The representation of three-dimensional object similarity in human vision

F. Cutzu and M. J. Tarr

Dept. of Cognitive and Linguistic Sciences, Brown University
Providence, Rhode Island 02912 USA

ABSTRACT

Outline shape carries a substantial part of the information present in an object view, and is more economical than classical representations such as geon-structural-descriptions or multiple-views. We demonstrate the utility of silhouette representations for a variety of visual tasks, ranging from basic-level categorization to finding the best view of an object. All these tasks necessitate the computation of silhouette similarity. We present an algorithm for estimating silhouette similarity and apply it to a number of simple but realistic vision problems.

Keywords: silhouettes, similarity, shape classification, natural categories

1. INTRODUCTION

1.1. Object identification: compensating for variable pose

The human visual system possesses an impressive ability to recognize and categorize complex three-dimensional objects under widely different viewing conditions. Although there are many sources for variability in the image, the majority of psychophysical studies have focused on how recognition performance is affected by viewpoint change. The picture that has emerged from these studies is however, not entirely clear (see Ref. 1 for a recent review). Some researchers found recognition performance to be essentially viewpoint-invariant[2,3]; on the contrary, others demonstrated significant effects of object rotation on recognition accuracy and speed.[4–8] Given these seemingly contradictory findings, the nature of the internal representations underlying the ability to recognize rotated objects has remained a point of controversy.

Drawing on the psychophysical results, a substantial variety of 3D object recognition and representation theories has been proposed in computational vision (see Refs. 9,10 for recent reviews). Simplifying somewhat, one can distinguish two main families of visual object representation schemes: theories predicting viewpoint-invariant recognition and theories predicting viewpoint-dependent recognition; basically, the proposed models differ either on the format of the proposed representation (either 3D-part-based or image-based) or on the coordinate system in which the representation is described (viewpoint dependent vs. viewpoint invariant). At one end of the gamut of theories of representation lie 3D, object-centered qualitative schemes (e.g., structural graphs composed of generic volumetric parts and spatial relationships), which are invariant to pose by design.[11,2] At the opposite end are the 2D, viewer-centered models treating recognition as a problem of interpolation of the space of all views of an object from a relatively small number of known data points.[12–14]

1.2. Object categorization: compensating for variable shape

All proposed representation schemes, however varied, are in fact different solutions for the problem of object identification under variable viewpoint. The identification of familiar objects under variable viewing conditions is only one of the challenges facing the human visual system. Tasks of at least equal ecological importance are the construction of categories and the classification of novel objects. Accomplishing these objectives requires the computation of similarity among internal representations (construction of categories) and between percepts and internal representations (classification of novel inputs). Few of the current object recognition schemes deal explicitly with these problems (although see 2); essentially, these models solve for input variability in *view space* rather than in *shape space*. In object identification the computational goal is achieving *object constancy*,[15,1] i.e. recognizing a previously seen object

Send correspondence to F.C. Email: florin@cog.brown.edu despite random image variations induced by variable pose or illumination. In categorization on the other hand, the goal is to compute similarity and assign distinct objects to the same class despite individual shape variations.

Numerous studies in cognitive science (see Rosch et al., 1976 for a review) reveal that in the hierarchical structure of object categories there exists a certain level, called basic level, which is the most salient according to a variety of criteria, such as the ease of recognition and preference of access. Taking as an example the hierarchy "quadruped, mammal, cat, Siamese," the basic level is that of "cat." A recognition task that implies more detailed distinctions than those required for basic-level categorization is said to occur at the subordinate level. Objects in the same basic-level (or subordinate-level) category tend to be *visually similar*,[17] and therefore such a category can be represented pictorially by averaging the images of its members.[16] Psychophysical experiments indicate that basic level recognition is an effortless act, largely unaffected by viewing conditions such as viewpoint.[2] Objects can be accurately categorized even when only impoverished visual information is provided – such as silhouettes or noisy images.[16]

1.3. Similarity and categorization

In contrast to computational vision research, similarity and categorization have been traditional subjects in mathematical psychology. A basic concept in similarity research is the notion of a multidimensional feature (psychological) space, upon which the geometrical model of similarity and the multidimensional scaling (MDS) techniques are based.[18–20] Postulating that perceptual similarities decrease monotonically with corresponding psychological space distances (the geometrical model of similarity), MDS algorithms[20] derive the feature space configuration of the stimulus-points from the table of the observed similarities. In this theoretical framework, represented objects are encoded as vectors of continuous features in a N-dimensional *psychological space* endowed with a metric (usually Euclidean or city-block).

The idea of expressing similarity as a decaying function of feature-space distance has been incorporated in Nosofsky's Generalized Context Model (GCM),[21,22] which is an exemplar-based model of categorization and recognition based on the MDS model. A central assumption of the GCM theory is that while classification is based on comparing summed similarities to exemplars of alternative categories, recognition depends on the summed similarity of the input item to all stored exemplars from all categories. The other major model of similarity and categorization based on the psychological space concept is the General Recognition Theory (GRT) developed by Ashby and his collaborators.[23,24] In the GRT scheme, the stimuli are represented by multivariate normal distributions in psychological space, rather than by single points; GRT can be viewed as an extension of the classical signal detection theory[25] to the multidimensional case.

Dissimilarities can not always be interpreted as distances in a metrical feature space, as assumed by the geometrical model of similarity. For example, Tversky (1977) showed that in certain cases perceptual similarities can be asymmetric, or that the triangle inequality can be violated. Such situations typically arise when subjects are operating in the semantic rather than the perceptual domain – comparing, for example, countries or people, and/or when one stimulus is much more prominent than the others. In these cases the geometrical, continuous feature MDS model of similarity must be abandoned in favor of set-theoretic models based on discrete features.[27,26] In general, however, when the stimuli have continuous features (as we assume for visual stimuli), are of equal salience, and are compared in a consistent context the metric model is appropriate and multidimensional scaling can be applied. The stimuli used in this study meet these conditions.

A general limitation of virtually all visual similarity research is the highly simplified nature of the experimental stimuli used to test the theories: semi-circles,[28] schematic faces,[29] dot patterns, semi-circles, rectangles.[30] Such 2D shapes have simple and obvious features (width and height for rectangles, radius and orientation for semi-circles), and thus the feature-space models of similarity can be applied with little difficulty. It is unclear whether this class of models can describe visual similarities for complex three-dimensional shapes.

2. PURPOSE

A comprehensive theory of object representation and recognition must provide an unified account of the effects of both pose and shape change on human performance. To date, the problem of recognizing novel objects at the category level has been largely neglected in computational vision; the present work represents an attempt to fill this gap (but see Ref. 31). The existing models of object recognition, for the most part, are by construction unsuitable for shape classification. Ironically, for artificial vision systems object identification is a simpler task than object classification, while for biological vision systems the reverse appears to be true (Ref. 10, pp. 159-160).

Categorization at the subordinate and basic levels may be based on visual similarity; therefore, to explain categorization, one must understand the computational basis of visual similarity. As explained above, the approach taken in categorization research is based on the concept of a stimulus feature space. Unfortunately, defining the complete set of features for the representation for complex 3D objects (as opposed to rectangles or semi-circles) is a computationally difficult endeavor. In contrast, it is believed that human visual categorization is a simple, elementary process.[2,16] Basic-level classification can be viewed as a first, rough step in a process leading to precise object identification. Recognizing only the class an input shape belongs to may suffice in many situations (recognizing a "cat" in the jungle is specific enough), therefore this first step is critical and is preferably carried out by simple, fast algorithms. In the present work we propose such an algorithm. Our results indicate that by computing similarities among *object silhouettes*, one can classify various objects at the basic level irrespective of their pose; when silhouettes of canonical views are used, even finer discriminations become possible.

3. RELATED WORK

Outline shape carries a substantial part of the information present in an object view. Data from a number of psychophysical studies indicates that silhouette shape information is important (and may be even sufficient) for successful object recognition. Indeed, Rosch et al. suggested that humans may be able to accurately categorize objects when shown only silhouettes.[16] Several recent studies[32,33] have demonstrated this – indicating that recognition performance can be predicted from changes in the shape of stimuli outline alone.

Computer vision researchers have long been aware of the utility of silhouettes for classification and recognition. A classical approach is based on the realization that the outline is a closed curve and thus can be written as a periodic function. The Fourier coefficients of this function can be used to represent the contour shape. Some schemes use the angle of the tangent to the curve versus arc length as the periodic contour function; others use the the sequence of the (x, y) coordinates of the contour pixels in the complex form $x + iy$ as to generate a periodic complex function. The discrete Fourier transform (DFT) of the periodic contour function is termed the Fourier Descriptor (FD) of the curve. Both the magnitudes and the phases of the DFT can be used, but in many cases the amplitude information suffices for rough shape classification. Fourier domain (or space domain) normalization steps are required to ensure the invariance of the FD to curve size and orientation and to contour starting point (the latter is necessary if phase information is to be used). The FD constitute a feature vector describing the contour; the similarity between two contours can be defined as the Euclidean distance in the space of Fourier coefficients.

Fourier descriptors of closed curves are invariant to rotations, scaling, translations[34,35]; recent models achieve invariance to more general, affine transformations.[36] In practical applications, Fourier descriptors have been successfully used to recognize handwritten characters[37] and airplanes based on their outline over a restricted range of viewpoints.[36] A number of psychophysical studies have shown that perceptual similarities for sets of simple synthetic 2D curves are consistent with distances in FD space.[38,39]

Another possibility is to compute a measure of area overlap between two shapes brought in register by a optimal similarity transformation; for example, such an approach is taken in Ref. 40.

A recent computational study by Ullman and Yolles (Ref. 10, pp. 172-178) employed a simple silhouette-based distance measure for the classification of cartoon-like images of common objects. The bounding object contours were represented as sets of elementary line segments characterized by two features, namely position and orientation. The algorithm, following the alignment of the bounding boxes of the two pictures, computed a feature space distance between the contours. Correspondence between contour elements was defined in by finding the minimal feature space distance. Computational experiments demonstrated that this pictorial classification system correctly assigned novel object images to the corresponding object classes stored in the database in most instances.

The silhouette similarity model used in this paper, presented in the next section, is related to the Ullman-Yolles scheme.

4. COMPUTATION OF SILHOUETTE SIMILARITY

Our interest in silhouettes is motivated by the idea that complete image analysis is unnecessary for structuring a high-dimensional shape space for recognition.[14] In particular, it is our hypothesis that a space sufficient for recognition at several levels of specificity may be organized on the basis of silhouette similarity. It is not our contention, however, that *only* the silhouette is represented – points in a silhouette-defined representation space may encode many properties of the object missing from the silhouette, such as texture or color. On the other hand, information in silhouettes may be sufficient for some recognition tasks, for example, basic-level classification (more detailed information may be necessary to discriminate between instances of a class). It is the validity of this hypothesis we hope to examine in our computational experiments.

When estimating the similarity between two shapes, it is desirable that irrelevant factors such as position, scale, orientation are eliminated. The silhouette similarity model used in this study is based on an affine distance measure[41] among point sets, which represents a generalization the Procrustes distance.[42] These distance measures are insensitive to similarity transformations (the Procrustes distance) and to affine transformations (the affine distance measure).

The idea of the similarity computation algorithm is simple: align the two silhouettes point sets by the best affine transformation and take the residual mismatch as a measure of dissimilarity.

As a first step, the silhouette was extracted from the object image using a standard boundary tracing algorithm. The (x, y) coordinates of the boundary pixels were then mapped, by linear interpolation, to an array of standard size (200×2), typically smaller than the length of the image contour. As a result, the silhouette was encoded by a sequence of 200 equally spaced (x, y) pairs. This normalization of the number of silhouette points allows on the one hand the computation of distance among boundaries of different lengths and on the other hand realizes a smoothing of the original contour extracted from the image.

To normalize the silhouette representation with respect to translations, the center of mass of the boundary point set was shifted to the origin of the coordinate system.

Next, all silhouettes were normalized with respect to their inertia moments. If the silhouette is represented by the array of boundary point coordinates $A = \begin{pmatrix} x_1 & \dots & x_n \\ y_1 & \dots & y_n \end{pmatrix}$, the matrix associated with the inertia tensor is given by $T(A) = \begin{pmatrix} \sum y_i^2 & -\sum y_i x_i \\ -\sum y_i x_i & \sum x_i^2 \end{pmatrix}$.

The diagonal elements of T are the moment of inertia coefficients and the off-diagonal elements are termed products of inertia. T is symmetrical, has real valued components and hence is hermitean and has real eigenvalues. Its eigenvectors are orthogonal and are termed the principal axes.

A linear transformation L was applied to A to render its inertia matrix unitary. The action of L can be decomposed in two steps. First, the silhouette A was rotated in the image plane to align its principal axes of inertia to the coordinate system axes; following this rotation step, the products of inertia become equal to zero. The rotated silhouette was then scaled independently along the X and Y directions to make the the two principal moments equal to one. Thus, for $a = LA$, the rotated and scaled version of A, the inertia tensor is characterized by a unitary matrix $T(a) = I$. This normalization step was necessary to insure that the distances computed following the next, alignment step, are symmetric.

Let a and b denote two normalized contours. The goal of the alignment stage was to match a and b by finding the best linear transformation P, with $a' = Pa$. "Best" was defined in terms of minimizing the distance $||a' - b||$ over all linear transformations and point correspondences between a and b. For a given correspondence, P is given by using the pseudo-inverse p^+, $P = bp^+$. Thus, finding the best P entails searching along the contour for the best correspondence.

Finally, the distance between the contours a and b is given by $D^2(a, b) = ||Pa - b||^2$; it represents the residual squared error after a and b have been brought into register by the best linear transformation. It can be shown that $D(a, b)$ is a metric.[41]

The scheme described here is closely related to the so-called Procrustes algorithm.[42] The main difference between the affine and Procrustes distance is that under the Procrustes scheme only similarity transformations can be used for aligning the two shapes.

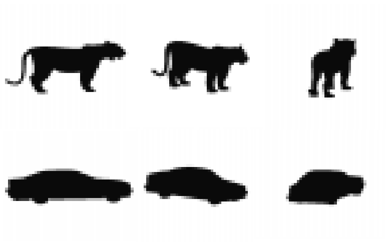

Figure 1. Examples of silhouettes used in the basic-level categorization experiment. Each object was imaged from three viewpoints.

5. RESULTS

The similarity measure defined in the preceding section was tested on contours extracted from images of common 3D objects. In the first series of computational experiments we explored basic-level categorization, a task in which humans perform with ease and accuracy and thus a good benchmark for the proposed similarity model.

5.1. Basic-level categorization

The silhouette similarity algorithm was tested in two categorization tasks of different degree of difficulty. The first experiment involved a number of substantially dissimilar object classes; the second task required considerably finer shape distinctions.

5.1.1. Automobiles, aircraft and animals

The objects used in this experiment were 21 3D models of cars (four models), trucks (three models), airplanes (three models), helicopters (two models) and mammals (cats, dogs, cattle, and the artificial animals illustrated in Figure 5 – two or three examples for each subclass).

Three views per object were employed; in no view were the objects severely foreshortened; however, the orientation range was not restricted to good viewpoints. Two typical examples are shown in Figure 1, which depicts the three images used for one of the objects in the cat subclass and the three images for one of the objects in the car subclass. A total of 63 images were tested. Silhouette similarities were computed for all image pairs and entered in a 63 × 63 table. The similarity table was analyzed by multidimensional scaling* and the resulting two-dimensional configuration is presented in Figure 2.

In an MDS configuration distance expresses dissimilarity: the longer the distance separating two points, the higher their dissimilarity. The goodness-of-fit parameter ranges between 0 and 1 and measures the how well the distances measured in the plane of the MDS configuration fit the similarity data.

Two conclusions could be drawn from the examination of the MDS solution:

1. Views of objects from the same basic-level category are more similar to each other than to views of objects from different categories. In other words, object pose is inconsequential for basic-level categorization and does not affect relative similarity to a large extent.
2. A superordinate† organization of the basic-level classes is also visible: cars are more similar to trucks than to helicopters, and all artificial objects are more similar to each other than to mammals. Also, the artificial animals, though they cluster together with the "real" mammals, form a distinct group. These results are somewhat surprising in that superordinate categories have often been assumed to have no unique visual representation. Thus, our data indicate that objects in different basic-level categories but in the same superordinate class do bear a certain resemblance that can be extracted from image silhouettes and be used to separate them from objects from different superordinate categories.

*Technically, MDS procedures find a monotonic function that transforms similarities into numbers interpretable as distances in a space of low dimensionality by minimizing the discrepancy (stress) between the distances in the target space and the distances implicit in the data. The target space is usually two or three dimensional and the output of the MDS algorithm includes the coordinates of the stimuli in this space.

†Basic-level classes aggregate into more abstract, superordinate categories (for example, "animals" or "vehicles"), which, unlike the basic or the subordinate levels, do not seem to be composed of visually similar objects.

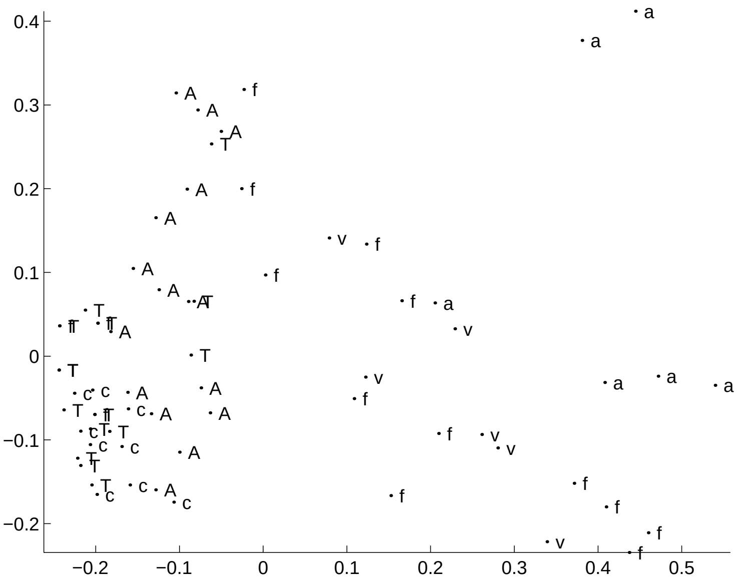

Figure 2. The 2D MDS solution for the basic-level categorization experiment. Each point represents a certain object view (silhouette). Three views per object were employed. Each category consisted of several objects. 'c' denotes cars and pickups, 'T' denotes trucks, 'A' denotes aircraft, 'f' denotes felines (cats), 'v' cattle and 'a' artificial animal models. Goodness-of-fit: 0.65

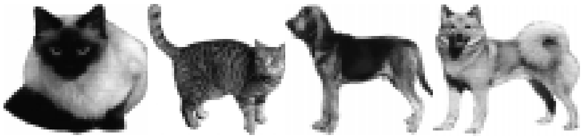

Figure 3. Typical examples of cat and dog images.

5.1.2. Cats and dogs

Although cats and dogs belong to different basic-level categories, they are quite similar in shape, especially if only the silhouette information is taken into account. Compounding to the difficulty is the variable geometry of the animals which substantially adds to image variability. Thus, discriminating cats from dogs based only on outline shape is not a trivial problem.

Interestingly, there is evidence that human infants as young as 3-months of age can successfully distinguish cats from dogs in color images. Quinn, Eimas, and Rosenkrantz[43] found that infants preferred viewing a new cat to a new dog after being shown a series of dogs or preferred viewing a new dog to a new cat after being shown a series of cats. They conclude that infants, who were unlikely to have much experience with either category, must be using perceptual features to perform this subtle basic-level classification.

We used the actual images of real cats (18 photos) and dogs (18 photos) of various breeds used by Ref. 43. The animals were photographed in different stances (sitting, standing, etc) and orientations with respect to the camera. Figure 3 displays some examples.

After the silhouettes were extracted from the images, all their pairwise similarities were computed and entered in a 36×36 table. The similarity table was analyzed by multidimensional scaling and the resulting configuration is shown in Figure 4.

Considering the computational difficulty of the task, the algorithm performed quite well (indeed the infants were not perfect either, and seem to show the same cat-dog asymmetry observed here). In the MDS solution the cats

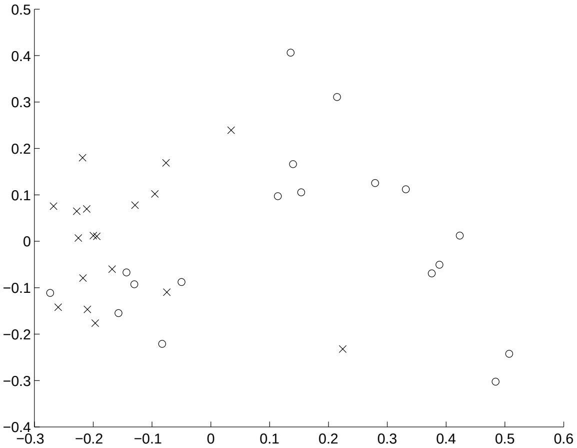

Figure 4. The 2D MDS solution for table of proximities among the cat and dog silhouettes. The dogs are denoted by 'o', the cats by 'x'. Goodness-of-fit: 0.74

tended to be better clustered than the dogs. The distances among the dog silhouettes had the larger variance, not surprising given that the different breeds of dogs were substantially more diverse than the breeds of cats. Importantly, the infants in Ref. 43 showed a similar sensitivity to the larger shape variability in the class of dogs – reducing this variability led to better classification performance by the infants. Finally, recent work by Quinn and Eimas (personal communication) indicates that infants seem to perform about as well in the cat-dog classification task when exposed only to silhouettes. This suggests that infants are making their judgments primarily on the information available in the silhouettes since they presumably have little knowledge of these categories in general and no knowledge of the specific images used in the study – thus, accessing shape memory through silhouettes is unlikely to have occurred.

5.2. Subordinate-level categorization

At the subordinate-level objects are highly similar and recognition is significantly dependent on viewpoint. Therefore, it seems unlikely that objects in the same subordinate class can be distinguished based on silhouette information alone.

We explored this issue in a series of computational experiments involving the synthetic animal shapes described in Ref. 44. The advantage in using these computer generated objects was that their 3D geometric similarity can be controlled parametrically. This was achieved by embedding the stimuli in a common objective *shape space* — a metric space spanned by a few dozen parameters, jointly controlling the appearance of each stimulus object.

The first stimulus set, the "triangle" configuration, consisted of seven objects, positioned at the vertices, barycenter and at the midpoints of the edges of an equilateral triangle in the parameter space, as shown on the left in Figure 5. The second set consisted of seven objects placed in a star-like pattern at the vertices, the barycenter, and at the midpoints of the segments joining the barycenter to the vertices of the triangle (shown on the right in Figure 5).

Four images (thus four silhouettes) were used for each object, resulting in $7 \times 4 = 28$ silhouettes for configuration triangle and $7 \times 4 = 28$ silhouettes for the star configuration. For each animal object, the four images displayed good views as well as and foreshortened views, as shown in Figure 6.

All pairwise silhouette distances were computed for each object set, and the similarity tables were analyzed by MDS. The MDS configuration is displayed in Figure 7.

For both the triangle and the star configuration the conclusion was the same: the similarity between images of different objects seen under corresponding viewpoints (say, profile views) is much higher than the similarity between different views of the same object. In other words, at the subordinate level the similarity between the profile views of objects A and B is substantially higher than the similarity between the profile and frontal views of object A. The same conclusion was reached in Ref. 44 by measuring image similarity in Gaussian receptive field space.

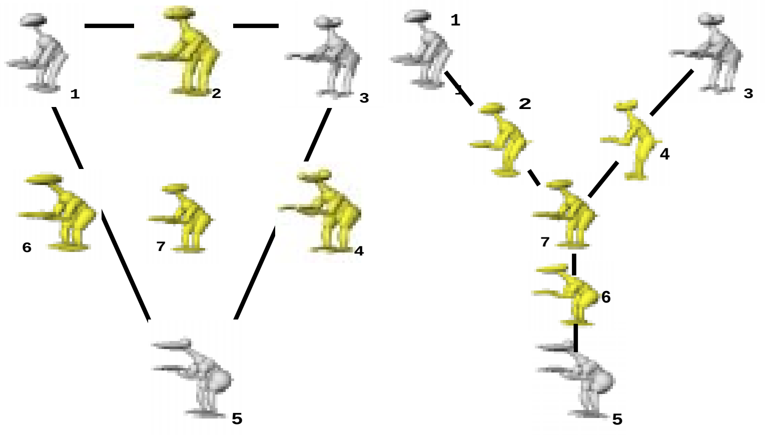

Figure 5. Animal shapes arranged in a triangle (left) and star (right) in shape parameter space.

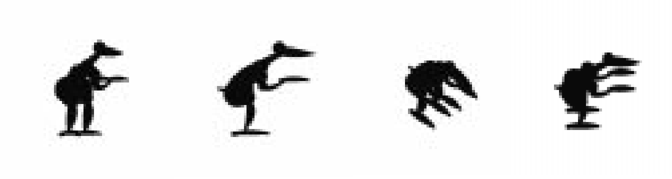

Figure 6. The four views of an animal model.

However, perceptual similarities measured for the same stimuli in Ref. 44 resulted in MDS configurations virtually identical to the corresponding shape space object configurations, indicating the visual proximities reflected shape space distances rather than view space distances.

We found that, in fact, the recovery of the shape space configuration (and thus of the psychological space configuration) is possible when using silhouette-based distances. The condition for recovery is to compute the silhouette distance measure among the *canonical views* of the objects in the set, and to exclude the silhouettes of the foreshortened views.[45,46] The silhouettes of the objects in the star and the triangle configurations were extracted from their canonical views (one silhouette per object). These views are illustrated in Figures 5 and 6 (the first silhouette from left).

The silhouette similarities were computed for all object pairs for both the star and triangle configurations and the resulting 7 × 7 distance tables were submitted to MDS analysis. As can be seen in Figure 8 the MDS solutions derived from silhouette distances are practically identical to the shape space configurations (Figure 5) and are very similar to the MDS solutions obtained from subject similarity data.[44]

If non-canonical, foreshortened views are chosen, the MDS solution derived from the silhouette distance table no longer bears resemblance to the star or triangle shape space configurations. The similarity among objects imaged from accidental viewpoints does not reflect the similarity computed by taking into account the 3D geometrical structure of the object. On the contrary, the similarity among objects imaged from canonical viewpoints closely approximates the similarity between their complete geometrical descriptions.

In 44 subjects judged perceptual similarities when shown the objects displayed in 5 revolving continuously on the screen, facilitating, via the kinetic depth effect the perception of the 3D geometry of the objects. The subjective similarities thus obtained were consistent with the geometrical, shape space distances among the stimuli. As demonstrated here, these data are consistent also with canonical view silhouette-based distances. It appears that silhouette distance can economically capture the similarity between the complete 3D geometrical descriptions of the respective objects if the proper views are chosen. Taking this argument in reverse, one may define the best view as the view whose silhouette allows the optimal recovery of the object's 3D similarity to other objects.

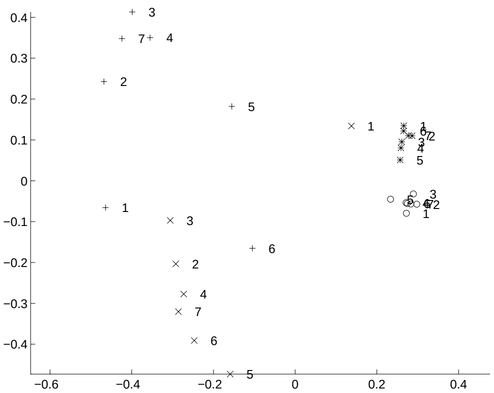

Figure 7. The 2D MDS solution for the seven objects arranged in a triangle in parameter space. Four views per object were employed. The numerical labels 1 to 7 indicate the object, and the four symbols 'x','+','*' and 'o' denote the four imaging viewpoints. The silhouettes clustered by object orientation, not by object identity.

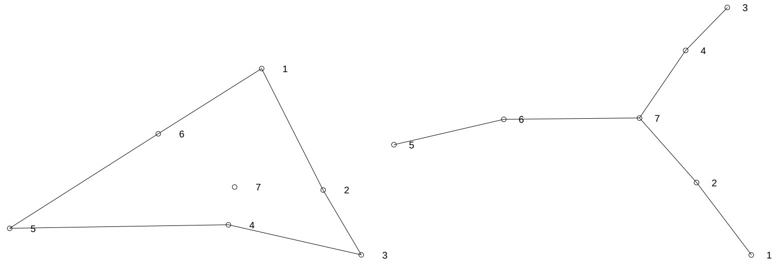

Figure 8. The 2D MDS solution for the table of similarities among the silhouettes extracted from the canonical views. Left: The seven silhouettes corresponding to the objects in the triangle configuration. Right: The seven silhouettes corresponding to the objects in the star configuration. The labels correspond to the object labels in Figure 5.

6. SIMILARITY OF THE VIEWS OF THE SAME OBJECT

We were also interested in the relationship between similarity and canonicalness. Specifically, the canonical view, as the object's most representative or typical, is maximally similar to the other views of the object (see Ref. 47,40 for a discussion of the relation between typicality and recognition). To investigate this issue, we reran the simulations described in the previous section on different views of a single object.

If the set of views under consideration is $\mathcal{V} = \{V_i\}$, $i = 1, \ldots, N$, the best view $V_c \in \mathcal{V}$ is the view for which the summed similarity (or typicality) $\tau_c = \sum_i S(V_c, V_i)$ (where S denotes a certain similarity measure, for example the silhouette-based distance), is maximal over the set $\mathcal{V}$. This interpretation of canonicalness is also compatible with the idea of goodness-of-view as the degree of viewpoint stability[48]: a stable view tends to be similar to the neighboring views on the viewing sphere, while an unstable view is special, very different from the others.

The viewing sphere of the test objects was sampled uniformly, by taking photographs from the vertices of a quasi-regular polyhedron with 42 vertices. After eliminating diametrically opposite viewpoints, a set of 21 distinct silhouettes remained. All pairwise silhouette distances were determined and the typicality was determined for each silhouette by applying the formula above. The results indicate that the good views are highly typical (high overall similarity), while the foreshortened views are atypical. Figure 9 displays typical (on the left) and atypical views (on the right) for two test objects. Note that these results are in agreement with psychophysical data we have recently collected on subjects' ratings of similarity for different views of airplane models.

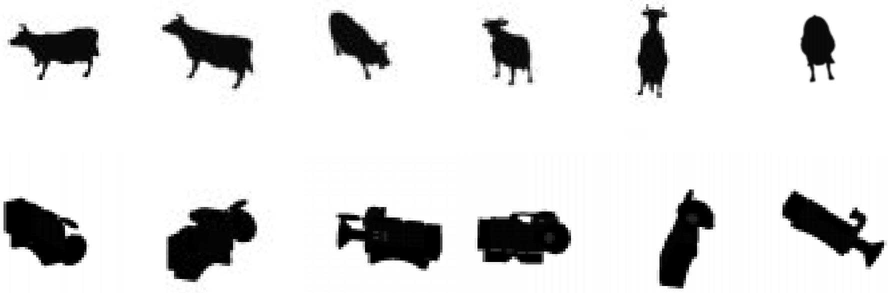

Figure 9. Typicality scores decrease from left to right. The most typical views are canonical views; the most atypical views are 'bad', unstable views.

7. CONCLUSIONS

Silhouette information can be used to do a lot of useful work in vision, from basic-level categorization to estimating view typicality. Fundamental to the computations involved in all these tasks is the measurement of similarity among silhouettes. The silhouette similarity measure presented in this paper, although simple, performed quite well in difficult categorization problems. In summary:

- One can tentatively define a principle for categorizing objects at the basic level: object identity (shape) is more important than object pose. In particular, we find that computations of silhouette similarity are sufficient for reasonably accurate basic-level classification. This is true even for more subtle discriminations, such as cats vs. dogs. These results are promising in that they are both more robust and more sensitive than other models of human basic-level classification.[2,49] Moreover, our results point the way to a possible organizing principle for representing objects in a high-dimensional shape space.
- We can also define a principle for subordinate-level shape contrasts: object pose is more important than object identity in the determination of perceptual similarity. Thus, variation in viewpoint is more likely to impact recognition at the subordinate level as compared to recognition at the basic level. This point has been substantiated by numerous psychophysical studies, where it is often the case that discriminating between visually-similar stimuli leads to larger viewpoint effects relative to discriminating between visually-dissimilar stimuli.[50–52]
- Although it is usually claimed that the different basic-level classes making up a superordinate class are not visually similar, our results, surprisingly, indicate that a degree of shape resemblance does exist.
- We have arrived at two different operational definitions of canonicalness: one defines the best view of an object in the context of other, similar objects from which it must be distinguished; the other defines the best view by considering the object in isolation. For the objects used in this study it appears that these definitions are consistent with each other and with the subjective notion of goodness of view.[45,46]

REFERENCES

1. P. Jolicoeur and G. K. Humphrey, "Perception of rotated two-dimensional and three-dimensional objects and visual shapes," in *Perceptual constancies*, V. Walsh and J. Kulikowski, eds., ch. 10, Cambridge University Press, Cambridge, UK, 1996. in press.
2. I. Biederman, "Recognition by components: a theory of human image understanding," *Psychol. Review* **94**, pp. 115–147, 1987.
3. I. Biederman and P. C. Gerhardstein, "Recognizing depth-rotated objects: Evidence and conditions for three-dimensional viewpoint invariance," *Journal of Experimental Psychology: Human Perception and Performance* **19**(6), pp. 1162–1182, 1993.

4. H. H. Bülthoff and S. Edelman, "Psychophysical support for a 2-D view interpolation theory of object recognition," *Proceedings of the National Academy of Science* **89**, pp. 60–64, 1992.
5. G. K. Humphrey and S. C. Khan, "Recognizing novel views of three-dimensional objects," *Canadian Journal of Psychology* **46**, pp. 170–190, 1992.
6. F. Cutzu and S. Edelman, "Canonical views in object representation and recognition," *Vision Research* **34**, pp. 3037–3056, 1994.
7. H. H. Bülthoff, S. Edelman, and M. J. Tarr, "How are three-dimensional objects represented in the brain?," *Cerebral Cortex* **5**, pp. 247–260, 1995.
8. M. J. Tarr, "Rotating objects to recognize them: A case study of the role of viewpoint dependency in the recognition of three-dimensional objects," *Psychonomic Bulletin and Review* **2**(1), pp. 55–82, 1995.
9. S. Edelman and D. Weinshall, "Computational approaches to shape constancy," in *Perceptual constancies: why things look as they do*, V. Walsh and J. Kulikowski, eds., Cambridge University Press, Cambridge, UK, 1996. in press.
10. S. Ullman, *High level vision*, MIT Press, Cambridge, MA, 1996.
11. D. Marr and H. K. Nishihara, "Representation and recognition of the spatial organization of three dimensional structure," *Proceedings of the Royal Society of London B* **200**, pp. 269–294, 1978.
12. T. Poggio and S. Edelman, "A network that learns to recognize three-dimensional objects," *Nature* **343**, pp. 263–266, 1990.
13. S. Ullman and R. Basri, "Recognition by linear combinations of models," *IEEE Transactions on Pattern Analysis and Machine Intelligence* **13**, pp. 992–1005, 1991.
14. S. Edelman, "Representation of similarity in 3D object discrimination," *Neural Computation* **7**, pp. 407–422, 1995.
15. R. Ellis, D. A. Allport, G. W. Humphreys, and J. Collis, "Varieties of object constancy," *Q. Journal Exp. Psychol.* **41A**, pp. 775–796, 1989.
16. E. Rosch, C. B. Mervis, W. D. Gray, D. M. Johnson, and P. Boyes-Braem, "Basic objects in natural categories," *Cognitive Psychology* **8**, pp. 382–439, 1976.
17. B. Tversky and K. Hemenway, "Objects, parts, and categories," *Journal of Experimental Psychology: General* **113**, pp. 169–193, 1984.
18. W. Torgerson, *Theory and Methods of Scaling*, John Wiley & Sons, Inc., New York, 1958.
19. R. N. Shepard, "Representation of structure in similarity data: Problems and prospects," *Psychometrika* **39**, pp. 373–421, 1974.
20. R. N. Shepard, "Multidimensional scaling, tree-fitting, and clustering," *Science* **210**, pp. 390–397, 1980.
21. R. M. Nosofsky, "Exemplar-based approach to relating categorization, idenitification, and recognition," in *Multidimensional models of perception and cognition*, F. G. Ashby, ed., pp. 363–394, Lawrence Erlbaum, Hillsdale, NJ, 1992.
22. R. M. Nosofsky, "Similarity scaling and cognitive process models," *arp* **43**, pp. 25–53, 1992.
23. F. G. Ashby, "Multidimensional models of categorization," in *Multidimensional models of perception and cognition*, F. G. Ashby, ed., pp. 449–484, Lawrence Erlbaum, Hillsdale, NJ, 1992.
24. N. A. Perrin, "Uniting identification, similarity and preference: general recognition theory," in *Multidimensional models of perception and cognition*, F. G. Ashby, ed., pp. 123–146, Lawrence Erlbaum, Hillsdale, NJ, 1992.
25. D. M. Green and J. A. Swets, *Signal detection theory and psychophysics*, Wiley, New York, 1966.
26. A. Tversky, "Features of similarity," *Psychological Review* **84**, pp. 327–352, 1977.
27. R. Beals, D. H. Krantz, and A. Tversky, "The foundations of multidimensional scaling," *Psychological Review* **75**, pp. 127–142, 1968.
28. R. M. Nosofsky, "Overall similarity and the identification of separable-dimension stimuli: a choice model analysis," *Perception and Psychophysics* **38**, pp. 415–432, 1985.
29. R. M. Nosofsky, "Tests of an exemplar model for relating perceptual classification and recognition memory," *Journal of Experimental Psychology: Human Perception and Performance* **17**, pp. 3–27, 1991.
30. W. T. Maddox and F. G. Ashby, "Comparing decision bound and exemplar models of categorization," *Perception and Psychophysics* **53**, pp. 49–70, 1993.

31. Y. Moses, S. Ullman, and S. Edelman, "Generalization to novel images in upright and inverted faces," *Perception* **25**, pp. 443–462, 1996.
32. W. G. Hayward, "Effects of outline shape in object recognition," *Journal of Experimental Psychology: Human Perception and Performance* **in press**, 1997.
33. W. G. Hayward and M. J. Tarr, "Testing conditions for viewpoint invariance in object recognition," *Journal of Experimental Psychology: Human Perception and Performance* **in press**, 1997.
34. C. T. Zahn and R. Z. Roskies, "Fourier descriptors for plane closed curves," *IEEE Transactions on Computers* **C-21**, pp. 269–81, March 1972.
35. E. Persoon and K. Fu, "Shape discrimination using fourier descriptors," *IEEE Transactions on Pattern Analysis and Machine Intelligence* **8**, pp. 388–397, May 1986.
36. K. Arbter, W. Snyder, H. Burkhardt, and G. Hirzinger, "Application of affine-invariant fourier descriptors to recognition of 3-d objects," *IEEE Transactions on Pattern Analysis and Machine Intelligence* **12**, pp. 640–647, July 1990.
37. G. H. Granlund, "Fourier processing for hand print character recognition," *IEEE Transactions on Computers* **C-21**, pp. 195–201, February 1972.
38. R. N. Shepard and G. W. Cermak, "Perceptual-cognitive explorations of a toroidal set of free-form stimuli," *Cognitive Psychology* **4**, pp. 351–377, 1973.
39. J. M. Cortese and B. P. Dyre, "Perceptual similarity of shapes generated from Fourier Descriptors," *Journal of Experimental Psychology: Human Perception and Performance* **22**, pp. 133–143, 1996.
40. M. A. Kurbat, E. E. Smith, and D. L. Medin, "Categorization, typicality, and shape similarity," in *Proceedings of the Sixteenth Anual Conference of the Cognitive Science Society*, Lawrence Erlbaum, (Hillsdale, NJ), 1994.
41. M. Werman and D. Weinshall, "Similarity and affine invariant distance between point sets," *IEEE Transactions on Pattern Analysis and Machine Intelligence* **PAMI-17**(8), pp. 810–814, 1995.
42. I. Borg and J. Lingoes, *Multidimensional Similarity Structure Analysis*, Springer, Berlin, 1987.
43. P. C. Quinn, P. D. Eimas, and S. L. Rosenkrantz, "Evidence for representations of perceptually similar natural categories by 3-month-old and 4-month-old infants," *Perception* **22**, pp. 463–475, 1993.
44. F. Cutzu and S. Edelman, "Faithful representation of similarities among three-dimensional shapes in human vision," *Proceedings of the National Academy of Science* **93**, pp. 12046–12050, 1996.
45. S. Palmer, E. Rosch, and P. Chase, "Canonical perspective and the perception of objects," in *Attention and Performance IX*, J. Long and A. Baddeley, eds., pp. 135–151, Lawrence Erlbaum, Hillsdale, NJ, 1981.
46. V. Blanz, M. J. Tarr, H. H. Bülthoff, and T. Vetter, "What object attributes determine canonical views?," Tech. Rep. 42, Max-Planck Institute for Biological Cybernetics, 1996.
47. R. M. Nosofsky, "Exemplar-based accounts of relations between classification, recognition, and typicality," *Journal of Experimental Psychology: Learning, Memory and Cognition* **14**, pp. 700–708, 1988.
48. D. Weinshall, M. Werman, and N. Tishby, "Stability and likelihood of views of three dimensional objects," in *Proc. 10th Israeli Symposium on Computer Vision and AI*, R. Basri, U. Schild, and Y. Stein, eds., pp. 445–454, 1993.
49. J. E. Hummel and I. Biederman, "Dynamic binding in a neural network for shape recognition," *Psychological Review* **99**(3), pp. 480–517, 1992.
50. M. J. Tarr and S. Pinker, "When does human object recognition use a viewer-centered reference frame?," *Psychological Science* **1**(42), pp. 253–256, 1990.
51. S. Edelman, "Class similarity and viewpoint invariance in the recognition of 3D objects," *Biological Cybernetics* **72**, pp. 207–220, 1995.
52. M. J. Tarr and H. H. Bülthoff, "Is human object recognition better described by geon-structural-descriptions or by multiple-views?," *Journal of Experimental Psychology: Human Perception and Performance* **21**(6), pp. 1494–1505, 1995.