuning of each neuron (henceforth termed “tuning size”). Tuning size is defined in terms of proportion of a whole face covered by a template (see Methods section), not in terms of number of pixels or degrees of visual angle. When tuning size is large, even without encompassing the whole face, the Composite Face Effect is found (Fig. 1c). A single change – reduction of tuning size – abolishes the Composite Face Effect, i.e. leads to “object-like” processing of faces. Thus, our results show that a seemingly qualitative difference between “face-like” and “object-like” behaviour could simply stem from a quantitative difference in one parameter of the underlying mechanisms.

“Holism” is a controversial psychological construct with multiple interpretations and putative mechanisms[4,20,21] for which a consensus has yet to emerge. Our simulation results promote one particular interpretation of holism, that it is simply the byproduct of having large tuning size – a theoretical clarification to earlier proposals[18,22,23]. The Composite Face Effect is found using each individual model neuron with large tuning size by itself (Fig. 1c inset), even though tuning size is less than half the whole face. Conversely, even though neurons with small tuning size collectively span the whole face, they do not produce the Composite Face Effect (Fig. 1c).

Since there is nothing qualitatively “whole”, “singular”, “unified”, “global” or “non-decomposable” about processing that uses large tuning size rather than small, the term “holistic” may be somewhat misleading (to the extent that it implies a qualitative difference, an absolute whole, or integration into a single representation). Our results do, however, indicate that holism is not an all-or-none phenomenon, but it is one that can vary continuously depending on tuning size (and be modulated by other factors).

In our simulations, tuning size is the sole change between the two conditions depicted in Fig. 1c (*Large* and *Small* tuning size), which suggests that decisional and attentional factors are not key. Rather, what matters is the amount of “perceptual integration”, as controlled by tuning size. Additionally, while detection and segmentation are important processes for accurate face recognition, the absence of explicit mechanisms for these in our simulations suggest that they are also not key factors relating to holism.

Tuning size also accounts for another key face-specific phenomenon, the Face Inversion Effect (FIE), whereby upside-down inversion disrupts face processing significantly more than object processing[23,24]. We found that when tuning size is reduced, the behavioural effect of inversion is also reduced (Fig. 2), akin to face processing becoming “object-like”. Our simulations also show that inversion reduces the mean response of each individual neuron (Fig. 3), illustrating the neural basis of the behavioural Face Inversion Effect[25].

The Face Inversion Effect has sometimes been associated with “configural” rather than “holistic” processing of faces, but their relationship is unclear[6,20,26]. Our simulation results demonstrate a link between these two notions, through the common causal factor of large tuning size. Because neurons with large tuning size cover more than individual face parts, they are more sensitive to the configuration of multiple parts, which is altered by inversion. For the exact same reason, these neurons are also more sensitive to information that comes from more distant regions of the face image (in the case of the

Composite Face Effect). The notion of large tuning size may also be able to account for another classic face-related phenomenon – sensitivity to spacing between face parts. The idea is that since each neuron's tuning is specified by a certain face template, any deviation from that template, such as changing the distance between the eyes, will reduce the neural response. If so, then large tuning size provides a unified account of this important trinity of face-specific effects.

By changing only tuning size and keeping everything else unchanged (Figs. 1-3), our simulations are able to sidestep a confound that is unavoidable for empirical experiments that investigate mechanisms underlying face versus object processing – the confound of different stimuli. Empirically, face and object stimuli elicit measurable differences, but do these stem from differences in physical stimulus properties, or from differences in processing mechanisms? Here, instead of changing the stimuli to produce measurable differences, we changed only the underlying processing but not the stimuli, so this confound is avoided.

Clearly, our simulations do not capture the full complexity of face processing, nor is tuning size necessarily the only difference between face and object processing mechanisms. However, we have shown that a change in tuning size alone can account for two phenomena commonly thought to be characteristic of face processing. Therefore, neither the Composite Face Effect nor Face Inversion Effect require face and object processing to be fundamentally different.

Our results suggest that both effects stem from a common cause: large tuning size. This is consistent with actual face-selective neurons being tuned to multiple face parts, but not necessarily the entire face[27,28]. Face recognition algorithms that uses corresponding features (large but not whole-face) show excellent performance[29]. Visual deprivation during infancy abolishes the Composite Face Effect[30], suggesting the possibility that the frequent close-up viewing of faces during normal infancy may cause a significant portion of face-sensitive neurons to have large tuning size – and could explain why, in practice, holism is face-specific.

Any organism, model or otherwise, is unique in some way. Likewise, among categories, faces may require unique mechanisms for gaze and expression processing. Nonetheless, processing of identity for faces and non-face objects may share enough similarities for faces to serve as a model category, accelerating progress in understanding and reverse-engineering visual object recognition.

## Methods Summary

**Model.** The HMAX model[15] simulates hierarchical processing in visual cortex. The model's lower two layers (*S1* and *C1*) contain neurons selective for various orientations. The upper two layers (*S2* and *C2*) contain model neurons that are tuned during an unsupervised template-learning process, performed prior to normal model operation. Template-learning simply involves storing "snapshots" of *C1* activity produced in response to some set of training images. Subsequently, these snapshots become templates that new images are matched against. This template-matching produces the *S2* layer, and pooling of the *S2* model neurons over all image positions and scales produces the *C2* layer.

**Tuning size.** Each small template is roughly the size of a face part (e.g. eye), while each large template covers multiple face parts but not the whole face. Since each template was learnt from a different part of a training image, even the small templates (collectively) spanned the whole face. Tuning size is defined as proportion of a whole face, not in pixels or visual angle.

**Face Inversion Effect.** Dissimilarity between two images was defined as the Euclidean distance between the two sets of *C2* layer responses. Fig. 2c shows the mean dissimilarity between all pairs of faces. Fig. 3a shows the mean response to all individual faces.

**Composite Face Effect (CFE).** On each trial, two composites are presented, and their top halves are judged to be same or different, ignoring the bottom halves. The Composite Face Effect is defined as a higher hit-rate[6] (i.e. accuracy on "*same*" trials) for misaligned than aligned composites. For each pair of composites, if their dissimilarity (Euclidean distance) is below some threshold, the composites are considered "*same*". For each model neuron type (e.g. small tuning size), the threshold is set so that the aligned, upright hit-rate is 75%, but results are robust to threshold used.

## References

1. Biederman, I. Recognition-by-components: a theory of human image understanding. *Psychol. Rev.* **94**, 115 (1987).

2. Diamond, R. & Carey, S. Why faces are and are not special: an effect of expertise. *J. Exp. Psychol. Gen.* **115**, 107 (1986).

3. Gauthier, I. & Logothetis, N. K. Is face recognition not so unique after all? *Cogn. Neuropsychol.* **17**, 125 (2000).

4. Maurer, D., Le Grand, R. & Mondloch, C. J. The many faces of configural processing. *Trends Cogn. Sci.* **6**, 255 (2002).

5. Robbins, R. & McKone, E. No face-like processing for objects-of-expertise in three behavioural tasks. *Cognition* **103**, 34 (2007).

6. McKone, E. & Robbins, R. in *The Oxford Handbook of Face Perception* (eds Calder, A. J., Rhodes, G., Johnson, M. H. & Haxby, J. V.) 149-176 (Oxford Univ. Press, 2011).

7. Kanwisher, N., McDermott, J. & Chun, M. The fusiform face area: a module in human extrastriate cortex specialised for face perception. *J. Neurosci.* **17**, 4302 (1997).

8. Tsao, D. Y., Freiwald, W. A., Tootell, R. B. H. & Livingstone, M. S. A cortical region consisting entirely of face-selective cells. *Science* **311**, 670 (2006).

9. Rossion, B. *et al.* The N170 occipito-temporal component is delayed and enhanced to inverted faces but not to inverted objects: an electrophysiological account of face-specific processes in the human brain. *Neuroreport* **11**, 69 (2000).

10. Pitcher, D., Walsh, V., Yovel, G. & Duchaine, B. TMS evidence for the involvement of the right occipital face area in early face processing. *Curr. Biol.* **17**, 1568 (2007).

11. Kanwisher, N. & Yovel, G. in *Handbook of Neuroscience for the Behavioural Sciences* (eds Berntson, G. G. & Cacioppo, J. T.) 841-858 (Wiley, 2009).

12. Fukushima, K. Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biol. Cybern.* **36**, 193 (1980).

13. Perrett, D. I. & Oram, M. W. Neurophysiology of shape processing. *Image Vision Comput.* **11**, 317 (1993).

14. Wallis, G. & Rolls, E. T. Invariant face and object recognition in the visual system. *Prog. Neurobiol.* **51**, 167 (1997).

15. Riesenhuber, M. & Poggio, T. Hierarchical models of object recognition in cortex. *Nature Neurosci.* **2**, 1019 (1999).

16. Cottrell, G. W., Branson, K. M. & Calder, A. J. in *Proceedings of the 24th Annual Conference of the Cognitive Science Society* (eds Gray, W. D. & Schunn, C.) 238-243 (Lawrence Erlbaum, 2002).

17. Young, A. W., Hellawell, D. & Hay, D. C. Configurational information in face perception. *Perception* **16**, 747 (1987).

18. Tsao, D. Y. & Livingstone, M. S. Mechanisms of face perception. *Annu. Rev. Neurosci.* **31**, 411 (2008).

19. Tanaka, J. W. & Gordon, I. in *The Oxford Handbook of Face Perception* (eds Calder, A. J., Rhodes, G., Johnson, M. H., Haxby, J. V.) 177-194 (Oxford Univ. Press, 2011).

20. Richler, J. J., Palmeri, T. J. & Gauthier, I. Meanings, mechanisms, and measures of holistic processing. *Front. Psychol.*, **3**, 553 (2012).

21. Piepers, D. & Robbins, R. A review and clarification of the terms “holistic”, “configural” and “relational” in the face perception literature. *Front. Psychol.*, **3**, 559 (2012).

22. Farah, M. J., Wilson, K. D., Drain, M. & Tanaka, J. N. What is “special” about face perception? *Psychol. Rev.* **105**, 482 (1998).

23. Rossion, B. & Gauthier, I. How does the brain process upright and inverted faces? *Behav. Cogn. Neurosci. Rev.* **1**, 63 (2002).

24. Yin, R. Looking at upside-down faces. *J. Exp. Psychol.* **81**, 141 (1969).

25. Yovel, G. & Kanwisher, N. The neural basis of the behavioural face-inversion effect. *Curr. Biol.* **15**, 2256 (2005).

26. Rossion, B. Picture-plane inversion leads to qualitative changes of face perception. *Acta Psychol.* **128**, 274 (2008).

27. Perrett, D. I., Rolls, E. T. & Caan, W. Visual neurones responsive to faces in the monkey temporal cortex. *Exp. Brain Res.* **47**, 329 (1982).

28. Freiwald, W. A., Tsao, D. Y. & Livingstone, M. S. A face feature space in the macaque temporal lobe. *Nature Neurosci.* **12**, 1187 (2009).

29. Viola, P. & Jones, M. J. Robust real-time face detection. *Int. J. Comput. Vision* **57**, 137 (2004).

30. Le Grand, R., Mondloch, C. J., Maurer, D. & Brent, H. P. Impairment in holistic face processing following early visual deprivation. *Psychol. Sci.* **15**, 762 (2004).

**Acknowledgements** We thank J. DiCarlo, R. Desimone and J. Leibo for comments on the manuscript; N. Kanwisher, P. Sinha, D. Sheinberg, I. Gauthier, G. Cottrell and the Perceptual Expertise Network for discussions on earlier versions of this work. This work was supported by grants from DARPA (IPTO and DSO), National Science Foundation (NSF-0640097, NSF-0827427), AFSOR-THRL (FA8650-05-C-7262). Additional support (to T. P.) was provided by: Adobe, Honda Research Institute USA, King Abdullah University Science and Technology grant to B. DeVore, NEC, Sony and especially by the Eugene McDermott Foundation. C. T. was supported by fellowships from A*STAR and the McGovern Institute for Brain Research.

**Author Contributions** C. T. and T. P. designed the study. C. T. performed experiments, analysed data and wrote the manuscript. C. T. and T. P. discussed the results and revised the manuscript.

## Figure Legends

**Figure 1. Tuning size determines whether processing is "face-like" or "object-like", as gauged by the Composite Face Effect (CFE).** **a,** Aligned composite faces. Top halves are identical, while bottom halves are different. People sometimes incorrectly perceive the two identical top halves as different. **b,** Misaligned composite faces. Human judgement of the top halves (as being identical) is significantly more accurate for misaligned than aligned composites. **c,** Simulations show that the CFE is produced by neurons with large – but not small – tuning size. **inset:** each individual neuron with large tuning size can produce the CFE. Error bars: ± 1 standard error.

**Figure 2. Tuning size accounts for the behavioural Face Inversion Effect (FIE). a, b,** Illustration of the FIE: dissimilarity between faces is more apparent for upright than inverted faces. **c,** Simulations show that decrease in dissimilarity varies with tuning size. **d,** FIE effect size (upright dissimilarity – inverted dissimilarity) varies with tuning size. Neurons with small tuning size show "object-like" processing, i.e. minimal inversion effect. Error bars: ± 1 standard error.

**Figure 3. Tuning size accounts for the neural Face Inversion Effect (FIE). a,** In terms of mean individual neuron response to single faces (as opposed to dissimilarities between pairs of faces; Fig. 2), tuning size also accounts for susceptibility to inversion. **b,** FIE effect size (upright response – inverted response) varies with tuning size. Error bars: ± 1 standard error.

## Methods

**Model.** The HMAX model[15] simulates hierarchical processing in primate visual cortex, reflecting the increase in neural tuning complexity and invariance up the hierarchy. The lowest levels correspond to orientation-selective cells in primary visual cortex, while the highest levels correspond to face-selective and object-selective cells in inferotemporal cortex.

We used the model implementation found at http://cbcl.mit.edu/jmutch/cns/. Of the four model layers, the orientation-selective lower two layers (*S1* and *C1*) contain model neurons tuned to Gabor patches of various orientations and spatial frequencies; the parameters have been pre-determined based on prior electrophysiological data. The upper two layers (*S2* and *C2*) contain model neurons that are tuned during an unsupervised template-learning process, performed prior to normal model operation. Template-learning simply involves storing "snapshots" of *C1* activity produced in response to some set of training images. In subsequent model operation, these snapshots act as templates that new images are matched against. The *S2* layer comprises the output of this template-matching process, and pooling of the *S2* model neurons over all image positions and scales (for invariance to these) produces the *C2* layer. If training images consist of faces, then *S2* and *C2* model neurons are face-selective. All simulations used 1000 *C2* model neurons.

**Tuning size.** The critical independent variable is "tuning size". Large, medium and small tuning sizes correspond respectively to *S2* tuning templates covering 12x12, 8x8 and 4x4

*C1* model neurons, all from the relatively coarse scale 7 (out of 9). At this scale, the entire face oval corresponds to 17x22 *C1* neurons, so each small template is roughly the size of a face part (e.g. eye, nose), while each large template covers multiple face parts but not the whole face.

Importantly, “tuning size” is defined as the proportion of a whole face covered by a template. This is not the same as “size” defined in terms of number of pixels or degrees of visual angle. In the human and primate visual systems (as well as our model), there exists some invariance to image scale. Therefore, a particular tuning size (e.g. half a face) can correspond to a range of physical sizes (in pixels or degrees of visual angle).

Since each template was learnt from a different (random) part of a training image, even the 1000 small templates (collectively) spanned the whole face – yet they did not produce a Composite Face Effect (Fig. 1c), thus ruling out some alternative accounts of mechanisms underlying “holistic processing”.

**Stimuli.** Stimuli were derived from 100 frontal-view male faces from the MPI database (http://faces.kyb.tuebingen.mpg.de/). Faces were downscaled by 25%, and then oval-cropped to remove outline and external features (e.g. hair). Faces were normalised so that all had the same pixel-value statistics (mean and variance). Odd-numbered faces were used for template-learning, even-numbered faces for normal operation. All faces were upright unless explicitly inverted. Note: in Figs. 1 and 2, backgrounds were cropped to save space and make face details more apparent.

**Face Inversion Effect.** Dissimilarity between two images was defined as the Euclidean distance between the two sets of *C2* layer responses. Fig. 2c shows the mean dissimilarity between all 1225 pairs of faces within each condition. Fig. 3a shows the mean response (averaged over all model neurons) to all 50 faces. Error bars were derived using 10,000 bootstrap runs.

**Composite Face Effect (CFE).** Composites were constructed by pairing the top of one face with the bottom of another (with a two-pixel gap). Only 20 faces were used; these were chosen prior to simulations, for behavioural replication of the CFE (not reported here).

On each trial, two composites are presented, and their top halves are judged to be same or different, ignoring the bottom halves. Only trials with identical top halves are analysed[6]. The Composite Face Effect is defined as a higher hit-rate (i.e. accuracy on these *"same"* trials) for misaligned than aligned composites.

To simulate human subjects looking and attending to the top halves, bottom-half pixel values are multiplied by 0.1, and faces shifted downwards so that the top halves occupy the center. To simulate subjects comparing composites, if the dissimilarity between composites (Euclidean distance between the two sets of *C2* layer responses) is below some threshold, the composites are considered *"same"*. For each model neuron type (e.g. small tuning size), the threshold is set so that the aligned, upright hit-rate is 75%, but the results are qualitatively robust to the threshold used. Error bars were derived using 1000 bootstrap runs.

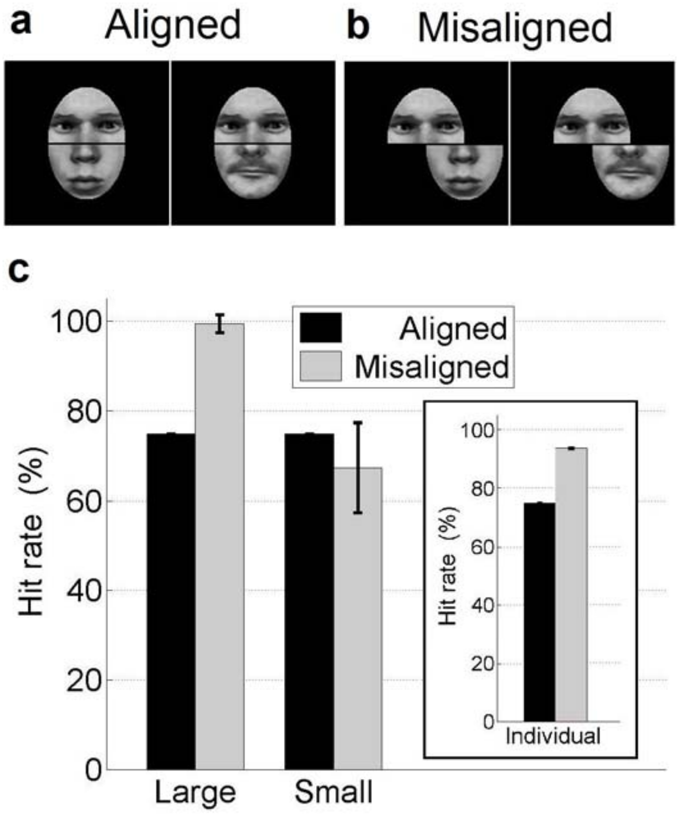

**Figure 1. Tuning size determines whether processing is “face-like” or “object-like”, as gauged by the Composite Face Effect (CFE)**.

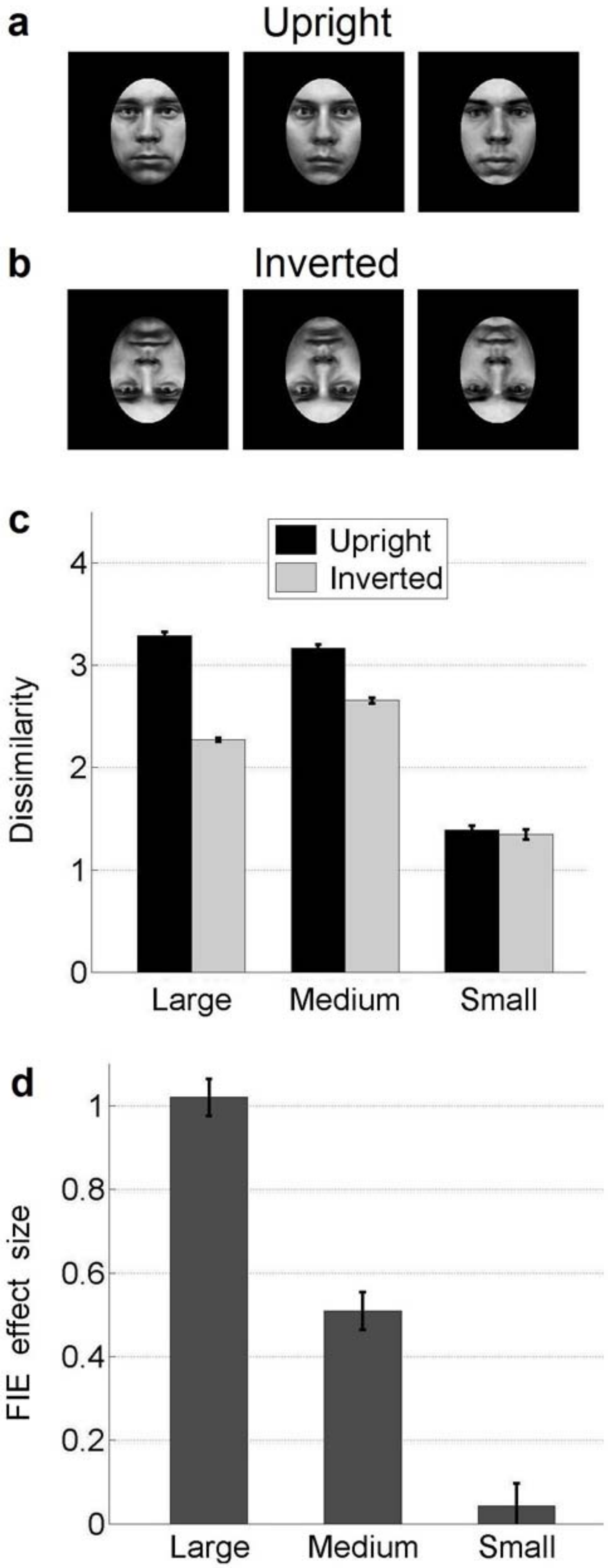

**Figure 2. Tuning size accounts for the behavioural Face Inversion Effect (FIE).**

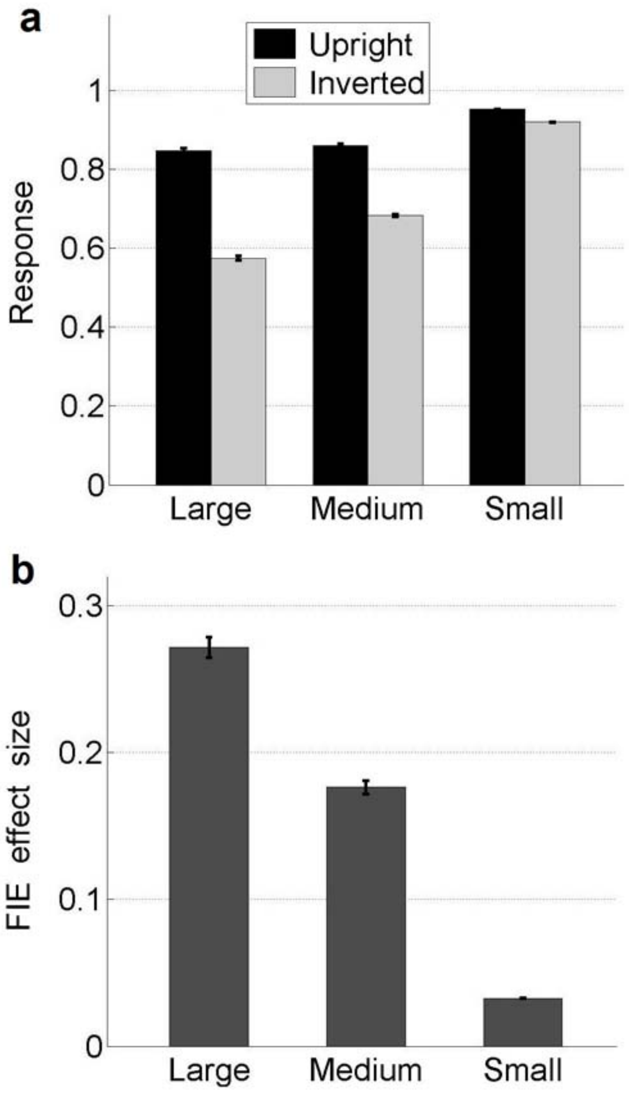

**Figure 3. Tuning size accounts for the neural Face Inversion Effect (FIE).**