Computer Science and Artificial Intelligence Laboratory

Technical Report

MIT-CSAIL-TR-2013-013
CBCL-312

June 18, 2013

Body-form and body-pose recognition with a hierarchical model of the ventral stream

Heejung Kim, Jeremy Wohlwend, Joel Z. Leibo, and Tomaso Poggio

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

Body-form and body-pose recognition with a hierarchical model of the ventral stream

Heejung Kim[Δ], Jeremy Wohlwend[Δ], Joel Z. Leibo[Δ,†], and Tomaso Poggio[Δ,†]

[Δ] *Center for Biological and Computational Learning, McGovern Institute for Brain Research, MIT*
[†] *To whom correspondence should be addressed: {jzleibo@mit.edu, tp@ai.mit.edu}*

Abstract

When learning to recognize a novel body shape, e.g., a panda bear, we are not misled by changes in its pose. A "jumping panda bear" is readily recognized, despite having no prior visual experience with the conjunction of these concepts. Likewise, a novel pose can be estimated in an invariant way, with respect to the actor's body shape. These body and pose recognition tasks require invariance to non-generic transformations [10, 16] that previous models of the ventral stream do not have. We show that the addition of biologically plausible, class-specific mechanisms associating previously-viewed actors in a range of poses enables a hierarchical model of object recognition to account for this human capability. These associations could be acquired in an unsupervised manner from past experience.

1 Introduction

A single object can have drastically different appearances depending on viewing conditions, e.g. viewing angle and illumination. Recognizing objects despite variability in viewing condition is one of the critical tasks of vision. In human visual experience these transformations of object appearance are hardly noticed — yet despite decades of work, the problem of how to get the same invariant recognition behavior from a machine remains unresolved.

During natural vision, images of objects undergo various transformations of their appearance. Transformations such as translation and scaling (2D affine transformations) are generic, i.e. they can be computed using only the information contained in the image itself. Other transformations such as 3D rotation and changes in illumination conditions do not have this property. They cannot be computed without additional information about the object's 3D structure or material properties. Invariance to generic transformations can be learned from experience with any objects, whereas invariance to non-generic transformations can only be acquired from experience with similar objects [10].

Many non-generic transformations are class-specific. Novel objects can be recognized invariantly to these transformations using prior knowledge of how other objects from the same class transform. For example, novel faces can be recognized invariantly to 3D rotation using features that are invariant to the 3D rotations of previously-viewed template faces.

Since the computations required to discount class-specific transformations differ for different object classes, it follows that the neural circuitry involved in computing invariance to transformations of some object classes must be separated from the circuitry involved in the analogous computation for other object classes. We have argued that this is the computational explanation for domain-specific processing modules in the ventral stream. [10, 16]

Within the ventral stream there are patches of cortex that show selective increases in BOLD response for specific classes of objects. These include regions that respond to faces—the fusiform face area (FFA), occipital face area (OFA), etc [6,7]—scenes—the parahippocampal place area (PPA) [5]—written words—the visual word form area (VWFA) [1], and bodies—the extrastriate body area (EBA) and the fusiform body area (FBA) [4, 13]. Many of these regions were shown to be necessary for recognition tasks with the objects they process by lesion studies [9, 11, 12, 18] and TMS [14, 15, 22].

We conjecture that domain-specific processing modules compute object representations that are invariant to class-specific transformations [10, 16]. To support this conjecture, we have built several hierarchical models that recognize objects of particular classes invariantly to class-specific transformations. In particular, we have studied viewpoint invariant face-identification [10], perspective-invariant scene identification [8], and font/case-invariant word recognition [2]. In the present report we describe an analogous model of body recognition and pose estimation. The analysis is purely computational but has implications for the interpretation of studies of this behavior and its associated modules – EBA and FBA.

2 Methods

2.1 Stimuli

Images of human bodies in various poses were used to train and test the model. 1408 3D object models of human body were created with DAZ 3D Studio and one 256*256 pixel greyscale image was rendered from each object with Blender.

The 1408 objects consisted of 44 differently shaped human bodies in 32 poses. The 44 bodies were either male or female, had varying degrees of body fat, muscularity, and limb proportion. The 32 poses were natural, commonly encountered poses such as waving, running, leaning, and clinging.

2.2 Task

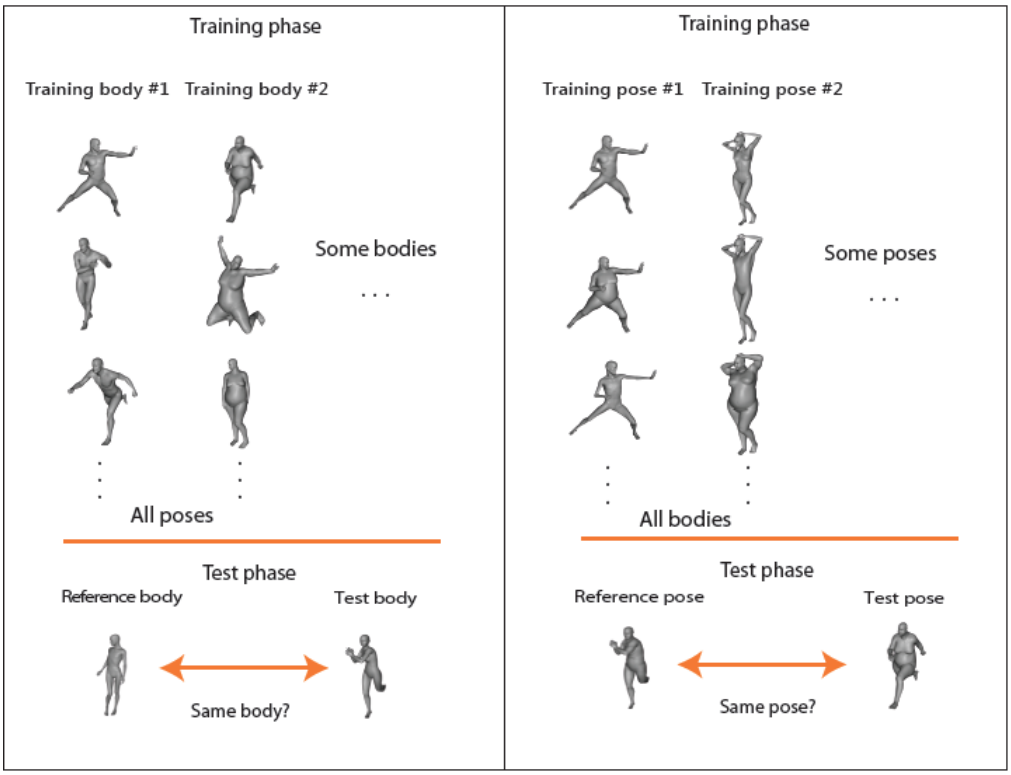

Figure 1: Left: Example images for the pose-invariant body-recognition task. The images appearing in the training phase were used as templates (in the sense of [16] and section 2.3 below). The test measures the model's performance on a same-different task in which a reference image is compared to a query image. 'Same' responses are marked correct when the reference and query image depict the same body (invariantly to pose-variation). Right: The body-invariant pose recognition task. This task is analogous to the other task except poses are used as templates instead of bodies and 'same' responses are correct if the reference and query pose match.

We modeled a same-different psychophysical test of initial invariance. A nearest-neighbor classifier ranked the similarity of a reference image to a set of testing images containing both the reference object under various transformations and distractor objects under the same transformations. None of the images used in the testing phase ever appeared in the training phase.

The pose-invariant body-recognition task requires the classifier to rank images of the same body as similar to one another despite variation in its pose. The body-invariant pose-recognition task requires the classifier to rank images of the same pose as most similar to one another despite variation in the actor.

2.3 Model

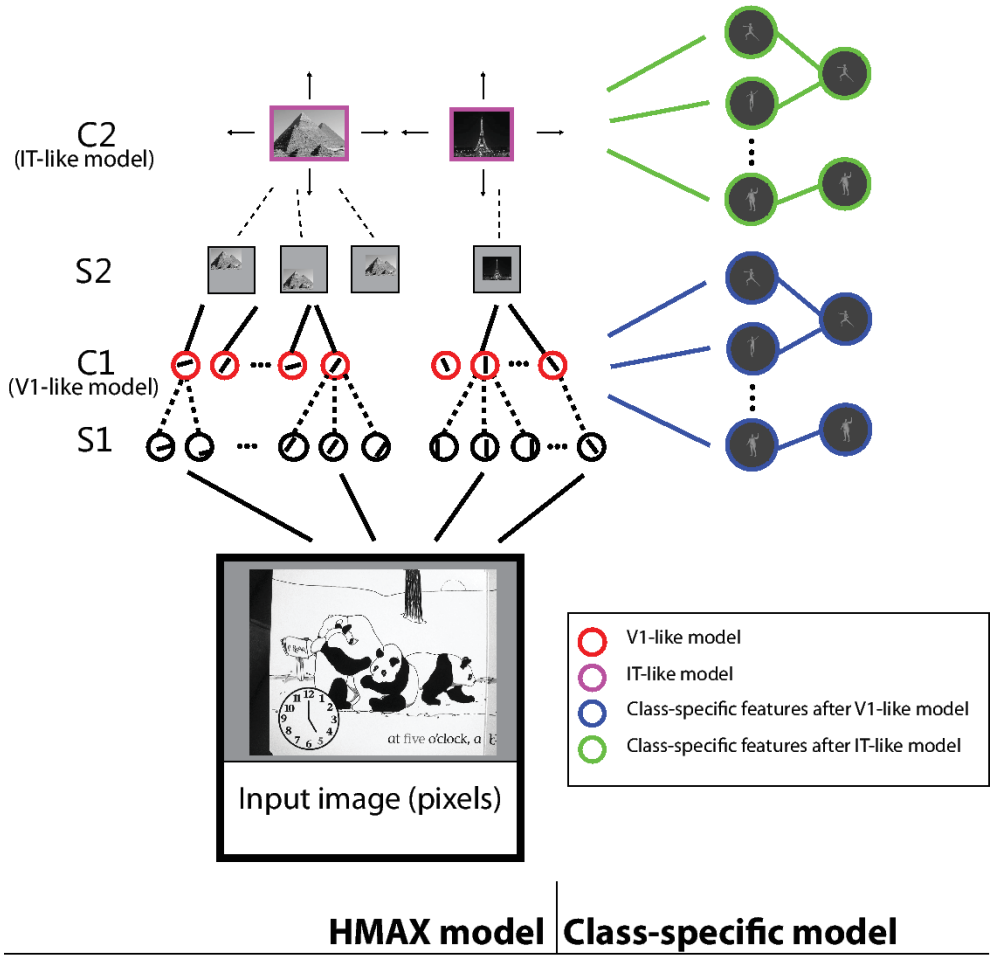

Figure 2: Illustration of two different class-specific network architectures. Blue: Class-specific model based on HMAX C1 responses. Green: Class-specific model based on HMAX C2 (global pooling).

Let $B = \{b_1, b_2, ..., b_n\}$ be a set of bodies and $P = \{p_1, p_2, ..., p_n\}$ be a set of poses. Let d be the dimensionality of the images. For the body-invariant pose recognition task we define the rendering function $t_b : P \rightarrow \mathbb{R}^d$, and analogously, $t_p : B \rightarrow \mathbb{R}^d$ for the pose-invariant body recognition task. In words, we say $t_p[b]$ renders an image of body b in pose p. In that case the argument b is the template and the subscript p indicates the transformation to be applied. Likewise, $t_b[p]$ renders an image of pose p with actor b, where p is the template and b the transformation. Our model aims to compute an invariant *signature* $\Sigma(\cdot)$. That is, a novel body b can be recognized invariantly if $\Sigma(t_{p_1}[b]) = \Sigma(t_{p_2}[b])$. Or analogously, for the body-invariant pose estimation task $\Sigma(t_{b_1}(p)) = \Sigma(t_{b_2}(p))$. In practice, signatures before and after a transformation need not actually be equal, they need only have less variance caused

by the transformation than by the differences between the objects to be recognized.

Following the theory of [16], the model pools inner products of the input image with a set of stored templates $\mathbb{T} = \{t_i(\tau_j)\}$. We use a gaussian radial basis function for this.

$$\langle x, t_i(\tau_j)\rangle = \exp\{\sigma * \sum((x - t_i(\tau_j))^2)\} \tag{1}$$

We obtain the signature vector $\Sigma : X \to \mathbb{R}^m$ by pooling the inner products of the input image with different renderings of the same template.

$$\sum(x) = \begin{pmatrix} \max(\langle x, t_1(\tau_1)\rangle, \langle x, t_2(\tau_1)\rangle, \ldots, \langle x, t_n(\tau_1)\rangle) \\ \max(\langle x, t_1(\tau_2)\rangle, \langle x, t_2(\tau_2)\rangle, \ldots, \langle x, t_n(\tau_2)\rangle) \\ \vdots \\ \max(\langle x, t_1(\tau_m)\rangle, \langle x, t_2(\tau_m)\rangle, \ldots, \langle x, t_n(\tau_m)\rangle) \end{pmatrix} \tag{2}$$

This model can be made hierarchical. It takes in any vector representation of an image as input. We investigated two hierarchical architectures built off of different layers of the HMAX model (C1 and C2b) [17].

The approach is equivalent to that of [16] and [10]. In those cases the rendering parameter is viewed as the parameter of a continuous transformation e.g., degrees rotated or distance translated; it could be interpreted as time in a training video. The rendering parameter does not have that interpretation in the present case.

3 Results

3.1 Model Performance

We evaluated the performance of two different class-specific architectures: one class-specific model operates on HMAX C1-encoded images (shown in blue), and the other is based on HMAX C2-encoded images (shown in green).

We tested each model on the same-different task of determining whether a given image has the same body or same pose as a reference image. We also tested two different layers of the generic HMAX model as a control [17]. We evaluated the performance of each model on images that were unused in the model-building phase. For the pose-invariant body recognition task, the template images were drawn from a subset of the 44 bodies—rendered in all poses. The test set contained images of 10 bodies that never appeared in the model-building phase—again, rendered in all poses. The body-invariant pose-recognition task was the other way around: the poses were split into template and testing sets and each was rendered with all bodies. We did 10 cross-validation runs with different randomly chosen training and test sets. The reported AUC was averaged over all runs.

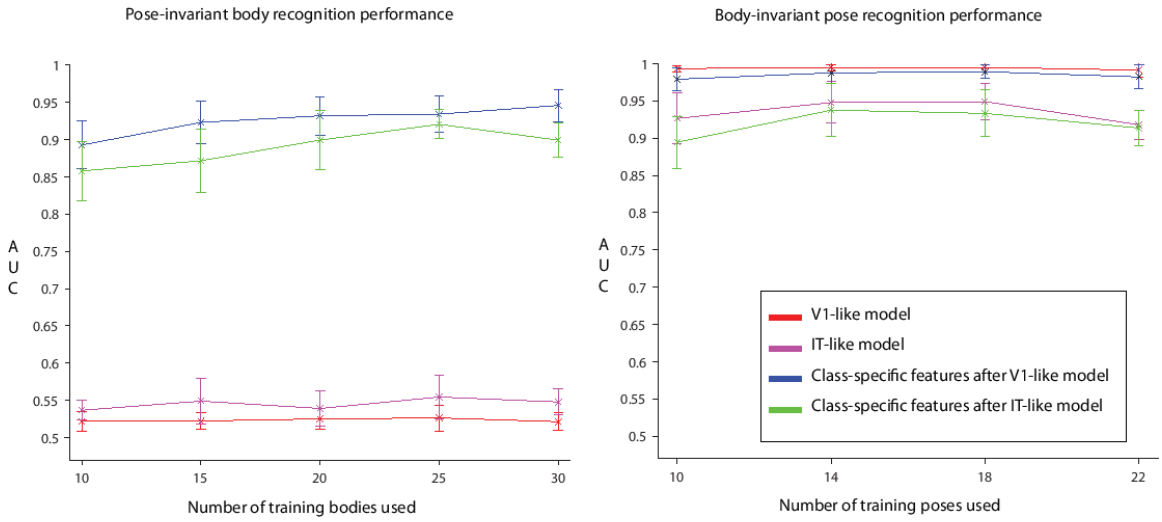

Figure 3: Model performance: area under the ROC curve (AUC) for the same-different task with 10 testing patterns (bodies or poses). Left: The X-axis indicates the number of bodies used to train the model. Bottom: The X-axis indicates the number of poses used to train the model. Performance was averaged over 10 cross-validation splits. The error bars indicate one standard deviation over cross-validation splits.

Figure 3-A shows the results on the pose-invariant body recognition task and figure 3-B shows body-invariant pose estimation results. We showed that simple methods (e.g., HMAX's C1 layer) are sufficient to achieve good pose-recognition (at least, in our simplified setting). However, on the body-recognition task, the HMAX models we tested perform almost at chance. The addition of the class-specific mechanism significantly improves performance on this difficult task. That is, models without class-specific features were unable to perform the pose-invariant body recognition task while class-specific features enabled good performance on this difficult invariant recognition task.

3.2 Human performance at pose-invariant body-identification

We tested 8 human observers on the same pose-invariant body-identification task. Subjects were seated in front of an 85 Hz monitor in a dark room. We ran 5 blocks of 120 trials with 5 minute breaks between blocks. Each trial began with a black fixation cross appearing in the middle of a grey background for 1s. The subjects then saw two images, first the "reference" image, followed by a second "query" image. After the second image disappeared, the subject indicated, by a button press, whether or not the reference and query images depicted the same body (invariantly to the body's pose). Each image presentation was followed by a mask image (a random binary pixel noise image) displayed for 50 ms. The reference image was always displayed for 60ms. We varied the presentation time of the query image in order to investigate task performance as a function of post-stimulus processing time. The correct response was "same" on 50% of the trials; all trial types were intermixed.

Figure 4 shows that human observers could perform this task at their performance ceiling ($\sim 70-75\%$) by 48ms after the onset of the query image. We argue, along the same lines as [17,21], that this pattern

of results suggests the brain is operating in a feedforward processing mode while subjects perform this task.

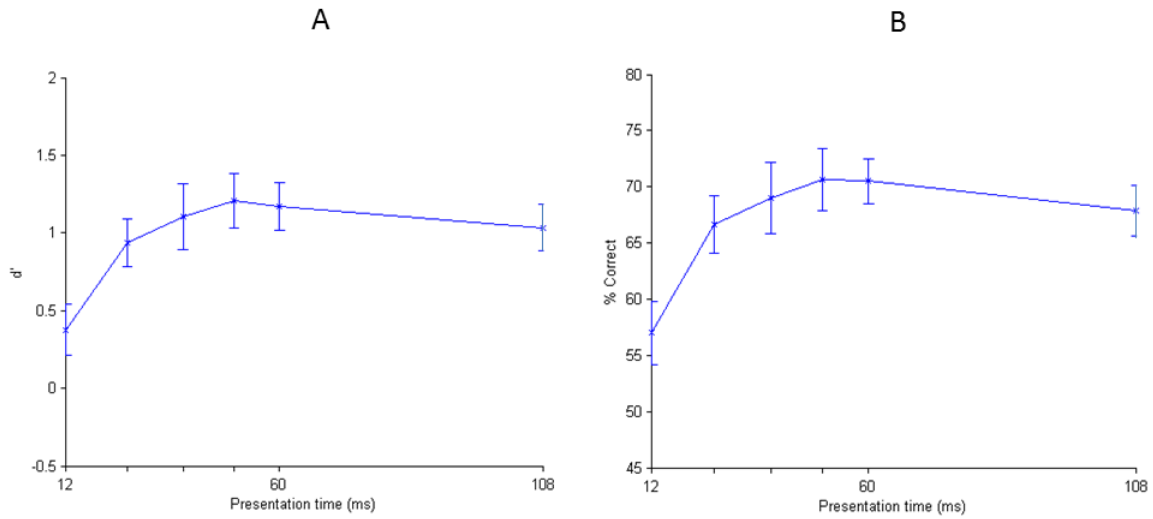

Figure 4: Panel A: Discriminability (d'). Panel B: Accuracy (% correct). Both X-axes indicate the presentation time of the second image (the query). N = 8 subjects, error-bars are standard error of the mean.

3.3 Body-size judgment and misjudgment: a model of anorexia nervosa?

Suchan et al. reported a reduction of gray matter density in the left EBA of women with anorexia nervosa (AN) as well as a decrease in connectivity between their left EBA and FBA. Both of these neural abnormalities were negatively correlated with performance on a body-size judgment task [19,20]. Additionally, Urgesi et al. also reported body-specific perceptual abnormalities in AN patients [23]. While none of these studies establish causality—it is unclear if the perceptual and brain abnormalities cause the eating disorder or the other way around—nevertheless, it is interesting to consider how a lesion in the body-processing module could affect performance on body-size judgments in the context of our computational model.

The procedure for these experiments was identical to the pose-invariant body recognition task except in this case the goal was to generalize over pose and over bodies with identical sizes. Note: "size" here corresponds to a subjective human judgment (made by the researchers); see figure 5-A for example images.

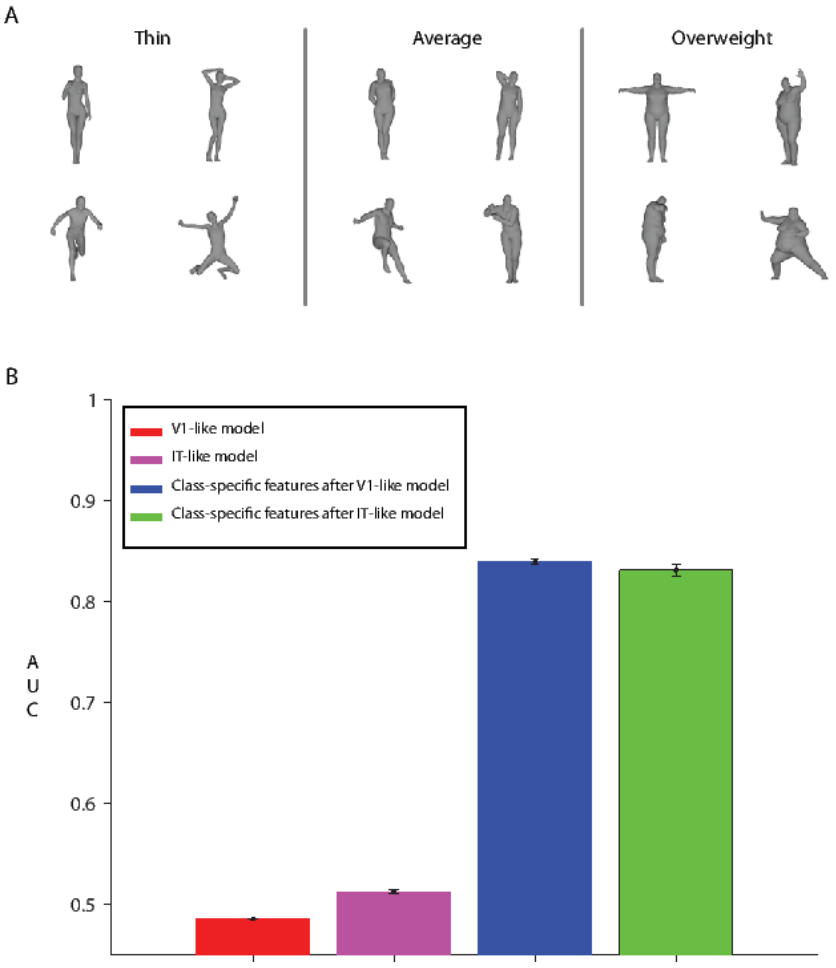

Figure 5: (A). Example images from each of the three body-size categories. (B). Concordance of the model's body-size judgments with the three categories. Pose-pooled signatures provide the dissimilarity metric for this test (blue and green bars). No supervised training was performed. The experiment was repeated 8 times using different bodies as templates. AUC was computed as in the same-different task, in this case, the model was marked correct whenever the two testing images were in the same size category.

Figure 5-B compares the concordance of the pose-invariant model's body-size judgments to human judgments. This simulation used the same class-specific models as in the pose-invariant body recognition task (pooling only over pose). These preliminary results suggest that the class-specific features for the pose-invariant body-recognition task also induce a certain similarity structure over the "body-size" dimension which is in accord with typical human judgments and different from the judgments of AN patients.

These results on body-size judgments should not be interpreted as a strong endorsement of the premature-at-best claim that AN (or related disorders) arise from a perceptual abnormality. They merely show a

potential mechanism by which an aspect of the behavior of patients with AN could arise from selective damage to a perceptual system. They do not even say anything about causality: in our model, we could regard the loss of class-specific cells as arising due to damage to the EBA or FBA; or alternatively, we could regard it as due to degeneration of the cells in those regions due to their lack of use, possibly due to a prior aversion to looking at body-stimuli. The contribution of these simulations on body-size judgement to the broader field of research on AN and related disorders is only to raise the possibility that body-perception abnormalities in these disorders may stem from selective damage to a body-processing specific module in the perceptual system. It should be considered in the context of the recent neuroimaging results showing abnormalities in these brain regions [19, 20, 24].

4 Discussion

Downing and Peelen (2011) argued that the EBA and FBA "jointly create a detailed but cognitively unelaborated visual representation of the appearance of the human body". These are perceptual regions—they represent body shape and posture but do not explicitly represent high-level information about "identities, actions, or emotional states" (as had been claimed by others in the literature cf. commentaries on [3] in the same journal issue). The model of body-specific processing we present here is broadly in agreement with this view of EBA and FBA's function. It computes, from an image, a body-specific representation that could underlie many further computations e.g. action recognition, emotion recognition, etc.

We previously conjectured that modularity in the ventral stream arises because of the need to discount class-specific transformations [10, 16]. We showed that humans can accurately perform the task of pose-invariant body-recognition with very short presentation times (performance ceiling is reached by 50ms). Standard feedforward models of the ventral stream cannot account for this human ability without the addition of cells that pool over pose transformations. When these class-specific cells are included, the task becomes relatively easy. This observation suggests that the underlying computational reason that the brain separates the processing of images of bodies from the processing of other images is the need to recognize specific people invariantly to their pose.

5 Acknowledgements

This report describes research done at the Center for Biological and Computational Learning, which is in the McGovern Institute for Brain Research at MIT, as well as in the Dept. of Brain and Cognitive Sciences, and which is affiliated with the Computer Sciences and Artificial Intelligence Laboratory (CSAIL). This research was sponsored by grants from DARPA (IPTO and DSO), National Science Foundation (NSF-0640097, NSF-0827427), AFSOR-THRL (FA8650-05-C-7262). Additional support 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.

References

[1] L. Cohen, S. Dehaene, and L. Naccache. The visual word form area. *Brain*, 123(2):291, 2000.

[2] D. R. Deo, J. Z. Leibo, and T. Poggio. A model of invariant text recognition in the ventral stream (in preparation).
[3] P. Downing and M. Peelen. The role of occipitotemporal body-selective regions in person perception. *Cognitive Neuroscience*, 2(3-4):186–203, 2011.
[4] P. E. Downing, Y. Jiang, M. Shuman, and N. Kanwisher. A cortical area selective for visual processing of the human body. *Science (New York, N.Y.)*, 293(5539):2470–3, Sept. 2001.
[5] R. Epstein and N. Kanwisher. A cortical representation of the local visual environment. *Nature*, 392(6676):598–601, 1998.
[6] N. Kanwisher, J. McDermott, and M. M. Chun. The fusiform face area: a module in human extrastriate cortex specialized for face perception. *The Journal of Neuroscience*, 17(11):4302, 1997.
[7] N. Kanwisher and G. Yovel. The fusiform face area: a cortical region specialized for the perception of faces. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 361(1476):2109, 2006.
[8] E. Y. Ko, J. Z. Leibo, and T. Poggio. A hierarchical model of perspective-invariant scene identification. Washington DC, 2011. Society for Neuroscience (486.16/OO26).
[9] A. Leff, G. Spitsyna, G. Plant, and R. Wise. Structural anatomy of pure and hemianopic alexia. *Journal of Neurology, Neurosurgery & Psychiatry*, 77(9):1004–1007, 2006.
[10] J. Z. Leibo, J. Mutch, and T. Poggio. Why The Brain Separates Face Recognition From Object Recognition. In *Advances in Neural Information Processing Systems (NIPS)*, Granada, Spain, 2011.
[11] V. Moro, C. Urgesi, S. Pernigo, P. Lanteri, M. Pazzaglia, and S. M. Aglioti. The neural basis of body form and body action agnosia. *Neuron*, 60(2):235–46, Oct. 2008.
[12] M. Moscovitch, G. Winocur, and M. Behrmann. What is special about face recognition? Nineteen experiments on a person with visual object agnosia and dyslexia but normal face recognition. *Journal of Cognitive Neuroscience*, 9(5):555–604, 1997.
[13] M. V. Peelen and P. E. Downing. Selectivity for the human body in the fusiform gyrus. *Journal of neurophysiology*, 93(1):603–8, Jan. 2005.
[14] D. Pitcher, L. Charles, J. T. Devlin, V. Walsh, and B. Duchaine. Triple dissociation of faces, bodies, and objects in extrastriate cortex. *Current biology : CB*, 19(4):319–24, Feb. 2009.
[15] D. Pitcher, V. Walsh, G. Yovel, and B. Duchaine. TMS evidence for the involvement of the right occipital face area in early face processing. *Current Biology*, 17(18):1568–1573, 2007.
[16] T. Poggio, J. Mutch, F. Anselmi, J. Z. Leibo, L. Rosasco, and A. Tacchetti. The computational magic of the ventral stream: sketch of a theory (and why some deep architectures work). *MIT-CSAIL-TR-2012-035*, 2012.
[17] T. Serre, A. Oliva, and T. Poggio. A feedforward architecture accounts for rapid categorization. *Proceedings of the National Academy of Sciences of the United States of America*, 104(15):6424–6429, 2007.
[18] B. Sorger, R. Goebel, C. Schiltz, and B. Rossion. Understanding the functional neuroanatomy of acquired prosopagnosia. *NeuroImage*, 35(2):836–52, Apr. 2007.
[19] B. Suchan, M. Busch, D. Schulte, Grönermeyer, Dietrich, S. Herpertz, and S. Vocks. Reduction of gray matter density in the extrastriate body area in women with anorexia nervosa. *Behavioural Brain Research*, 206(1):63–67, 2010.
[20] B. Suchan, D. Soria Bauser, M. Busch, D. Schulte, Grönemeyer, Dietrich, S. Herpertz, and S. Vocks. Reduced connectivity between the left Fusiform Body Area and the Extrastriate Body Area in Anorexia Nervosa is associated with body image distortion. *Behavioural Brain Research*, 241(1):80–85, 2012.
[21] S. Thorpe, D. Fize, and C. Marlot. Speed of processing in the human visual system. *Nature*, 381(6582):520–522, 1996.
[22] C. Urgesi, G. Berlucchi, and S. M. Aglioti. Magnetic stimulation of extrastriate body area impairs visual processing of nonfacial body parts. *Current Biology*, 2004.
[23] C. Urgesi, L. Fornasari, L. Perini, F. Canalaz, S. Cremaschi, L. Faleschini, M. Balestrieri, F. Fabbro, S. M. Aglioti, and P. Brambilla. Visual body perception in anorexia nervosa. *International Journal of Eating Disorders*, 45(4):501–511, 2012.
[24] S. Vocks, D. Schulte, M. Busch, Grönermeyer, Dietrich, S. Herpertz, and B. Suchan. Changes in neuronal correlates of body image processing by means of cognitive-behavioural body image therapy for eating disorders: a randomized controlled fMRI. *Psychological Medicine*, 41(88):1651–1663, 2011.