Functional Parts

Joshua Tenenbaum
Department of Brain and Cognitive Sciences
Massachusetts Institute of Technology
Room E10-120
Cambridge, Massachusetts 02139
jbt@psyche.mit.edu

Abstract

Previous work in visual cognition has extensively explored the power of parts-based representations of objects for recognition, categorization, and functional reasoning. We propose a novel, parts-based representation of objects, where the parts of an object are found by grouping together object elements that move together over a set of images. The distribution of object configurations is then succinctly described in terms of these functional parts and an orthogonal set of modal transformations of these parts. If the distribution has a natural set of principal axes, the computed modes are stable and functionally significant. Moreover, the representation is always unique and robustly computable because it does not rely critically on the properties of any particular element in any particular instance of the object. Most importantly, the representation provides a set of direct cues to object functionality without making any assumptions about object geometry or invoking any high-level domain knowledge. This robustness and functional transparency may be contrasted with standard representations based on geometric parts, such as generalized cylinders (Marr and Nishihara, 1978) or geons (Biederman, 1987), which are sensitive to accidental alignments and occlusions (Biederman, 1987), and which only support functional reasoning in conjunction with high-level domain knowledge (Tversky and Hemenway, 1984).

Geometric Parts and Functional Parts

Previous work in visual cognition has extensively explored the power of parts-based representations of objects. Representations that make explicit the parts of objects and the relationships between parts have been hypothesized to underlie visual object recognition (e.g. Marr and Nishihara, 1978; Hoffman and Richards, 1982; Biederman, 1987; Shapira and Ullman, 1991; Dickinson, Pentland, and Rosenfeld, 1992), as well as more abstract operations such as categorization and reasoning about function from structure (e.g. Tversky and Hemenway, 1984; Stark and Bowyer, 1991). Conventional notions of *part* in visual cognition are geometric. Objects in a static scene are typically segmented into parts at points of extreme curvature, either cusps or concavities, and a set of deformable models are then fit to the extracted parts. Proposed models for geometric parts include two-dimensional contour primitives such as codons (Hoffman and Richards, 1982), and three-dimensional volumetric primitives such as generalized cylinders (Marr and Nishihara, 1978), superquadrics (Pentland, 1986), or geons (Biederman, 1987).

Tversky and Hemenway (1984) argue that parts bridge the gap between structure and function because they are both perceptually salient and functionally significant aspects of objects. But in general, the geometric parts that are perceptually salient in a static scene do not necessarily correspond to the functional aspects of objects in a simple way. For instance, compare the drawings of two everyday objects in Figure 1. The geometric parts of a hand and a fork are very similar, consisting in both cases of a broad body and several long, thin projections. But some crucial functional information is missing from this static scene. Whereas the tines of a fork are all rigidly joined to the body of the fork, the fingers of a hand can move independently of the palm and of each other, to assume a wide range of possible configurations. That is, the geometric parts of a fork always function together, but the geometric parts of a hand can and often do function independently. We could say that the hand has several distinct functional parts, while the fork really has only one functional part.

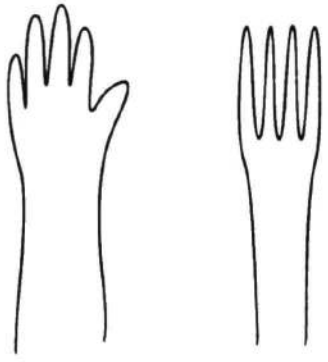

Figure 1. A hand and a fork have similar geometric part structures, but different functionality.

This distinction between geometric parts and functional parts is fundamental to how we perceive, think about, and even talk about objects. The Gestalt psychologists showed that even very dissimilar or disconnected visual elements may be grouped by correlated motion, or "common fate" (Kohler, 1947). Developmentally, infants parse the world into objects based on functional properties such as correlated motion before they are sensitive to object boundaries defined by static, geometric properties (Spelke, 1990). Our adult language also supports the distinction between functional and geometric parts. Consider the linguistic representations of

the parts of a hand and the parts of a fork. "Finger" and "palm" are common, basic-level terms, whereas "tine" is a much less ordinary term, and there is no term to describe the "palm" of a fork. Moreover, the individual fingers of a hand have individual, functionally-oriented names, such as "index finger", "pointer", or "ring finger", while one never even refers to an individual tine of a fork. We do not mean to push this point on language too far, but simply to motivate the proposal that functional parts, and not just geometric parts, should provide the link between vision and cognition.

Now, if the functional parts of objects are not perceptually salient in scenes such as Figure 1, must we abandon the attractive position of Tversky and Hemenway (1984), that parts bridge the gap between structure and function by virtue of their perceptual salience and functional significance? No, but we do have to revise what we mean by "perceptually salient," to include information that is unavailable in single, static images, and that only becomes salient over a set of images. This paper proposes a novel, parts-based representation of objects, where the parts of an object are found by grouping together object elements that move together over a set of images. In the simplified, two-dimensional cases considered here, "object elements" are edge elements, or "edgels" for short, but in the more general, three-dimensional case, functional parts may be extracted by grouping surface elements that move together. This approach assumes that the problem of edgel correspondence across the set of images has already been solved by lower level processes, so that we know which edgels correspond in any two configurations of an object.

Figure 2 illustrates the basic idea: a simplified, two-fingered hand assumes several different configurations as the fingers pivot, and since the fingers can move independently, the orientations of the edgels are perfectly correlated only within each finger. When a fork assumes different configurations, however, the tines cannot pivot independently, so the orientations of all edgels are always perfectly correlated. By looking at the covariance patterns of edgel orientations, the visual system can infer the different functional part structures of the hand and the fork.

Of course, there is more to function than just correlated motion. A part's geometric shape certainly places strong constraints on its possible functions. Nonetheless, correlated motion is one of the most direct, and hence one of the most important, visual cues to functionality. The links between shape or texture and function may be somewhat subtle, but motion indicates the basic units of function without requiring any high-level knowledge. Functional part-based representations should not replace geometric part-based representations, but rather supplement them for the purposes of functional reasoning. The goal of this preliminary paper is to demonstrate how much useful information about function can be extracted from patterns of edgel covariance, without making any assumptions about object geometry or invoking any high-level domain knowledge.

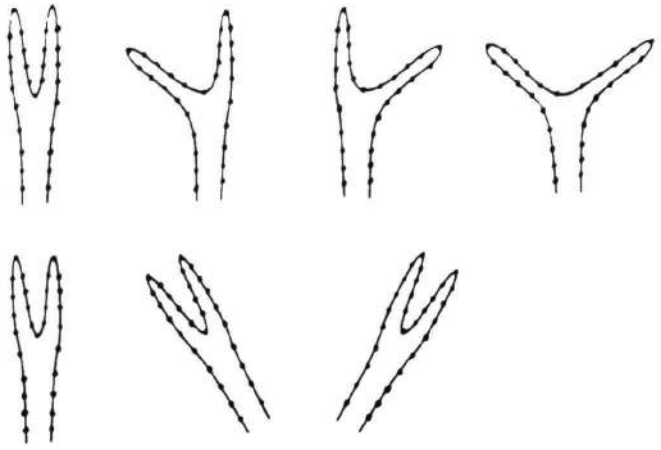

Figure 2. As functional parts are articulated by rotating, the orientations of their component edgels covary. The orientations of hand edgels are only correlated within each finger, while the orientations of all fork edgels are perfectly correlated. Dots schematically separate individual edgels along contours.

Constructing Functional Representations

We begin with a set Σ of N instances of an object in various configurations (e.g. the set of two-fingered hands in Figure 2). We represent the ith instance of an object as a vector, $\mathbf{v}_i$, of edgel properties. For example, $\mathbf{v}$ might be a list of edgel lengths and orientations. If the object consists of K edgels, each with P properties, then $\mathbf{v}_i$ has dimension KxP, and the set of instances is a (KxP) by N matrix, defined by $[\mathbf{v}_1\ \mathbf{v}_2 \ldots \mathbf{v}_N]$. Ultimately, we seek a representation that most simply accounts for the variability in object configuration across Σ. Let $\bar{\mathbf{v}}$ be the average instance, defined by

$$\bar{\mathbf{v}} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{v}_i$$

The ith instance then differs from the average by $\mathbf{u}_i = \mathbf{v}_i - \bar{\mathbf{v}}$. Then the full variability across is embodied in the edgel covariance matrix

$$\mathbf{C} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{u}_i\mathbf{u}_i^T$$

However, a functional parts-based representation can describe this variability across Σ much more succinctly, by making only two weak assumptions about how an object's configuration may vary.

1. An object is composed of one or more functional parts, each of which always moves or deforms as a unit.

2. When a part moves or deforms, i.e. when a parameter of a part changes, a corresponding change occurs in some parameter of each of the part's component edgels.

These two conditions respectively provide the motivation and

the basis for extracting functional parts by grouping together edgels that move together. Note first that the edgel parameters must reflect the transformations by which we expect functional parts to be articulated. If parts are articulated by rotation, then we need to represent edgel orientation explicitly in the instance vectors $\mathbf{v}_i$, because the orientation of all edgels in a part will covary as the part rotates. If parts are articulated by translation, then we need to represent edgel position; if parts are articulated by stretching, then we need to represent edgel length; and so on. Functional parts become simply clusters of covariance. Then, instead of characterizing the variability across Σ in terms of how the properties of each edgel covary with the properties of every other edgel, as in the full edgel covariance matrix $\mathbf{C}$, we group edgels into functional parts and describe the same variability in terms of how the properties of these parts covary. We use the term function in a very specific sense to refer to the allowed relations between functional parts that determine the possible configurations of the object. Since the number of functional parts will generally be much less than the number of edgels, a representation in terms of functional parts and functions amounts to a significant simplification of the full edgel covariance matrix $\mathbf{C}$.

Mathematically, we simplify $\mathbf{C}$ by means of principal components analysis (PCA). PCA and other modal representations of shape have recently become popular in the computer vision literature (Cootes *et al.*, 1992; Sclaroff and Pentland, 1993), but this paper is the first effort to make functional parts and functions explicit through PCA. If all the edgels within a single functional part move as a unit, then all the rows (and columns) of $\mathbf{C}$ corresponding to these edgels will be essentially the same. Thus, $\mathbf{C}$ will have a dominant block structure which can be extracted by finding its dominant eigenvectors. The eigenvectors will correspond to functions and will articulate the individual functional parts.

The easiest way to appreciate the special structure of $\mathbf{C}$ is to consider a simple example. Figure 3 shows six instances of a stick-figure hand with two "fingers". The orientation of the left finger is randomly distributed about -30° with a variance of 30°, and the orientation of the right finger is randomly (and independently) distributed about +30°, also with variance 30°. Figure 4 shows the distribution of finger orientations for 30 such objects. Each finger in Figure 3 is represented by ten edgels, with constant length but varying orientation. Because parts are only articulated by rotation, we only need to keep track of the orientation of each edgel. So each instance is represented by a 20 dimensional vector of edgel orientations. In order to demonstrate that the representation is robust in the presence of perceptual noise, the orientation of each edgel is randomly perturbed about the true finger orientation, with a variance of 7.5° (25% of the whole finger variance).

The full edgel covariance matrix, $\mathbf{C}$, is depicted graphically in Figure 5. The block structure corresponding to the two functional parts is clearly evident, and these functional parts can be extracted by standard techniques of cluster analysis.

Figure 3. Six instances of a stick-figure hand with two fingers. Dots separate the 20 individual edgels.

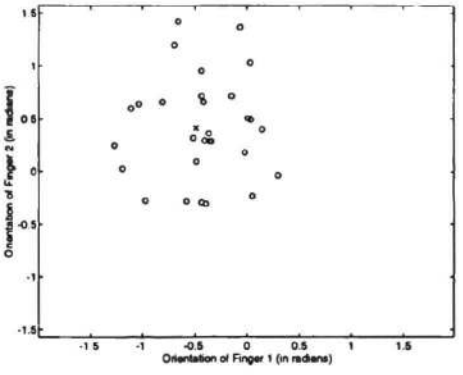

Figure 4. Distribution of finger orientations for 30 two-fingered hands. The sample mean is indicated by 'x'.

Having found the functional parts, we can then completely characterize the variability in the original set of 30 instances merely by specifying how these two parts interact. PCA provides a natural framework for this step. When the eigenvectors of $\mathbf{C}$ are computed, we find that the first two principal components dominate the other modes, accounting for over 95% of the variance in the original data. This is to be expected, given that we began with essentially two independent degrees of freedom, plus noise. Figure 6 shows how the mean shape deforms in the two significant modes. Note that the modes articulate the two functional parts, reflecting the block structure of $\mathbf{C}$. Starting from only the orientations of 20 edgels in each of 30 object configurations, we have recovered a concise representation of the object's functionality in terms of its two functional parts and a set of two orthogonal transformations of these parts.

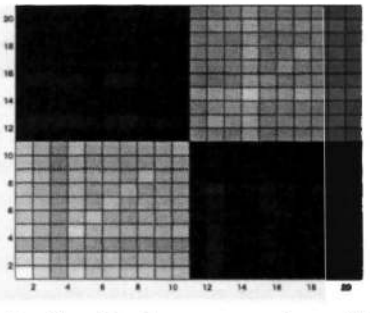

Figure 5. Graphical representation of $\mathbf{C}$, the edgel covariance matrix for the two-fingered hand. For each pair of 20 edgels, the shading of a square indicates the magnitude of covariance. White indicates maximal covariance, black indicates weak covariance.

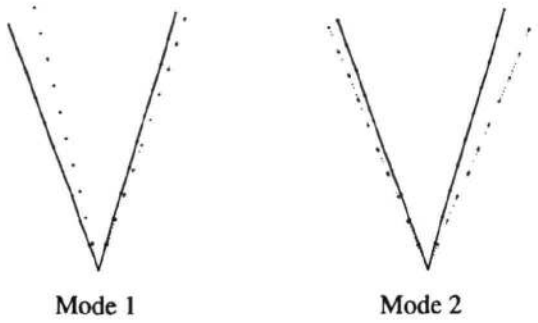

Figure 6. The two significant principal components of the distribution in Figure 4, depicted as deformations of the mean shape.

Functional Modes

The modes, or principal components, of C account for how the functional parts interact to produce the range of object configurations observed in Σ. Ideally, we would like to identify these modes with functions; that is, we would like the observed modes to have some functional significance. Looking at Figure 6, it is not at all clear what functional significance can or should be attributed to these modes. We might view the individual modes as roughly representing the function of the hand as a whole, with the first mode representing "hand orientation" and the second mode representing "hand shape," the relative angle between the two fingers. But this interpretation is not particularly compelling, because while both fingers do participate in each mode, they do not participate equally. In mode 1, the left finger pivots significantly more than the right finger, and vice versa for mode 2. Thus neither mode truly reflects a simple parameter of the whole hand. Alternatively, we might view the individual modes as roughly representing the function of individual parts, with mode 1 being the "left finger" mode and mode 2 being the "right finger" mode. But this interpretation is even less compelling, because although one finger predominates in each mode, both fingers also participate significantly in each mode. In fact, not only do the computed modes for this object defy a straightforward functional interpretation, they are also quite unstable. The modes are liable to be completely different if we compute them again for a different set of 30 instances drawn from the same distribution.

The functional ambiguity and the instability of the modes are deeply related. Both difficulties result from the fact that the generating distributions of the object's two functional parts are completely independent and have equal variance. PCA will only give a stable, functionally meaningful solution when the distribution of object configurations scatters more in some directions than in others, giving a natural set of principal axes. The functional character of the solution will depend on the orientation of this distribution, which is not well-determined for data that scatter roughly equally in all directions, as in Figure 4.

When the distribution of object configurations strongly reflects a single functional structure, PCA will find it. Consider a similar two-fingered hand in which both fingers still move independently, but each finger has a different, characteristic way of moving. Specifically, the joint on finger 1 is stiffer than the joint on finger 2, so that the distribution of finger 1 orientation has only half the variance (15°) of the finger 2 distribution (30°). The bivariate distribution for 30 instances of this object is shown in Figure 7. Now the data have a natural set of principal axes, corresponding to the two fingers. As expected, the computed modes align with these axes and individually articulate the two functional parts (Figure 8).

Now consider the case in which the two fingers have equally stiff joints, but they are not articulated independently. For example, the whole hand may pivot about the vertical with a variance of 30°, but the hand shape, i.e. the angle between the two fingers, may have a variance of only 15°. The bivariate distribution for 30 instances of such an object is shown in Figure 9. Now the natural set of principal axes corresponds to two properties of the whole hand, rather than to properties of the individual fingers, which are not functionally independent units of this object. As expected, the computed modes articulate these two functional properties, hand orientation in mode 1 and hand shape in mode 2 (Figure 10).

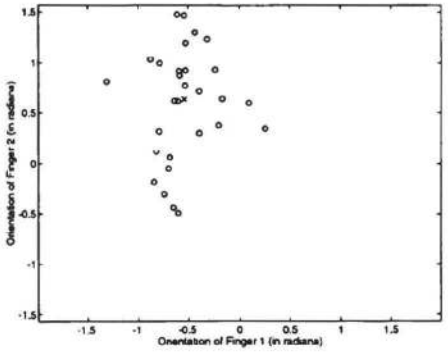

Figure 7. Distribution of finger orientations for two-fingered hands, with finger 1's joint stiffer than finger 2's joint.

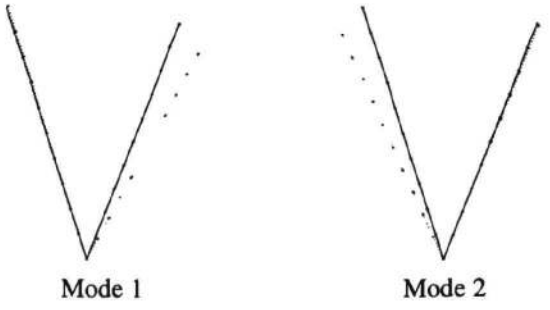

Figure 8. The two significant principal components of the distribution in Figure 7, depicted as deformations of the mean shape.

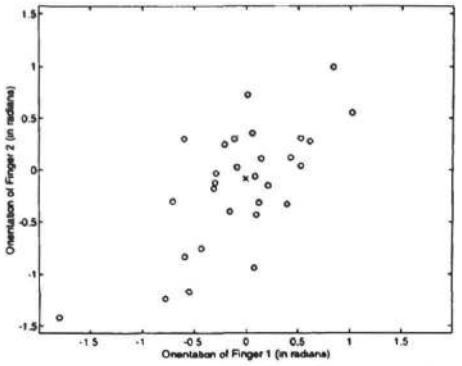

Figure 9. Distribution of finger orientations for two-fingered hands, with the variance in hand orientation significantly greater than the variance in hand shape.

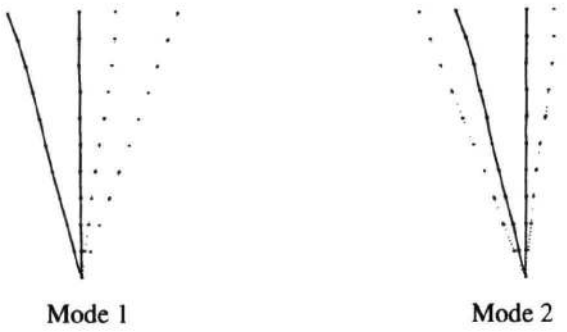

Figure 10. The two significant principal components of the distribution in Figure 9, depicted as deformations of the mean shape.

A Non-Trivial Example

Finally, in case these examples seem too simplistic, we give a non-trivial example which demonstrates the power of a representation based on functional parts. Figure 11 shows 6 instances of a jointed rod composed of six, separately moving segments. The first segment is constrained to lie within 15° of the horizontal, and each successive segment is constrained to lie within 15° of the previous segment. Each segment is represented by ten edgels, again with constant length but varying orientation, and thus each instance is represented by a vector of 60 edgel orientations. The orientation of each edgel is randomly perturbed about the true segment orientation, with a variance of 3°. Because these six-segment objects have more inherent variability than the two-fingered hands, we carry out our analysis on a larger population of 180 instances.

Figure 12 shows the covariance matrix and Figure 13 shows the six significant principal components. Notice that the modes correspond roughly to the physical modes of string vibration, and that they are ordered by frequency. Also note that individual segments are only weakly articulated in the strongest modes, and consequently the full edgel covariance matrix does not have the desired block structure. However, we can obtain the desired block structure by reweighting the contribution to **C** of each principal mode to be independent of eigenvalue. Normally, we have $\mathbf{C} = \Phi\,\Omega\,\Phi^T$, where Φ is an orthonormal matrix of eigenvectors and Ω is a diagonal matrix of eigenvalues, ordered by frequency of vibration, which weights the contribution to **C** of each eigenvector in Φ. Suppose that by thresholding on eigenvalue, we have determined that only the first M modes are functionally significant. Then define Ω' by replacing the first M diagonal entries of Ω with 1, and the remaining diagonal entries with 0. The resulting matrix $\mathbf{C}' = \Phi\,\Omega'\,\Phi^T$, weights the contribution of each functional mode equally, regardless of eigenvalue, but completely eliminates higher frequency modes that are presumably due to noise. If the between-part variations are significantly larger than the within-part variations, then this eigenvalue threshold will ensure that **C'** displays the desired block form that makes these parts explicit. Figure 14 shows that this transformation indeed captures the functional part structure of the six-segment jointed rod. Finally, we note that although discussion in this paper is restricted to a linear analysis of simple artificial objects, ongoing work is providing evidence that functional part-based representations play a crucial role in visual cognition for real, everyday objects such as hands and faces. We are also exploring the application of nonlinear dimensionality reduction techniques (e.g. Hinton and Zemel, 1994), in order to extend the range of potential part structures beyond the limitations of PCA.

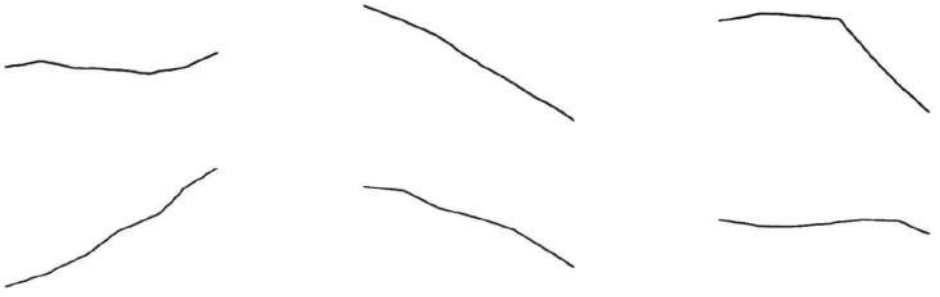

Figure 11. Six instances of a six-segment, 60-edgel rod. Each segment is constrained to lie within 15° of the preceding segment. For clarity, individual edgels are not shown.

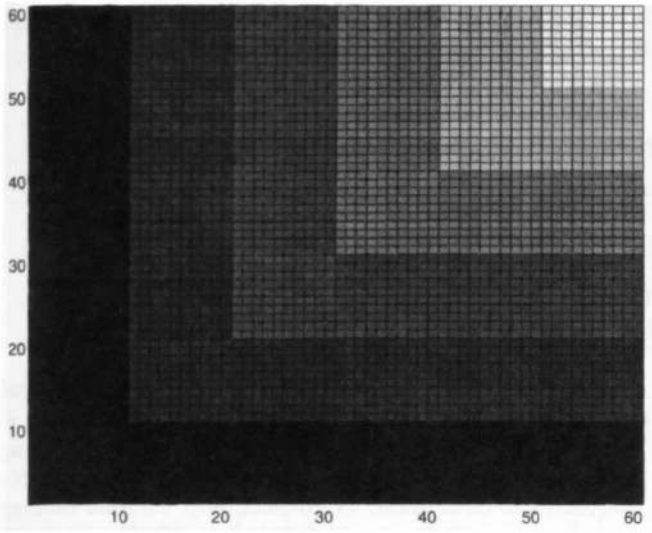

Figure 12. Graphical representation of **C**, the edgel covariance matrix for the six-segment, 60-edgel rod.

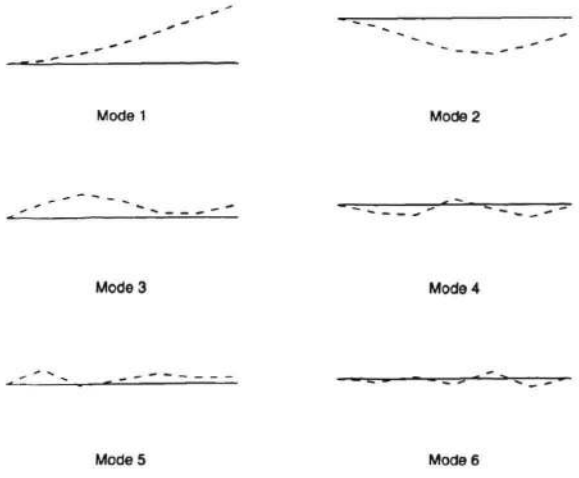

Figure 13. The six significant modes of the covariance matrix in Figure 12, depicted as deformations of the mean shape. Note that the modes are naturally ordered by frequency, with the dominant modes expressing the lowest frequency deformations.

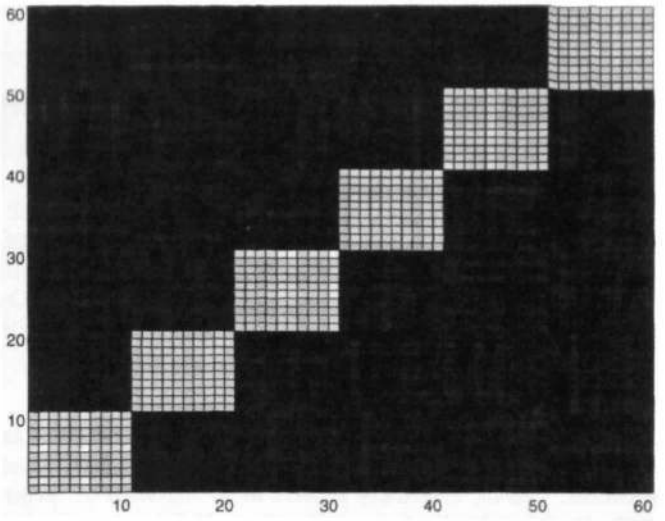

Figure 14. **C'**, the reconstructed covariance matrix with the contributions of all six functional modes weighted equally, independently of eigenvalue.

Acknowledgments

Emanuel Todorov, Yair Weiss, Stephen Gilbert, Chris Moore, Sandy Pentland and Whitman Richards provided many helpful suggestions during the course of this work. Joshua Tenenbaum is a Howard Hughes Medical Institute Predoctoral Fellow.

References

Biederman, I. 1987. Recognition-by-components: a theory of human image understanding. *Psychological Review*, 94, 115-147.

Cootes, T., Taylor, C., Cooper, D., & Graham, J. 1992. Training models of shape from sets of examples, in *Proc. BMVC 1992* (pp. 9-18). Springer-Verlag.

Dickinson, S., Pentland, A., & Rosenfeld, A. 1992. 3-D shape recovery using distributed aspect matching. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 14, 174-198.

Hoffman, D. & Richards, W. 1982. Representing plane curves for visual recognition. MIT A.I. Lab Memo 630.

Hinton, G. & Zemel, R. 1994. Autoencoders, minimum description length and helmholtz free energy, in J. Cowan, G. Tesauro, and J. Alspector (Eds.), *Advances in Neural Information Processing Systems 6*. San Mateo, CA: Morgan Kauffman.

Kohler, W. 1947. *Gestalt psychology*. New York: Liveright Publishing Co.

Marr, D. & Nishihara, K. 1978. Representation and recognition of the spatial organization of three-dimensional shapes. *Proc. R. Soc. Lond. B*, 200, 269-294.

Pentland, A. 1986. Perceptual organization and the representation of natural form. *Artificial Intelligence*, 28, 293-331.

Sclaroff, S. & Pentland, A. 1993. Modal matching for correspondence and recognition. MIT Media Lab Perceptual Computing Section Technical Report No. 201, May 1993.

Shapira, Y. & Ullman, S. 1991. A pictorial approach to object classification, in *Proc. 12th International Joint Conference on Artificial Intelligence*. San Mateo, CA: Morgan Kauffman.

Spelke, E. 1990. Origins of visual knowledge, in D. Osherson, S. Kosslyn, & J. Hollerbach (Eds.), *An Invitation to Cognitive Science*, vol. 2. Cambridge, MA: MIT Press.

Stark, L. & Bowyer, K. 1991. "Form and function": a theory of purposive, qualitative 3-D object recognition, in Y. Feldman & A. Bruckstein (Eds.), *Artificial Intelligence and Computer Vision*. Elsevier.

Tversky, B. & Hemenway, K. 1984. Objects, parts, and categories. *Journal of Experimental Psychology: General*, 110, 169-191.