

1 Neural Construction of 3D Medial Axis from the Binocular Fusion of 2 2D MAs

3

4 Wei Qiu, Yasuhiro Hatori, and Ko Sakai*

5 Department of Computer Science, University of Tsukuba

6 1-1-1 Tennodai, Tsukuba, Ibaraki 305-8347, Japan

7 *: corresponding author: sakai@cs.tsukuba.ac.jp, +81-29-853-5348

8

9

10 Abstract

11 The perceptual constancy of shape, including view invariance, is an amazing property of the
12 visual system. Cortical representation by the medial axis (MA) is an attractive candidate for
13 maintaining the constancy of a wide range of arbitrary shapes. Recent physiological studies
14 have reported that neurons in the primary visual cortex (V1) show a response to
15 two-dimensional (2D) MAs, and those in the inferior temporal cortex (IT) are selective to
16 three-dimensional (3D) MAs. However, little is known about the neural mechanisms
17 underlying the transformation of 2D to 3D MAs. As a first step toward investigating the
18 cortical mechanism, we have proposed as a hypothesis that a pair of monocular 2D MAs is
19 fused to generate a 3D MA. We examined the computational plausibility of the hypothesis;
20 specifically, whether an energy-based fusion model is capable of generating 3D MAs. We
21 generated blob-like, physiologically plausible 2D MAs, and used a standard energy model to
22 detect the disparity between a pair of 2D MAs. The model successfully generated 3D MAs for
23 a variety of objects that included typical shape characteristics. A reconstruction test showed
24 that the computed 3D MAs captured the essential structure of the objects with reasonable
25 accuracy and view invariance. These results indicate that the fusion of monocular blob-like
26 2D MAs is capable of generating a reasonable 3D MA within the framework of the energy
27 model.

28

29

30 Keyword: vision, shape perception, cortical representation, medial axis, stereo

31 1 Introduction

32

33 Robust perception of the shape of objects is an amazing property of the visual system.
34 Although the view and size of an object on a retinal image change dramatically as we see the
35 object from different directions and distances, our visual system perceives a stable, invariant
36 shape for the object. The representation of shape in the visual cortex should play a crucial
37 role in realizing such invariance in shape perception. An object-centered representation that
38 describes shape as a spatial arrangement of parts has been supported widely by
39 psychological and physiological studies [1,2,3], as it has the ability to establish the
40 perceptual constancy of shape, including view and distance invariance. The medial axis (MA)
41 is considered suitable for a parts-based representation among theorists [4,5]. MA
42 representation encodes each part of the object with a medial line that is derived from the
43 local symmetry of the part. This representation, based on an object-centered coordinate, is
44 independent of view and capable of describing shape efficiently using two types of
45 parameters: the spatial arrangement and relative length of the axes corresponding to the
46 parts [4,6,7]. MA is an attractive candidate for the cortical representation of shape, as a
47 robust and efficient coding scheme [8].

48 Recently, Hung et al. showed that a number of neurons in the inferior temporal
49 cortex (IT) encode three-dimensional (3D) MA configurations, supporting the idea that the
50 MA plays a critical role in the representation of shape in the ventral pathway [9]. IT has been
51 reported to encode the 3D structure of shape [10,11], but little was known about the
52 representation scheme for 3D shapes. The selectivity for 3D MA configurations reported
53 recently in IT has provided crucial direct evidence to support MA coding for the cortical
54 representation of shape. A recent fMRI study has also reported the cortical representation of
55 MA structure in the ventral stream [12]. However, the computational processes that
56 constitute the 3D MA along the ventral pathway remain unknown. One of the keys to
57 understanding these processes lies in the lower cortex: cells in the primary visual cortex (V1)
58 show strong responses to the MA of a textured figure [13,14]. Computational studies have
59 shown that the MA response in V1 can be generated by simultaneous arrival of traveling
60 spikes that are initiated by nearby V1 cells [14], or from onset synchronization of
61 border-ownership (BO)-selective cells in V2 [15,16]. These computational studies have also

62 reported that the generated MA encodes arbitrary two-dimensional (2D) shapes. These
63 studies note that the MAs were not like thin skeletons as previous studies have assumed,
64 but rather, the MAs were elongated blobs with spatial extent. This blob-like MA is expected
65 to be robust for 3D construction. Because the structure of skeletons is sensitive to the
66 direction of view (binocular difference) and noise in the contours, small changes in view and
67 contour dramatically alter the structure of skeleton-like MAs, leading to erroneous stereo
68 matching. However, blob-like MAs are expected to be insensitive to such changes [17].
69 Investigating the fusion of blob-like MAs rather than conventional skeleton-like MAs is
70 essential. The intermediate areas of the ventral visual pathway such as V4 are known to play
71 a crucial role in the binocular fusion of object shapes [18,19]. A certain translation function
72 that takes place along the ventral pathway may contribute to the construction of the 3D MA
73 observed in IT from the 2D MAs observed in V1.

74 We investigated the cortical mechanisms underlying the construction of 3D-shape
75 representation, by focusing on blob-like 2D MAs and their fusion along the ventral pathway.
76 Fusion of 2D MAs based on their disparity is a plausible candidate mechanism for filling the
77 gap between the 2D MA in the primary cortex and the 3D MA in the higher cortex. It is
78 conceivable that the 2D MAs resulting from the left and right retinal images are fused in an
79 intermediate-level area by a process based on disparities in the 2D MAs, thereby establishing
80 a 3D MA in IT. An alternative mechanism for the construction is that the MA responses in V1
81 are binocular with absolute disparity, and are thus “3D MA segments.” The 3D MA segments
82 in V1 would then be integrated along the visual pathway to establish a global 3D MA with
83 relative disparities in IT. Although a number of V1 cells are selective to the binocular
84 disparity of contours, it is not at all certain whether cells responding to MAs are selective for
85 the binocular disparity of the local MA. V1 cells could respond to the depth of contours, but
86 not necessarily to that of the MA. Specifically, the depths of both sides of an object as well
87 as its MA are generally different. This concept is illustrated by a cuboid with a different
88 depth for each side of the object; for example, the left side is nearer and the right side is
89 farther (see Figure 1A). Although the depths of these sides can be determined correctly, the
90 depth of the MA is inherently ambiguous; the MA could be located anywhere between the
91 two sides and there is no way to determine its depth from the depth of the sides. On the
92 other hand, in the former case involving 2D MAs, the local disparities between the 2D MAs
93 could be integrated without ambiguity (see Figure 1B). This idea appears to be consistent

94 with the tuning of three-dimensional orientation in the macaque V4 [19]. In the present
95 study, we focused on the fusion of monocular 2D MAs that are formed in V1, and are fused
96 along the ventral pathway based on the disparities between the axes, to generate a 3D MA
97 in IT.

98 Physiological evidence for the generation process of a 3D MA has not been available.
99 As a first step toward investigating our hypothesis, we conducted computational studies to
100 determine whether the fusion of monocular, blob-like (physiologically plausible) 2D MAs is
101 capable of generating a 3D MA, and how accurately this method would work. Specifically, we
102 constructed a fusion model based on a standard energy model [20] that is thought to
103 capture the essential functions of physiological properties in early- to intermediate-level
104 visual areas. We examined whether the model is capable of generating a correct 3D MA, and
105 whether the computed 3D MA captures the essential structure that is sufficient for the
106 reconstruction of a 3D shape. Our simulation results showed that the model was capable of
107 generating 3D MAs for a variety of shapes including those of natural objects. The results also
108 showed that the reconstruction of 3D shapes based on the computed 3D MAs was successful,
109 with similar levels of accuracy for various shapes with different degrees of shape complexity,
110 which is one of the most remarkable features of the visual system. Furthermore, we tested
111 view invariance of the model in terms of the reconstruction error. Similar reconstruction
112 errors were observed for images from different views, suggesting that the representation of
113 a 3D MA from the fusion of 2D MAs has invariance to rotation. View invariance has been
114 reported in MA-selective cells in IT [9]. Our results indicate that the energy-based fusion of
115 monocular blob-like 2D MAs is capable of generating a 3D MA with robustness in terms of
116 shape complexity and view invariance. Therefore, the generation of a 3D MA from the fusion
117 of 2D MAs is a plausible candidate for the cortical mechanisms underlying the
118 representation of 3D shape.

119

120 2 The model

121

122 To investigate whether the fusion of physiologically plausible 2D MAs is capable of
123 generating a correct 3D MA, and whether the computed 3D MA captures the structure

124 essential for the reconstruction of 3D shape, we constructed a computational model and
125 conducted simulations. An outline of the model is illustrated in Figure 1. The model is
126 composed of two stages: (i) the detection of monocular 2D MAs based on the distances from
127 surrounding contours, and (ii) the generation of a 3D MA from the disparities between the
128 two 2D MAs (Figure 1B). A unit in the first stage computes the distances between the unit
129 and the points on the contours surrounding the unit, and evaluates how much the unit is
130 similarly distant from the surrounding contours by taking pairwise differences between the
131 distances. Units with small differences (similar distances) tend to be located around local
132 symmetry axes, thus their locations are highly likely a part of the 2D MA. The second stage
133 fuses a pair of 2D MAs using a standard energy model to generate a 3D MA. Note that the
134 model includes neither the representation nor the reconstruction of a 3D object. We
135 conducted the reconstruction in the Results section solely for the evaluation of the
136 computed 3D MA.

137

138 2.1 The detection of 2D MA

139 A computational study by Hatori and Sakai has shown that onset synchronization of
140 BO-selective cells appears to generate V1 activities in response to 2D MAs [16,21].
141 BO-selective cells on the contour of a figure depolarize if the figure is located on their
142 preferred side [22]. The spikes from BO-selective cells, which are initiated at the same time
143 and travel at the same speed, reach the center of the figure at the same time. Temporal
144 integration of the traveling spikes would result in strong responses of cells located at the
145 center of the figure and along the axes of local symmetry, generating the V1 activity
146 corresponding to the MA. The magnitude of the activity depends on how much the cell is
147 similarly distant from the contours. Taking into account the essence of their idea, the
148 present model computes the possibility of being a 2D MA based on distances from the
149 surrounding contours. Although Hatori's model was capable of processing multiple objects,
150 we limited our model to dealing with a single object for the sake of simplicity. We computed
151 an index that describes how much a cell is similarly distant from the contours. If the value of
152 the index exceeds a certain threshold, we consider it as an indication of the MA.

153 The input to the model was a pair of stereo images with a spatial resolution of $200 \times$
154 200 pixels (considered as 5×5 degrees of visual angle). To evaluate the similarity of
155 distances from nearby contours, we measured the distance, $dist(p, q_i)$, between a point

156 within a figure, p , and every point on the contour, q_i :

$$157 \quad \text{dist}(p, q_i) = \| p - q_i \| , \quad \text{Eq. 1}$$

158 where $\| \cdot \|$ represents the Euclidean distance between the two points. The distances
159 between p and q_i were measured for every 5° (Figure 1C):

$$160 \quad q \in \mathbf{Q}, \quad \text{where } \mathbf{Q} = \{q_i | \angle q_i p q_{i+1} = 5^\circ\} . \quad \text{Eq. 2}$$

161 The equidistance index, $E(p)$, is given by a mean of the pairwise differences in the distance
162 between p and q_i :

$$163 \quad E(p) = \frac{1}{|\mathbf{Q}|} \left\{ \sum_{i=1}^{|\mathbf{Q}|} \sum_{j=i+1}^{|\mathbf{Q}|} s(\text{dist}(p, q_i) - \text{dist}(p, q_j)) \right\} , \quad \text{Eq. 3}$$

164 where $|\mathbf{Q}|$ indicates the number of the elements of \mathbf{Q} . To reproduce the nonlinearity of
165 neural responses, we introduced a sigmoidal function for $s(\cdot)$:

$$166 \quad s(x) = 1 - \frac{1 - e^{-\frac{x}{w}}}{1 + e^{-\frac{x}{w}}} , \quad \text{Eq. 4}$$

167 where a constant, w , controls the rate of sigmoidal decay. Throughout the simulations, we
168 set w to 6 so that the decay is 10% if the difference in distance is 18 pixels.

169 We computed the equidistance index for all points within a figure. A unit with a
170 higher index value is likely to be located around the local axes of symmetry. We consider
171 that units with an index value higher than or equal to a threshold, $E_{threshold}$, correspond to
172 the MA. Therefore, an index to represent how much a unit is likely to be part of the MA is
173 given by the equidistance index with a threshold, $E_{threshold}$:

$$174 \quad \mathbf{MA_index}(p) = \begin{cases} E(p) , & \text{if } E(p) \geq E_{threshold} , \\ 0 & , \text{ otherwise } , \end{cases} \quad \text{Eq. 5}$$

175 We chose empirically $E_{threshold} = 0.26$. This value is crucial for the formation of MA.
176 Although this threshold could be fixed for all stimuli, we chose to fine-tune the value within
177 15% because details of the formation of 2D MAs are not the focus of our study. To avoid an
178 abrupt distribution of the MA, we introduced Gaussian smoothing to **MA_index**:

$$179 \quad \mathbf{MA}(p_x, p_y) = (\mathbf{MA_index} * \mathbf{Gauss})(p_x, p_y) , \quad \text{Eq. 6}$$

180 where (p_x, p_y) is the spatial position of a point p , and $*$ and \mathbf{Gauss} represent convolution
181 and a 2D Gaussian with $\sigma_x = \sigma_y = 2$ pixels, respectively. The optimal size of the Gaussian

182 could be different among objects depending on their spatial extent. However, our test
 183 showed that the computed MAs were barely sensitive within the range of $3x\sigma$. As previously
 184 noted, we define a 2D MA as a set of points that are located nearly equidistant from
 185 surrounding contours. Therefore, our 2D MA is a fat region with spatial extent, which is
 186 distinct from an engineering MA that is defined by skeletons. A 2D MA was computed for
 187 each ocular image. A binocular pair of 2D MAs was used to generate a 3D MA as described in
 188 the next section.

189

190 2.2 The detection of 3D MA

191 To obtain a 3D MA from a pair of monocular 2D MAs, we computed disparities between the
 192 two axes. Figure 1D shows a diagram of the computation. We used a standard energy model
 193 for binocular disparity [23]. We assumed that a fusion mechanism similar to the energy
 194 model might take place along the ventral pathway probably in intermediate-level visual
 195 areas. The model consists of a cascade of simple- and complex-type cells with half-wave
 196 rectification. A pair of 2D MAs was used as input, and the disparities were determined as
 197 described below.

198 A model complex cell consists of a pooling of four quadrature pairs of model simple
 199 cells (Figure 1D (i)) with a particular binocular disparity. The response of a pair of simple cells,
 200 $O^1(x, y)$ was computed by the convolution of a monocular image (2D MA) and an oriented
 201 Gabor function with a particular orientation, phase, and disparity. We summed up the
 202 responses for the right and left images, and passed them through a half-wave rectification
 203 step (Figure 1D (ii)):

$$204 \quad O^1(x, y) = \begin{cases} \mathit{sum}_{\mathit{simple}}(x, y) , & \text{if } \mathit{sum}_{\mathit{simple}}(x, y) \geq 0 , \\ 0 & , \text{ otherwise } , \end{cases} \quad \text{Eq. 7}$$

205 where

$$206 \quad \mathit{sum}_{\mathit{simple}}(x, y) = (\mathit{nMA}_{\mathit{left}} * \mathit{Gabor}_{\mathit{left}})(x, y) + (\mathit{nMA}_{\mathit{right}} * \mathit{Gabor}_{\mathit{right}})(x, y). \quad \text{Eq. 8}$$

207

208 $\mathit{nMA}_{\mathit{left}}$ and $\mathit{nMA}_{\mathit{right}}$ represent a normalized 2D MA for the left and right images,
 209 respectively. $\mathit{MA}(x, y)$ of an image was normalized to its maximum value so that $\mathit{nMA}_{\mathit{left}}$
 210 and $\mathit{nMA}_{\mathit{right}}$ range between 0 and 1. $\mathit{Gabor}_{\mathit{left}}$ and $\mathit{Gabor}_{\mathit{right}}$ represent the
 211 oriented receptive field in V1 for the left and right images, respectively. A detailed

212 description of $Gabor_{left}$ and $Gabor_{right}$ is given in Appendix A.

213 The response of a model complex cell, $O^2(x, y)$, was given by the summation of
 214 squared outputs of the four quadrature pairs of the model simple cells, O^{1, ϕ_i} (Figure 1D
 215 (iii)):

$$216 \quad O^2(x, y) = \sum_{i=1}^4 \left(O^{1, \phi_i} \right)^2. \quad \text{Eq. 9}$$

217 To establish orientation-invariant selectivity, we pooled three types of complex cells with
 218 distinct optimal orientations ($\theta = 0, \pi/6, \pi/3$) by using a winner-take-all mechanism.
 219 Although the three channels for orientation appear fewer than those in V1, we chose three
 220 for the sake of simplicity. The response of the winner complex cell with disparity, ψ_j , is
 221 given by (Figure 1D (iv)):

$$222 \quad O^{3, \psi_j}(x, y) = \max_k \left(O^{2, \theta_k}(x, y) \right). \quad \text{Eq. 10}$$

223 The model has 11 distinct disparity channels ($j = 1-11$), resulting in the range of disparity
 224 between 0 and 10 pixels. The disparity of a location is given by a winner-take-all mechanism,
 225 that is, the preferred disparity of a cell with the strongest response among the 11 disparity
 226 channels is chosen as the disparity of the location (Figure 1D (v)):

$$227 \quad \mathbf{disp}(x, y) = \underset{\psi_j}{\operatorname{argmax}} \left\{ O^{3, \psi_j}(x, y) \right\}. \quad \text{Eq. 11}$$

228 We defined horizontal disparity as:

$$\mathbf{disparity}(x, y) = \begin{cases} ((\mathbf{disp} * \mathbf{Gauss})(x, y), & \\ \quad \quad \quad \text{for } (x, y | \mathbf{nMA}_{right}(x, y) > N_{threshold}) \quad , & \\ -1 \quad \quad \quad , \quad \text{otherwise} \quad \quad \quad , & \end{cases}$$

229 Eq. 12

230 where $N_{threshold}$ indicates the threshold for eliminating unnecessary smoothing. We set
 231 $N_{threshold}$ to 0.1, however, the results were similar when the threshold is less than or equal
 232 to 0.3. $\mathbf{Gauss}(x, y)$ represents the Gaussian for smoothing with $\sigma_x = \sigma_y = 3$ pixels. The
 233 optimal size (σ) of the Gaussian could depend on the size of an object. However, our test
 234 showed that the size of the Gaussian was relatively insensitive to the results; an enlargement
 235 of 50% did not alter the results. The relation between the disparity of a location,

236 ***disparity***(x, y), and the depth of 3D MA, ***depth***(x, y), is given by:

$$\mathbf{depth}(x, y)^2 - \left(\frac{\mathbf{delta}}{\mathbf{disparity}(x, y)/f} + 2 * \mathbf{fix} \right) * \mathbf{depth}(x, y) + \mathbf{fix}^2 = 0 ,$$

237 for ***disparity***(x, y) > 0 , Eq. 13

238 ***depth***(x, y) = 0, for ***disparity***(x, y) = 0,

239 where f is the focal length (5 cm; 142 pixels) and \mathbf{delta} is the distance between the two
240 eyes (8 cm; 227 pixels). \mathbf{fix} is the distance between a fixation point and the frontal plane
241 including the eyes. The nearest point of an object was chosen as the fixation point, and its
242 depth was considered zero. The depth of 3D MA in the model is given by:

$$\mathbf{MA}_{\mathbf{depth}}(x, y) = \frac{\mathbf{depth}(x, y)}{r} ,$$

244 for (x, y | ***disparity***(x, y) > 0) , Eq. 14

245 where r represents the ratio between the size of the real object and its projection onto the
246 retina (image).

247 To evaluate the model, we reconstructed the shape from the computed 3D MA, as
248 described in the Results section. For the reconstruction, we needed the distances between
249 the MA and the surrounding contours as well as the location of the MA. Because the model
250 does not compute the distances, we preserved the distances between the 2D MA and the
251 contour of the object in the right image for the purpose of evaluation. This procedure
252 assures consistency and objectivity in the determination of the distances, and adequately
253 evaluates the location of the MA.

254

255 3 Results

256

257 We constructed a computational model that generates a 3D MA from the fusion of
258 physiologically plausible 2D MAs, based on a standard energy model [20] that is thought to
259 capture the essential functions of physiological properties in early- to intermediate-level
260 visual areas. We examined whether the model is capable of generating a correct 3D MA, and
261 whether the computed 3D MA is adequate for the reconstruction of a 3D shape. The model

262 has two novel characteristics: (1) the 2D MA is defined as a set of points that are nearly
263 equidistant from surrounding contours, thus, our 2D MA has a spatial extent, unlike a
264 skeleton as defined in engineering; (2) we detected the binocular disparities between such
265 “fat” 2D MAs using an energy model. We performed the simulations of the model with a
266 variety of 3D objects that included distinct features of shape. Firstly, we present the results
267 of examples with elementary geometric features such as a capsule and a cuboid. Secondly,
268 we present the results for typical geometric features, such as a variation in thickness and a
269 bend, together with other complex features. We also present the results for pairs of real
270 images. For the evaluation of the computed MA, we reconstructed a 3D shape based on the
271 MA, and computed the reconstruction accuracy. To thoroughly test the reconstruction of the
272 3D shape, we examined the reconstruction error using three criteria: depth from the eyes,
273 3D shape (relative depth), and the shape of the 2D projection with respect to the eyes
274 (comparable with the retinal images). We also present the results for testing view invariance
275 of the computed MA.

276

277 3.1 The proposed 2D MA

278 Retinal images of an object can be noisy for various reasons, such that contours of an object
279 might be deformed. However, our visual system is capable of generating a stable percept of
280 the object’s shape. The representation of shape in the cortex appears to be robust with
281 respect to noise on the contour. In contrast, the skeletal representation that is used in
282 engineering is sensitive to noise on the contour. An example is given in Figure 2, which
283 shows MAs and their reconstruction, with and without noise. In Figure 2A, the top and
284 bottom panels show the rectangles without and with noise, respectively. Here, we
285 introduced two notches as contour noise. The two engineering MAs for the rectangles with
286 and without noise appear very different (correlation = 0.77), as shown in Figure 2B (left). The
287 change in the MA structure that is caused by slight noise often produces considerable
288 differences in MAs between the left and right images, which could be a major reason that
289 binocular fusion of the engineering MA is difficult. On the other hand, the physiological MA
290 appears to be stable with respect to noise, and produces a robust structure in the presence
291 of noise.

292 To demonstrate the insensitivity of a physiologically plausible MA with respect to
293 contour noise, we computed the MAs for the same two stimuli used above, with and

294 without noise, and compared the two MAs. Figure 2C (left panels) shows the computed 2D
295 MAs. The MAs for the object with and without noise were very similar (correlation = 0.99),
296 indicating that the physiologically plausible MA produces a stable structure insensitive to
297 noise on the contour. To demonstrate the accuracy in the reproduction of the original image
298 from the physiologically plausible 2D MAs, we reconstructed the object shape from the
299 computed MA. The reconstruction was conducted by placing circles for all points on the MA,
300 with the radius of the circles equal to the distance to the nearby contour [16]. Figure 2C
301 (right) shows the reconstructed shape from the computed physiologically plausible MAs.
302 Although the reconstruction is not as ideal as that from the engineering MAs (the
303 reconstruction errors (see [16] eq.10) for the engineering MAs were 0.05 regardless of noise),
304 the rough shape appears to be reproduced (the errors were around 0.09). The result
305 suggests that the physiologically plausible 2D MA produces a stable structure that is capable
306 of representing an object's shape with robustness.

307

308 3.2 The generation of 3D MAs for elementary shapes

309 The computed 2D and 3D MAs for a capsule, the simplest shape for representation by a MA,
310 are shown in Figure 3. Input images for the left and right eyes are shown in Figure 3A. The
311 computed 2D MAs for each eye is shown in Figure 3B. We observe a rod-like MA elongated
312 along the major axis of the capsule. The 2D MA for the left eye appears slightly tilted
313 compared with that for the right eye, indicating that the top side (in 2D image) of the
314 capsule is farther than the bottom side. We set the fixation point (depth = 0) at the bottom
315 end of the major axis, such that the disparity increases toward the top side. The computed
316 3D MA is shown in Figure 3C. We observed a rod-like MA with its depth increasing toward
317 the top side, showing agreement with the shape of the capsule.

318 We computed 2D and 3D MAs for a cuboid, which is another elementary shape with
319 sharp corners (Figure 3D). The computed 2D MA for the cuboid is shown in Figure 3E.
320 Similarly to the capsule, the tilt of the 2D MAs (Figure 3E) is slightly different between the
321 left and right images (the left MA is more tilted). We set the fixation point (depth = 0) at the
322 nearest corner of the cuboid, such that the disparity increases toward the top side. The
323 computed 3D MA is shown in Figure 3F. We observe a vase-like MA with its depth increasing
324 toward the top side, showing agreement with the shape of the cuboid. These results show
325 that the model computed reasonable 3D MAs for elementary shapes with simple structure.

326

327 3.3 Evaluation by the reconstruction of 3D shape for elementary shapes

328 To evaluate the adequacy of the computed 3D MA, we reconstructed a 3D shape based on
329 the 3D MA, and computed how accurately the computed 3D MA is capable of reproducing a
330 3D shape in terms of its depth and shape. For the reconstruction, we needed the distances
331 between the MA and the surrounding surface, as well as the location of the MA. The model
332 focuses on the location of the MA, and it does not determine the distances to the surface.
333 For the 3D reconstruction, we used the Euclidean distance between the MA and the nearest
334 contour that is stored separately from the model, as described in the Model section. We
335 reconstructed the 3D shape by placing a number of overlapping spheres along the 3D MA.
336 The centers of the spheres were aligned with the MA, and the radii were set equal to the
337 distance to the nearby contour.

338 We evaluated quantitatively the accuracy of the reconstruction in terms of depth
339 and shape. The reconstruction error for *depth* was defined as the difference in the depth
340 maps between the original and the reconstruction:

$$341 \quad Error_{depth} = \sqrt{\frac{\sum_{x,y} \{D_{original}(x,y) - D_{reconstruct}(x,y)\}^2}{\sum_{x,y} D_{original}(x,y)^2}}, \quad \text{Eq. 15}$$

342 for $(x, y \mid D_{original}(x, y) \cap D_{reconstruct}(x, y) \neq \phi)$,

343 where $D_{original}(x, y)$ and $D_{reconstruct}(x, y)$ represent the depth map of the original and
344 reconstruction, respectively. The depth map indicates the distances of all points on the
345 object surface from the eye. This index computes the difference in depth for all points where
346 the original and the reconstruction overlap. To evaluate the reconstruction of *shape*, we
347 introduced an index, $Error_{shape}$, which was defined by the normalization of $Error_{depth}$ to
348 the maximum depth within each map. This normalization cancels out the absolute depth so
349 that shape (or relative depth) is evaluated. Note that $Error_{shape}$ estimates the shape of
350 the front side, not the overall 3D shape, because the model does not estimate the back side
351 of an object. These error indices become zero when the reconstruction is perfect (equal to
352 the original), and one when the reconstruction is twice as large as the original.

353 Figure 4 shows the reconstruction of the two elementary shapes, the capsule and
354 cuboid. The columnar shape and the rounded ends of the capsule were reconstructed

355 smoothly (Figure 4A). Figure 4B shows the difference in depth from the viewing point. The
356 overall difference, $Error_{depth}$, was 0.78. As we discuss later, the error appears to be caused
357 by the simplification of the energy model in which only one and three channels are provided
358 for spatial frequency and orientation, respectively. Figure 4C shows the difference in shape
359 (relative depth). The overall difference, $Error_{shape}$, was 0.16, indicating that the model
360 successfully reproduced the shape with rounded surfaces. Further evaluation of the errors is
361 discussed in the next section. Because the shape of the capsule is composed of a set of
362 spheres, it was expected that the reconstruction from overlapping spheres along the 3D MA
363 would reproduce the shape of the capsule with high accuracy. A cuboid with sharp corners
364 was expected to be difficult for the model. Figure 4D shows the reconstruction of a cuboid.
365 Although the reconstructed shape is somewhat rounded compared with the original cuboid,
366 we can still observe corners that are a crucial feature of a cuboid. $Error_{depth}$ for the
367 cuboid was 0.79, indicating a level of accuracy similar to the capsule. The reconstruction of
368 the surface was fairly successful with $Error_{shape}$ of 0.52. These results indicate that the 3D
369 MA computed by the model is fairly capable of representing the shape of objects with
370 elementary shapes.

371

372 3.4 Evaluation of 3D MA for shapes with typical features

373 To investigate the accuracy of the model for the representation of 3D shape, we
374 performed simulations of the model with a variety of objects with distinct features. In this
375 section, we report in detail the results of three examples with typical features: (i) a shape
376 with varying thickness along its major axis, (ii) a shape with a curved axis, and (iii) a
377 combination of multiple features. Overall evaluation of the model for various shapes is
378 discussed in the next section.

379 Figure 5A shows the results of a vase whose radius varies along the major axis. The
380 reconstruction from the computed 3D MA shows the depth increasing along the major axis
381 from the center to both ends, indicating successful reproduction of the crucial features of
382 the vase. $Error_{depth}$ and $Error_{shape}$ were 0.57 and 0.62, respectively. The vase was
383 expected to be easy for the model, similarly to the capsule, because both surfaces are
384 rounded. However, the reconstruction errors were still larger than the cuboid that consists
385 of flat surfaces and sharp corners. This large error is attributable to the failure of the
386 reproduction along the top and bottom of the vase where surfaces splay out. This change is

387 barely detected by the disparity in the contours of the vase (the boundary between the vase
388 and background) that extends horizontally at the top and bottom ends. Because the overall
389 surface (except for both the ends) was reproduced successfully, the results indicate that the
390 3D MA computed from the model is capable of representing a shape with varying thickness
391 along the axis.

392 Figure 5B shows the results for a golf club that contains a bent axis and a flat surface.
393 In both the reconstruction and in the computed 3D MA, we can observe a sharp bend at the
394 middle of the head of the club. The depth of the reconstruction increases from the bottom
395 end toward the top end, which is consistent with the structure of the original shape.
396 $Error_{depth}$ and $Error_{shape}$ were 0.85 and 0.64, respectively, similar to the range for other
397 stimuli. The major cause of the error was the flatness of the club head. As discussed with the
398 cuboid, we reconstruct objects using spheres along the axis, so that the reconstruction of a
399 flat surface is difficult. These results indicate that the computed 3D MA is capable of
400 representing an object shape that contains a sharp bend along the major axis.

401 Figure 5C shows the results for a cow that has a complex structure. The head and
402 body of the cow appear to be reproduced smoothly and successfully. The values of
403 $Error_{depth}$ and $Error_{shape}$ were 0.62 and 0.68, respectively. The failure of the
404 reconstruction of the legs was a major source of the error. Because the present model has
405 only a single frequency, small parts are disregarded. The results show that the model is
406 capable of representing a complex structure with an error similar to that of simple structures,
407 which is consistent with the characteristics of the human visual system. This result supports
408 the robustness of the model in its representation of shape.

409 To examine the representation of shape from real images (not created by CG), we
410 conducted a simulation of the model using stereo photographs that may include a variety of
411 noise. A pair of convergent stereo images of a miniature duck was taken, as shown in Figure
412 6A, and used as an input stimulus. The fixation point was set at the center of the duck's
413 chest. Figure 6B-D shows the results of the simulation. The depth of the computed 3D MA,
414 as shown in Figure 6C, increases as it diverges from the center of the chest. Figure 6D shows
415 the reconstruction in which the shape of the head and body of the duck appear to be
416 reasonably reproduced. These results suggest that the model is capable of generating a
417 reasonable 3D MA from real images.

418

419 3.5 Overall evaluation of the reconstruction error

420 We evaluated quantitatively the accuracy of the reconstruction in terms of depth and 3D
421 shape. The reconstruction error for depth, $Error_{depth}$, represents how accurately the
422 absolute depth is reproduced by taking the difference between the depth maps of the
423 reconstruction and the original, and the error for shape, $Error_{shape}$, represents how
424 accurately 3D shape is reproduced by canceling out the absolute depth. We reconstructed
425 the 3D shape from the 3D MA for a variety of objects, in addition to those with typical
426 shapes as shown above. The eight input stimuli, the computed 3D MAs, and the
427 reconstructions of shape are shown in Figure 7. The $Error_{depth}$ and $Error_{shape}$ for all
428 objects (including those shown in the previous sections) are shown in Table 1. The mean and
429 SD of the depth error were 0.69 and 0.13, respectively, indicating that the capability of the
430 model to represent 3D depth is relatively independent of the complexity of the shape and
431 structure of the object. The mean and SD of the shape error were 0.70 and 0.34, respectively.
432 It appears that low errors were observed for the objects whose surface is smoothly rounded
433 or relatively simple when viewed from the designated eye position. The duck showed the
434 worst error among these objects, because the width of its neck differed between the left
435 and right images so that the shape of their 2D MAs were very distinct; this discrepancy
436 caused the failure of binocular fusion leading to an inaccurate representation of depth in the
437 3D MA. These results indicate that the proposed model is capable of representing the rough
438 shape of various 3D objects. Given the limited number of frequency and orientation
439 channels (1 and 3 for frequency and orientation, respectively), the reproduction should be
440 considered surprisingly successful.

441

442 3.6 Evaluation by the reconstruction of 2D stimulus (frontal projection)

443 As an evaluation of the internal representation of the model, we examined the capability of
444 the model to reconstruct the original input stimulus from the computed 3D MA. We defined
445 the error in 2D projection, $Error_{2D}$, as an index to indicate how accurately the model is
446 capable of reproducing the 2D shape:

$$447 \quad Error_{2D} = \frac{|S_{original} - S_{reconstruct}|}{S_{original}}, \quad \text{Eq. 16}$$

448 where $S_{original}$ and $S_{reconstruct}$ indicate the surface areas that are projected onto an eye

449 (the 2D area seen from a viewing point) of the original and reconstruction, respectively. This
450 index is important in that it shows the capability of the model to reconstruct the original
451 stimulus from the internal representation of the model. The index becomes zero when the
452 reconstruction is perfect, and one when the reconstruction is twice as large as the original.
453 The errors of all objects are shown in Figure 8. The mean and *SD* of the error were 0.14 and
454 0.05, respectively. The error was less than 20% for most of the objects except the horse and
455 the elephant whose legs were too thin to be reproduced. We also calculated separately the
456 errors for the over- and underestimation of the areas (the positive and negative parts of the
457 index). The results are shown in white and black in the insets of Figure 8, and the values are
458 given in Table 2. The model appears to show overestimation where the contour of an object
459 is concave, and underestimation where the part is small. Because the shape is reconstructed
460 by superimposing spheres, concave regions tend to be masked by the spheres
461 (overestimated). Small parts are often missed because the present model consists of a single
462 spatial frequency channel. If multiple frequency channels were provided, the model would
463 be capable of detecting these small parts and avoid underestimation. Multiple frequency
464 channels would also be helpful in reducing the overestimation caused by concave surfaces.
465 These results support the capability of the model to represent object shape.

466

467 3.7 View invariance of the reconstruction

468 IT neurons that are selective for 3D MA showed view-invariant responses [9]. The human
469 visual system also shows view invariance in its perception of object shape, although the
470 reaction time often varies. It is expected that view invariance is an inherent characteristic of
471 the representation by the MA. Here, we evaluated whether our model reproduces view
472 invariance in the reconstruction error. We computed the 3D MA and the reconstruction
473 error for a series of images viewed from distinct directions. Specifically, we used the stimuli
474 of a cow viewed from its side, tail, and an in-between position. The input stimuli are shown
475 in Figure 9 (generated by rotating the cow shown in Figure 5C), together with the computed
476 3D MAs and the reconstructions. The head and body of the cow were reproduced in all views,
477 although mostly its thin legs were not. The error in depth, $Error_{depth}$, for each view was
478 0.54 (Figure 9A), 0.40 (Figure 9B), and 0.62 (Figure 9C), respectively, and the mean of the
479 three was 0.52. The error in shape, $Error_{shape}$, for each view was 0.56 (Figure 9A), 0.75
480 (Figure 9B), and 0.83 (Figure 9C), respectively, and the mean of the three was 0.71. Both

481 errors in depth and surface show small variation: all views show errors that are within 10%
482 of the means. This result shows that the model is capable of reproducing shapes from a
483 variety of viewpoints with similar amounts of error. This view invariance is consistent with
484 the characteristics of IT neurons tuned to a 3D MA configuration, and also human vision.
485

486 4 Discussion

487
488 Numerous studies have suggested object-centered coordinates for the cortical
489 representation of shape [1,2,3]. Although theoretical studies have favored the advantages of
490 the MA representation for more than three decades, only a few physiological studies have
491 reported supportive results[13]. Recently, an electrophysiological study has provided direct
492 evidence that neurons in IT show selectivity for the 3D MA configuration [9]. However, the
493 mechanisms by which the 3D MA is constructed through the visual pathway have not been
494 clarified. The present study examined neural processes for the generation of a 3D MA. A
495 physiological study has reported that neurons in V1 respond to the medial region of a
496 textured figure[13]. Such a response in V1 could be produced by the synchronization of
497 BO-selective neurons in V2, and the 2D MA has been reported to be capable of coding object
498 shape [16]. We focused on the transformation of the 2D MA reported in V1 into the 3D MA
499 observed in IT. The latency of V1 cells that respond to the edges of an object range between
500 40 and 60 ms [13, 24], and that to 2D MA range between 90 and 110 ms [13, 24]. The onset
501 latency of IT cells is generally more than 90 ms [25] and the latency for 3D MA is considered
502 to be much longer than 90 ms. Given this time constraint, afferent connections appear to
503 play a crucial role in the transformation from 2D MAs to a 3D MA, probably in combination
504 with efferent connections. Therefore, the present study investigated the generation of a 3D
505 MA by the binocular fusion of 2D MAs. Note that the present model does not account for
506 these latencies. It is expected to further study the temporal properties of the representation
507 of 3D shape.

508 In the process of binocular fusion, it is crucial to determine whether the responses
509 of V1 cells to MA are monocular or binocular. If the MA is monocular, a retinal image of an
510 object is transformed into a monocular 2D MA by V1 cells, and then the fusion of a binocular

511 pair of 2D MAs generates a 3D MA based on the disparity between the 2D MAs. On the other
512 hand, if the MA in V1 is binocular, a local segment of the 3D MA is produced from a
513 binocular pair of local contours of the object image, and the integration of the 3D MA
514 segments generates a global 3D MA in IT. Consider the case where a contour of one side of
515 an object is nearer than the fixation point, and that of the other side is farther. In the
516 monocular case, the disparity-selective cells fuse a pair of 2D MAs based on the disparity
517 between the axes to generate a 3D MA that represents correct depth. On the other hand, in
518 the binocular case, the fusion of a binocular pair of local contours would be extremely
519 difficult because the fusion requires V1 cell that is tuned to near on one side and far on the
520 other side, and that signals depth at the middle of the two. An alternative would consider
521 feedback from higher cortical areas to V1. Because disparity-selective cells in V1 detect local
522 disparity and the higher cortical regions are required to produce global depth, a higher
523 region such as V4 and IT would generate 3D contours and send feedback to generate local,
524 binocular 2D MAs in V1. Although feedback may play an important role, an assumption of
525 such complex pathways prevents the construction of a plausible computational model. In
526 the present study, we focused on the monocular case, and proposed the hypothesis that a
527 pair of 2D MAs that encode monocular projections of object shape is fused to generate a 3D
528 MA, as a first step toward understanding the transformation of MA from V1 to IT.

529 We defined the physiologically plausible 2D MA to mimic the activities of V1 cells
530 responding to 2D MA. The physiologically plausible 2D MA is capable of representing the
531 outline of an object with around 10% error. A major downside of MA representation in
532 general could be high sensitivity to noise on contours. In the real world, a retinal image of an
533 object often includes noise on contours for a variety of reasons. In fact, an engineering MA
534 that is defined by a set of axes (skeleton) often changes considerably in response to noise, so
535 that even the graph structure that represents the object shape varies. Given that the visual
536 system is able to perceive shape with stability and robustness in the presence of noise, the
537 engineering MA may not be a suitable candidate for cortical representation. In the present
538 study, we propose that the physiologically plausible MA overcomes this disadvantage. To
539 reproduce V1 responses to 2D MA, we defined the physiologically plausible MA as having an
540 equal distance between the point under examination and nearby contours. Specifically, we
541 computed the equality of the distances from the contours, and determined the region of the
542 MA by setting a threshold for equality. Because of this processing, the MA is defined by a set

543 of points whose distances from the contours are similar, but not exactly equal, giving it
544 robustness with respect to noise on contours. Therefore, the physiologically plausible MA is
545 inherently more robust than the engineering MA, at the expense of accuracy. The
546 physiologically plausible MA appears meaningful in terms of its stability and robustness.

547 The present model uses a standard energy model [23] to determine the binocular
548 disparities of physiologically plausible 2D MAs. The disparity at each location along the axis is
549 detected by using a winner-take-all mechanism among disparity-selective cells, each of
550 which is tuned to a distinct disparity. For the sake of simplicity, the model has only a single
551 spatial frequency channel and three orientation channels. Therefore, the accuracy for
552 disparity detection is very limited, and much lower than that of the visual system. It should
553 be noted that a model with this simple structure is capable of generating a 3D MA whose
554 disparity varies reasonably according to the original shape, and reproduces the overall form
555 of the original object. These results support the plausibility of binocular fusion of
556 physiologically plausible MAs using the energy model.

557 We constructed the model for the generation of a 3D MA based on the binocular
558 fusion of physiologically plausible 2D MAs, and examined whether this model is suitable for
559 the representation of 3D shape. We computed a 3D MA from a number of objects, with a
560 variety of shape characteristics, including natural objects. The model was capable of
561 generating a reasonable 3D MA for a wide range of objects. We also reconstructed the 3D
562 shape of the test objects based on the computed 3D MA. The model showed excellent
563 reconstruction accuracy for somewhat rounded objects such as a capsule, and reasonable
564 accuracy for all other objects including those with sharp corners, flat surfaces, and complex
565 structures. Given the limited number of frequency and orientation channels, the
566 reproduction should be considered as surprisingly successful. Furthermore, the simulation
567 results showed view invariance in the reconstruction, which is consistent with the results of
568 physiological experiments [9]. These results show that a model based on the fusion of a
569 binocular pair of physiologically plausible 2D MAs generates a reasonable 3D MA with
570 robustness in representing the 3D structure independent of viewpoint, indicating the
571 plausibility of the model as a candidate for the cortical computation of 3D MA.

572

573 **Acknowledgments**

574 This work was supported by Grants-in-Aid from the JSPS (22300090, 26280047) and that
575 from Scientific Research on Innovative Areas “Shitsukan” (23135503, 25135704) of MEXT
576 Japan.
577

578 References

- 579 [1] I. Biederman, P.C. Gerhardstein, "Recognizing Depth-Rotated Objects: Evidence and Conditions
580 for Three-Dimensional Viewpoint Invariance", *Journal of Experimental Psychology: Human*
581 *Perception and Performance*, vol.19, pp.1162-1182 (1993)
- 582 [2] K. Tsunoda, Y. Yamane, M. Nishizaki, M. Tanifuji, "Complex objects are represented in macaque
583 inferotemporal cortex by the combination of feature columns", *Nature Neuroscience*, vol.4,
584 pp.832-838 (2001)
- 585 [3] Y. Yamane, K. Tsunoda, M. Matsumoto, A.N. Phillips, M. Tanifuji, "Representation of the Spatial
586 Relationship Among Object Parts by Neurons in Macaque Inferotemporal Cortex", *Journal of*
587 *Neurophysiology*, vol.96, pp.3147-3156 (2006)
- 588 [4] D. Marr, H.K. Nishihara, "Representation and Recognition of the Spatial Organization of
589 Three-Dimensional Shapes", *Proceedings of the Royal Society of London, Series B, Biological*
590 *Sciences*, vol.200, pp.269-294 (1978)
- 591 [5] H. Blum, "Biological shape and visual science. I.", *Journal of Theoretical Biology*, No.38,
592 pp.205-287 (1973)
- 593 [6] I. Biederman, "Recognition-by-Components: A Theory of Human Image Understanding",
594 *Psychological Review*, vol.94. pp.115-147 (1987)
- 595 [7] I. Kovacs, B. Julesz, "Perceptual sensitivity maps within globally defined visual shapes", *Nature*,
596 vol.370, pp.644-646 (1994)
- 597 [8] J. Feldman, M. Singh, "Bayesian estimation of the shape skeleton", *Proceedings of the National*
598 *Academy of Sciences of the United States of America*, vol.103, pp.18014-18019 (2006).
- 599 [9] C.C. Hung, E.T. Carlson, C.E. Connor, "Medial Axis Shape Coding in Macaque Inferotemporal
600 Cortex", *Neuron*, vol.74, pp.1099-1113 (2012)
- 601 [10] P. Janssen, R. Vogels, Y. Liu, G.A. Orban, "Macaque Inferior Temporal Neurons Are Selective for
602 Three-Dimensional Boundaries and Surfaces", *Journal of Neuroscience*, vol.21, pp.9419-9429
603 (2001)
- 604 [11] P. Janssen, R. Vogels, G. A. Orban, "Three-Dimensional Shape Coding in Inferior Temporal
605 Cortex", *Neuron*, 27, pp.385-397 (2000)
- 606 [12] M. D. Lescroart and I. Biederman, "Cortical representation of medial axis structure", *Cerebral*
607 *Cortex*, vol.23, pp.629-637 (2013)
- 608 [13] T.S. Lee, D. Mumford, R. Romero, V.A.F. Lamme, "The role of the primary visual cortex in higher

609 level vision”, *Vision Research*, vol.38, pp.2429-2454 (1998)

610 [14] B.B. Kimia, “On the Role of Medial Geometry in Human Vision”, *Journal of Physiology-Paris*,

611 vol.97, pp.155-190 (2003)

612 [15] K. Sakai, H. Nishimura, “Surrounding Suppression and Facilitation in the Determination of

613 Border Ownership”, *Journal of Cognitive Neuroscience*, vol.18, pp.562-579 (2006)

614 [16] Y. Hatori, K. Sakai, “Early representation of shape by onset synchronization of

615 border-ownership-selective cells in the V1-V2 network” *Journal of Optical Society of America, A*,

616 vol.31, pp.716-729 (2014)

617 [17] C. A. Burbeck, S. M. Pizer, “Object representation by Cores: identifying and representing

618 primitive spatial regions” *Vision Research*, 35(13), pp.1917-1930 (1995)

619 [18] H. M. Shiozaki, S. Tanabe, T. Doi I. Fujita, “Neural Activity in Cortical Area V4 Underlies Fine

620 Disparity Discrimination” *Journal of Neuroscience*, vol.32, pp.3830-3841 (2012)

621 [19] D. A. Hinkel, C. E. Connor, “Three-dimensional orientation tuning in macaque area V4”, *Nature*

622 *Neuroscience*, vol.5, pp.665-670 (2002)

623 [20] E. H. Adelson and J. R. Bergen, “Spatiotemporal energy models for the perception of motion”,

624 *Journal of Optical Society of America, A*, vol.2, pp.284-299 (1985)

625 [21] Y. Hatori, K. Sakai, “Robust Detection of Medial-Axis by Onset Synchronization of

626 Border-Ownership Selective Cells and Shape Reconstruction from its Medial-Axis”, *Advances in*

627 *Neuro-Information Processing, Lecture Notes in Computer Science*, vol.5506, pp.301-309 (2009)

628 [22] H. Zhou, H. S. Friedman, R. von der Heydt, “Coding of border ownership in monkey visual cortex”,

629 *Journal of Neuroscience*, vol.20, pp.6594-6611 (2000)

630 [23] I. Ohzawa, G.C. DeAngelis, R.D. Freeman, “Stereoscopic depth discrimination in the visual cortex:

631 neurons ideally suited as disparity detectors”, *Science*, vol. 249, pp.1037-1041 (1990)

632 [24] J. Poort, F. Raudies, A. Wannig, V.A.F Lamme, H. Neumann, P.R. Roelfsema, “The Role of

633 Attention in Figure-Ground Segregation in Areas V1 and V4 of Visual Cortex”, *Neuron*, vol.75,

634 pp.143-156 (2012)

635 [25] J. Bullier, “Communication between Cortical Areas of Visual System”, L.M. Chalupa, J.S. Werner

636 Eds., *The Visual Neurosciences*, vol.1, pp. 522-540, THE MIT PRESS.

637 <http://graphics.im.ntu.edu.tw/~robin/courses/cg03/model/>

638 <http://web.engr.oregonstate.edu/~mjb/cs519/Obj>

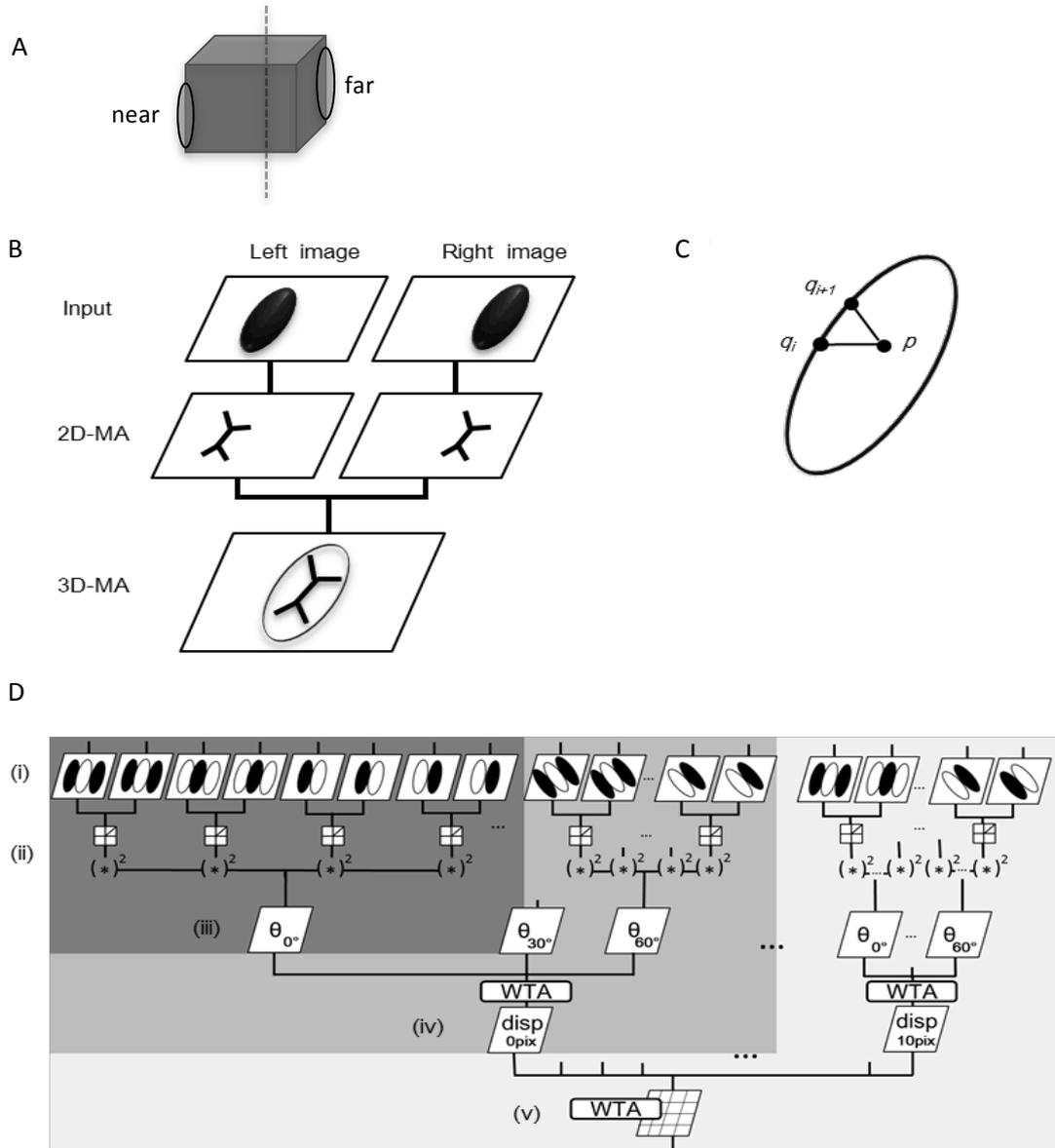
639 <http://groups.csail.mit.edu/graphics/classes/6.837/F03/models>

640 <http://users.cms.caltech.edu/~njlitke/meshes/toc.html>

641

642 Figure Captions

643



644
 645
 646
 647
 648
 649
 650
 651

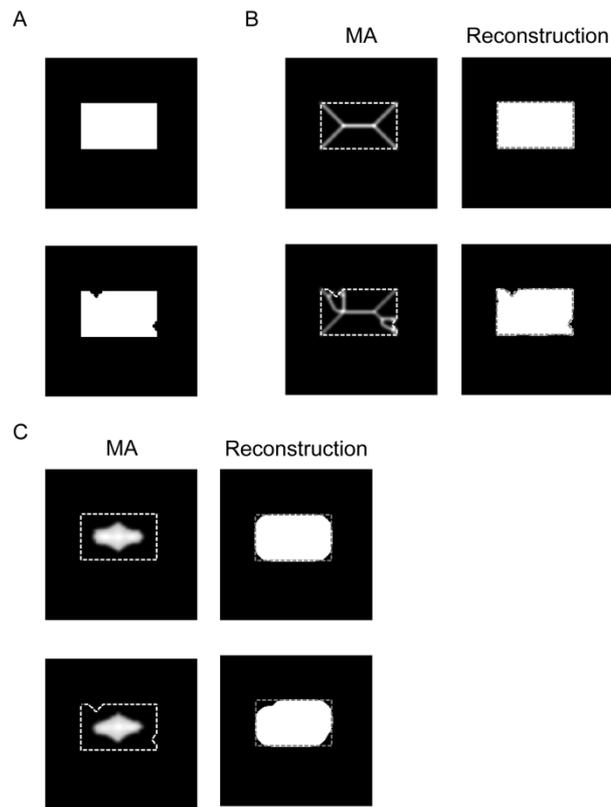
Figure 1.

652
653
654
655
656
657
658
659
660
661
662
663
664
665
666
667
668

Figure 1.

(A) An illustration showing that, in general, the depths of the sides and the MA of an object could be different. In this example, the right and left sides of the cube are far and near, respectively, with respect to the vertical dotted line. (B) The model comprises two distinct stages: detection of monocular 2D MAs based on the distances from surrounding contours, and generation of a 3D MA from the disparities between the two 2D MAs. (C) A 2D MA is defined as a set of points (e.g., p) equidistant from nearby contours (q_i, q_{i+1}). (D) A detailed illustration of the model. Activities of a pair of simple cells with a certain phase difference (e.g., an in-phase pair for disparity=0) are summed (i), and pass through a half-squaring computation (ii). A model complex cell pools four quadrature pairs of simple cells whose preferred orientation is one of three orientations (0, 30, or 60°; iii). The responses of three complex cells with a distinct preferred orientation are integrated by winner-take-all (iv). There are eleven channels with distinct phase differences, corresponding to eleven distinct disparities. The optimal disparity at each spatial position is chosen from the eleven distinct disparities by winner-take-all (v).

669
670
671

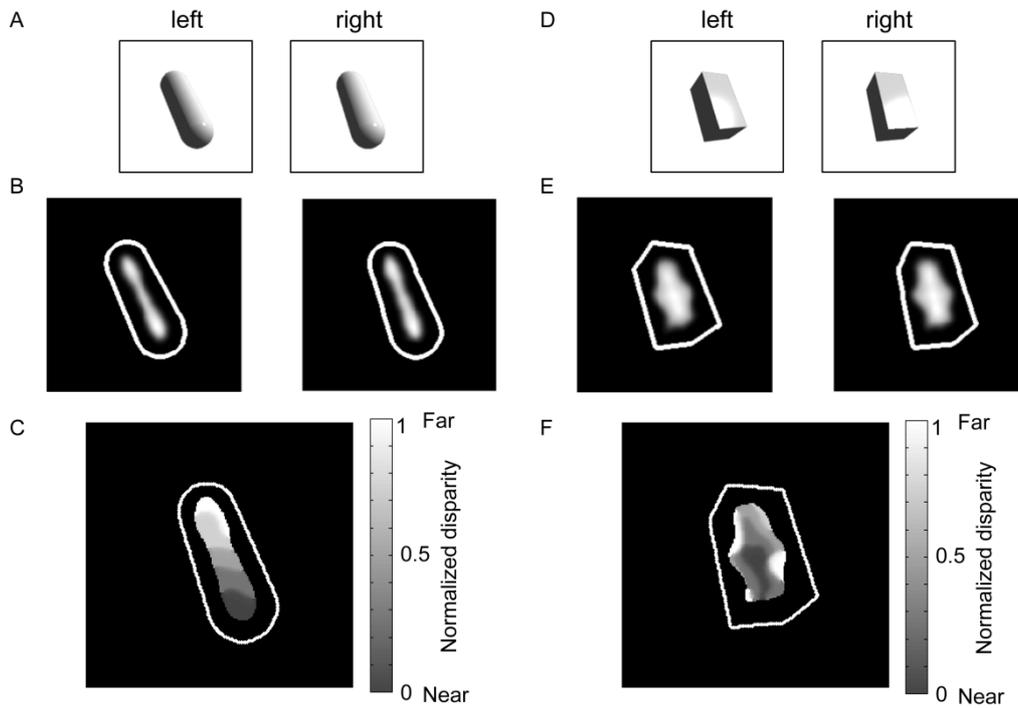


672
673

674 Figure 2.

675 (A) Stimuli used for the computation of 2D MAs, with and without noise on the contour
676 (bottom and top, respectively). (B) The engineering MAs (left) and their reconstructions
677 (right). Dotted lines indicate the object contour (shown for presentation purposes). The two
678 engineering MAs were different, with a correlation coefficient of 0.77. The reconstructed
679 images were accurate with reconstruction errors of 0.05 for both stimuli. (C) The biological
680 MAs (left) and their reconstructions (right), with (bottom) and without (top) contour noise.
681 The two biological MAs were similar with a high correlation coefficient of 0.99. Although the
682 reconstructions were less accurate (errors of 0.09) than those of engineering MAs,
683 reasonable shapes were achieved.

684
685



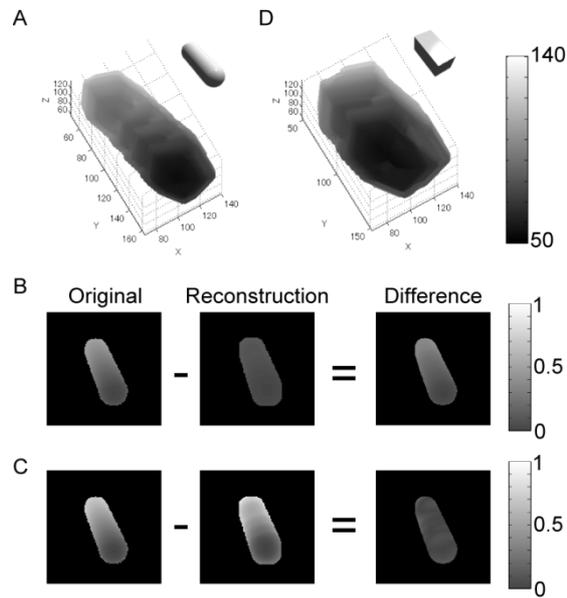
686
687

688 Figure 3.

689 Computation of 3D-MAs from binocular 2D-MAs (A-C for a capsule; D-F for a cuboid). (A, D)
690 Two input images for the left and right eyes. (B, E) The computed 2D MAs. White lines
691 indicate object contours for presentation purposes, and are not computed by the model.
692 (C, F) The 3D MA fused from the binocular 2D-MAs. The disparity computed by the model is
693 plotted in grey. White/dark gray represents a far/near disparity. (A-C) The fixation point
694 (depth = 0) was set at the bottom end of the major axis, so that the disparity increases
695 toward the top. The computed 3D-MA shows a smooth gradient for the disparity consistent
696 with the ground truth. (D-F) The fixation point was set at the nearest corner of the cuboid,
697 so that the disparity increases toward the top. The disparity in the 3D-MA is somewhat
698 complicated because of the sharp corners.

699

700
701

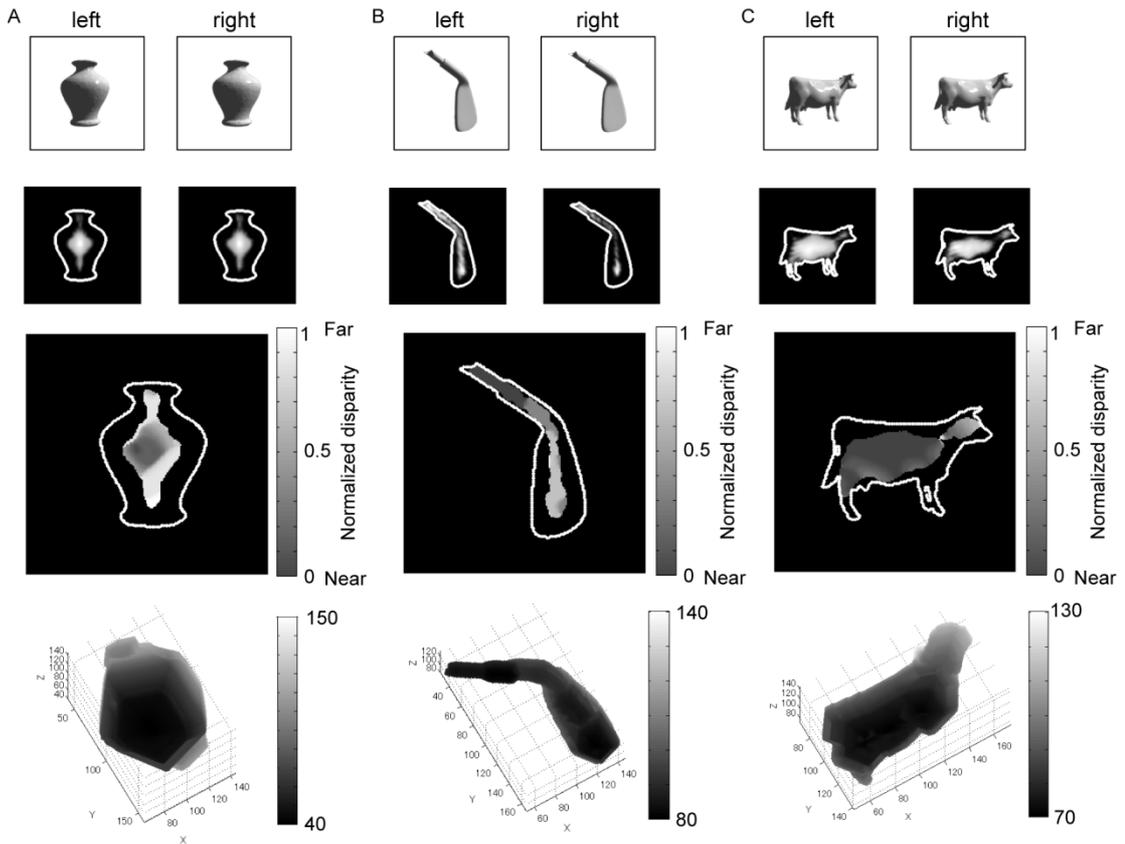


702
703
704
705
706
707
708
709
710
711
712

Figure 4.

(A) Reconstructed shape of a capsule. The reconstruction was given by the superposition of overlapping spheres onto the computed 3D MA. For details, see *the model* section. The x-y axes represent the plane projected onto a camera. The z-axis and grey represent depth (a larger value indicates farther away). (B) Evaluation of the difference in *depth* between the original and the reconstruction of the capsule. The right panel shows the difference in grey (between 0 and 1). The overall error for depth was 0.78. (C) Evaluation of the difference in *shape* (relative depth) between the original and the reconstruction of the capsule. The error for shape was 0.16. (D) Reconstructed shape of a cuboid.

713



714

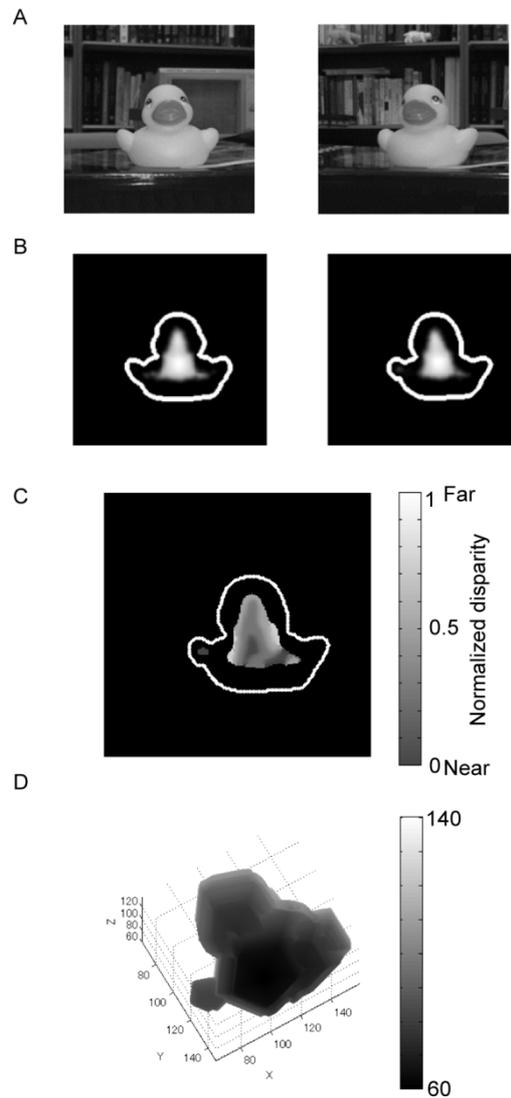
715

716 Figure 5.

717 Simulation results for the stimuli with typical features. The top row shows the binocular
718 stimuli for a vase (A), a golf club (B), and a cow (C). The second row shows the computed
719 2D-MAs. The third row shows the computed 3D MAs. Conventions are the same as in Figure
720 3. The bottom row shows the reconstructed shapes from the 3D MAs. Conventions are the
721 same as in Figure 4. The errors for depth were 0.57 (A), 0.85 (B), and 0.62 (C), and the errors
722 for shape were 0.62 (A), 0.64 (B), and 0.68 (C).

723

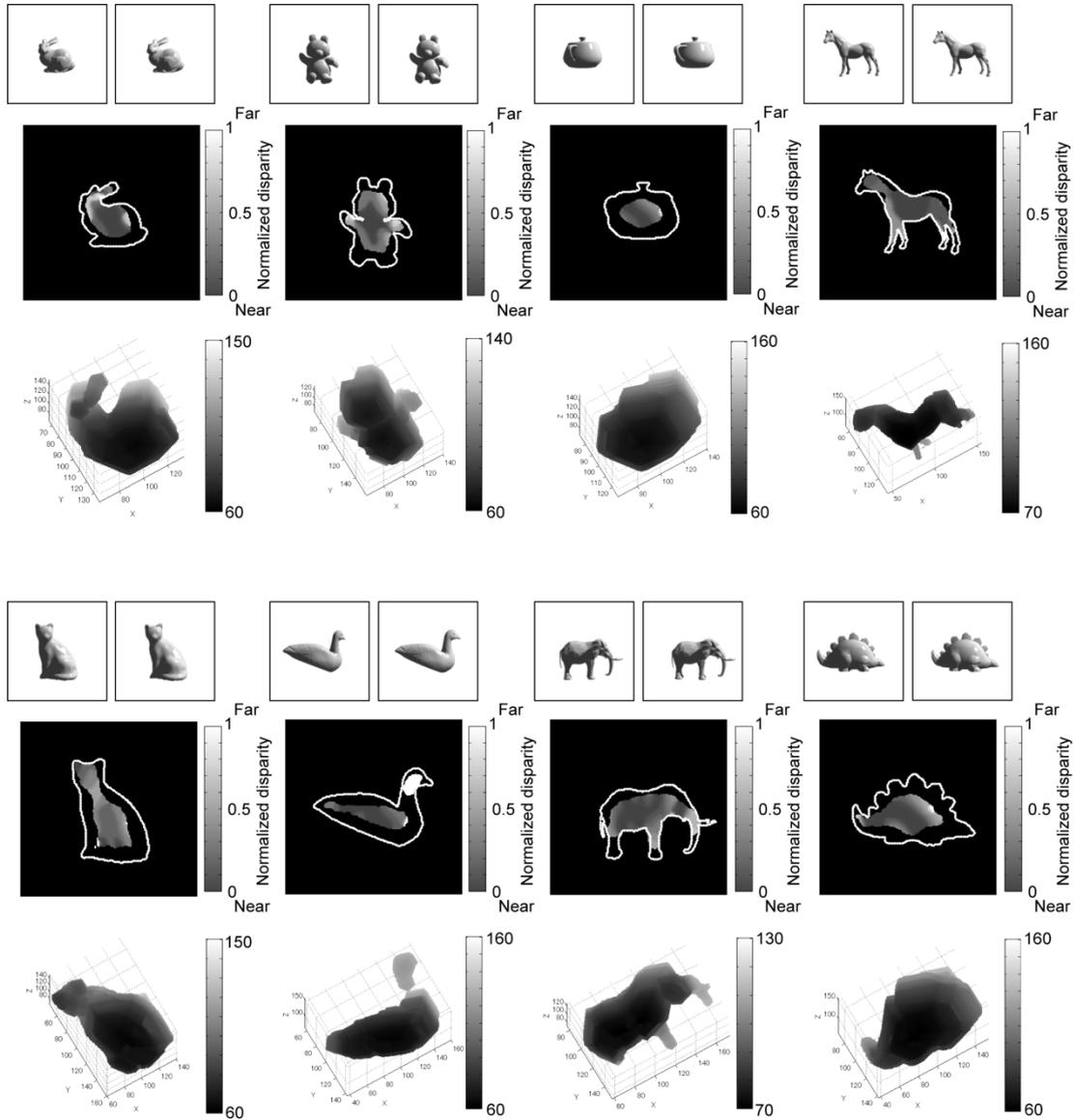
724
725



726
727
728
729
730
731
732
733
734
735

Figure 6.

Simulation results for real stereo images of a duck. Conventions are the same as in Figure 5. (A) The images of the duck taken using a stereo camera with the fixation point set at the center of the front of the body. (B) The computed 2D MAs. (C) The 3D MA obtained from the binocular 2D-MAs. The computed depth increases as it departs from the center of the front of the body. (D) The reconstructed shape from the 3D MA. The head and body of the duck are visible.



737

738

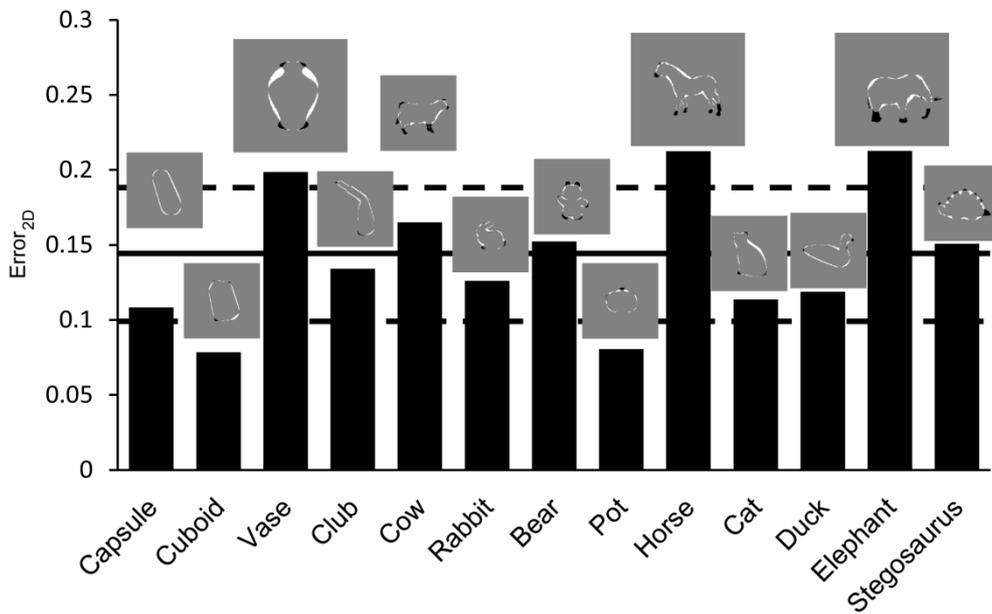
739 Figure 7.

740 Simulation results for other stimuli such as a rabbit, a bear, a pot, a horse, a cat, a duck, an
 741 elephant, and a stegosaurus. Conventions are the same as in Figure 5. The errors for
 742 reconstruction are summarized in Table 1. All shapes were reasonably reconstructed,
 743 including those with complex shapes.

744

745

746

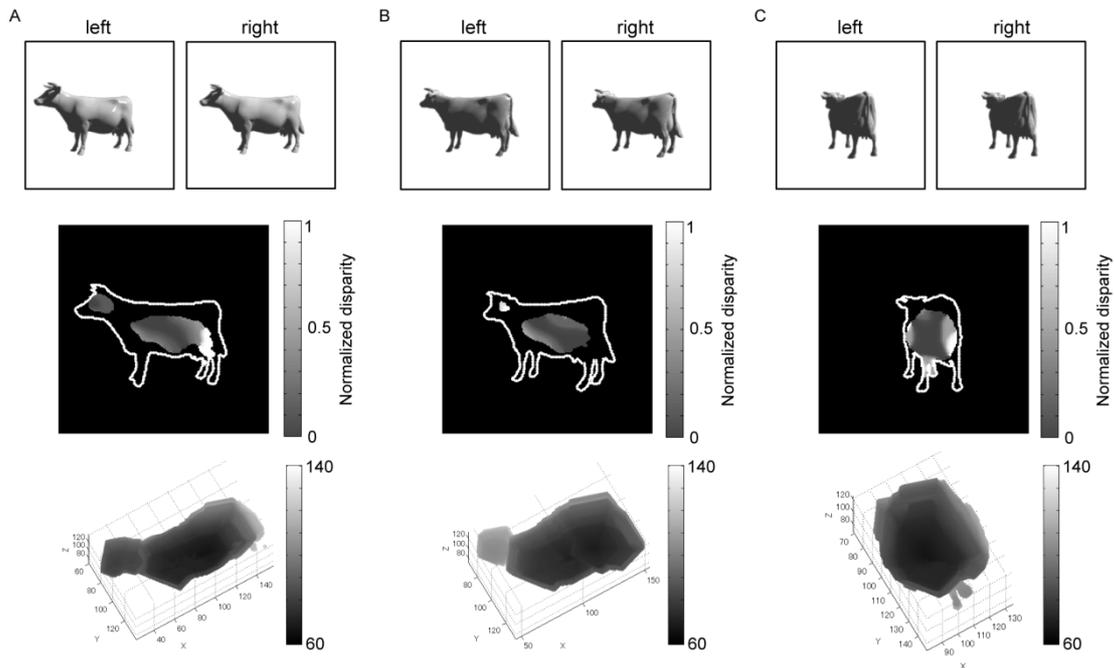


747

748 Figure 8.

749 Evaluation by the reconstruction of 2D images (surface area). The differences in the surface
750 areas between the original and reconstructed shapes ($Error_{2D}$) are plotted. The surface area
751 was determined by projecting a 3D shape onto a camera. The solid line and dotted lines
752 indicate the mean and SD of the errors, respectively. The errors were less than 20% except
753 for the horse and elephant. The over- and under-estimation of the areas are shown in the
754 insets by white and black, respectively. Overestimation is often observed around concave
755 contours, whereas underestimation occurs around small parts.

756



758

759

760 Figure 9.

761 View invariance of the reconstruction. Conventions are the same as in Figure 5. (A-C) The
 762 simulation results for a cow viewed from three different directions. The top row shows the
 763 stereo stimuli that were viewed from distinct points. The middle row shows the 3D MAs
 764 computed from the binocular 2D MAs. The bottom row shows the reconstructed shapes
 765 computed from their 3D MAs. The depth errors were 0.54 (A), 0.40 (B) and 0.62 (C). The
 766 shape errors were 0.56 (A), 0.75 (B) and 0.83 (C). Both types of error in the reconstruction
 767 show small variation, indicating that the model is capable of reproducing reasonable shapes
 768 regardless of viewpoint.

769

770 Table 1.

771 Reconstruction errors in depth and shape for all stimuli

772

Stimulus	Depth error	Shape error
Capsule	0.7771	0.1559
Cuboid	0.7879	0.5244
Vase	0.5822	0.6298
Club	0.8469	0.6408
Cow	0.6176	0.6803
Rabbit	0.5861	0.6577
Bear	0.7205	1.0237
Pot	0.7412	0.4924
Horse	0.7412	0.6748
Cat	0.6860	0.7660
Duck	0.6173	1.5153
Elephant	0.9318	1.0348
Stegosaurus	0.4262	0.4011
Mean	0.6891	0.7043
<i>SD</i>	0.1325	0.3368

773

774

775

776 Table 2.

777 Reconstruction errors in 2D projection for all stimuli

778

Stimulus	Error_{2D}	Overestimated	Underestimated
Capsule	0.1084	0.1084	0
Cuboid	0.0785	0.0703	0.0083
Vase	0.1986	0.1635	0.0351
Club	0.1341	0.1224	0.0117
Cow	0.1651	0.0700	0.0951
Rabbit	0.1260	0.1015	0.0245
Bear	0.1522	0.0713	0.0809
Pot	0.0806	0.0612	0.0194
Horse	0.2124	0.1347	0.0778
Cat	0.1136	0.0962	0.0174
Duck	0.1188	0.1066	0.0121
Elephant	0.2126	0.1101	0.1025
Stegosaurus	0.1507	0.0913	0.0594
Mean	0.1424	0.1006	0.0419
<i>SD</i>	0.0451	0.0291	0.0362

779

780

781 Appendix A

782

783 ***Gabor_{left}*** and ***Gabor_{right}*** represent the oriented receptive field of the model simple
 784 cells for the left and right images, respectively:

$$\mathbf{Gabor}_{left_{\theta, \phi_i}}(x, y) = \frac{1}{2\pi} \cos\left(2\pi \frac{(x - x_0) \cos(\theta) - (y - y_0) \sin(\theta)}{\lambda} + \phi_i\right) * e^h,$$

785

Eq. 17

$$\begin{aligned} \mathbf{Gabor}_{right_{\theta, \phi_i, \psi_j}}(x, y) \\ = \frac{1}{2\pi} \cos\left(2\pi \frac{(x - x_0) \cos(\theta) - (y - y_0) \sin(\theta)}{\lambda} + \phi_i + \psi_j\right) * e^h, \end{aligned}$$

786

787

Eq. 18

788 where

$$789 \quad h = - \left[\left\{ \frac{(x - x_0) \cos(\theta) - (y - y_0) \sin(\theta)}{\sigma_x} \right\}^2 + \left\{ \frac{(x - x_0) \sin(\theta) + (y - y_0) \cos(\theta)}{\sigma_y} \right\}^2 \right],$$

790 where x_0 and y_0 represent the center of the Gabor filters, and θ , λ , σ_x , and σ_y show
 791 the orientation, wavelength and SDs of the Gabor filters, respectively. ϕ_i represents the
 792 phase of the left receptive field, and ψ_j represents the ocular difference in phase. We set
 793 λ , σ_x and σ_y to 20, 8 and 8 pixels, respectively, so as to mimic V1 cells (λ and $\sigma_x(\sigma_y)$
 794 equal to 0.5 and 0.2 degree in visual angle, respectively). ϕ_i ($i = 1-4$) were set to 0, $\pi/2$,
 795 π and $3\pi/2$. ψ_j ($j = 1-11$) ranged between 0 and 2π in increments of $\pi/11$. The size
 796 (spatial extent) of the Gabor filter was set to 40×40 pixels (1×1 degree) so as to be
 797 consistent with that of the receptive field of V1 neurons [25].

798