Representation of similarity as a goal of early visual processing

Yair Weiss
Dept. of Brain and Cognitive Sciences and the Media Lab
MIT E15-384, Cambridge, MA 02139, USA
yweiss@media.mit.edu

Shimon Edelman
Dept. of Applied Math. and Comp. Science
The Weizmann Institute of Science
Rehovot 76100, Israel
edelman@wisdom.weizmann.ac.il

August 17, 1994

Abstract

We consider the representational capabilities of systems of receptive fields found in early mammalian vision, under the assumption that the successive stages of processing remap the retinal representation space in a manner that makes objectively similar stimuli (such as different views of the same 3D object) closer to each other, and dissimilar stimuli farther apart. We present theoretical analysis and computational experiments that compare the similarity between stimuli as they are represented at the successive levels of the processing hierarchy, from the retina to the nonlinear cortical units. Our results indicate that the representations at the higher levels of the hierarchy are indeed more useful for the classification of natural objects such as human faces.

1 Motivation

Systems of receptive fields (RFs) are probably the most prominent and ubiquitous computational mechanism employed in biological information processing, and, in particular, in vision. A natural question suggested by the hierarchy of RF types found in the visual pathway is, what is it good for? It may seem that the answer is to be found, jointly, in the many models of visual function based on population coding of various stimulus qualities, especially as some of these models draw explicit parallels between the representations they employ and the RFs found in biological vision. However, mere invocation of the idea of population coding, if not accompanied by a computational (in the sense of Marr, 1982) statement of what it is that the visual system does with its representations, simply begs the question:

> At all levels of the visual system, complex objects appear to be coded by the activity of populations, or networks, of cells, and the representation of a particular object may be widely distributed throughout one or more visual areas. That said, the goal of the anatomical pathway for object recognition becomes less obvious. The photoreceptors are a population of cells, for example, and they are necessarily capable of coding, by their population response, any conceivable stimulus. Why are subsequent populations needed? (Desimone and Ungerleider, 1989, p.268).

A number of recent works that do address the computational problem of representation tend to employ information-theoretic terms such as redundancy reduction and efficient coding (Field, 1994; Daugman, 1988; Atick, 1992). In this paper we suggest an alternative approach, based on the observation that object classification (which may be considered an ultimate goal of vision) requires faithful representation of true similarity between shapes (Edelman, 1993). This observation leads to the hypothesis that successive stages of early visual processing remap the retinal space in a manner that makes objectively similar shapes closer to each other.

The paper is organized as follows. In section 2, we consider a general formulation of the issue of similarity under different representations. Section 3 contains an experimental evaluation of several similarity measures on a database of face images, and, in particular, a comparison of the similarity induced by realistic RFs with that of two control cases. Section 4 contains a discussion of the results and lists directions for future research.

2 The effect of the choice of representation on similarity between images

Recent theories of object recognition based on view interpolation (Poggio and Edelman, 1990) or on linear combination of views (Ullman and Basri, 1991) have underscored the possibility of representing 3D shapes by collections of their views, or images. However, in models of biological vision, the notion of an *image* is ill-defined at any stage past the projection of the world onto the photoreceptor sheet in retina. At all the subsequent levels, the visual system has at its disposal only the activities evoked by this input image in the units of the preceding level, and it can compare two images only by comparing these population activity vectors. Thus, a representation scheme together with a metric in the representation space naturally induce a measure of similarity among (input) images. For a recognition scheme such as view interpolation to succeed, this induced or proximal similarity between image representations must correspond in a principled manner to the objective or distal similarity between objects that give rise to the images. Consequently, we propose to compare various representation schemes according to the similarity measures they induce.

2.1 Formulation

As customary in models of early vision, we treat the early representations as linear systems with an optional pointwise nonlinearity applied to the output of the system.[1] Following statistical convention, we refer to systems with a pointwise nonlinearity as generalized linear systems. To visualize the effect of representation on similarity, it is useful to think of images as vectors in a high-dimensional space X, with the number of dimensions equal to the number of pixels (Intrator et al., 1992; Intrator and Gold, 1993). The representation process is then considered as matrix multiplication, with the representation of an image x calculated by an application of a linear operator L, as Lx. For example, consider a scheme wherein images are represented by the activities of n linear units. The activity of the i'th linear unit is given by:

$$o_i(x) = < x, RF^i > = \sum_{k,m} x_{km} RF^i_{km} \tag{1}$$

In this case, the matrix L is a rectangular matrix with D columns and n rows. The i'th row contains the vector representation of the i'th receptive field (RF^i). The similarity between two images is found by multiplying each of them by the matrix L and using the metric in R^n to determine the distance between the two resulting vectors.[2] This measure should correspond as closely as possible to the objective similarity between the objects that gave rise to the two images.

2.2 Linear representations

How can the choice of representation affect the similarity between two images as it is perceived by the system? The first case to consider is obviously that of a singular matrix L. In this case, two distinct images x_1 and x_2 may be mapped to (represented by) identical vectors, introducing a metamery. The representation would then be trivially invariant to the transformation from x_1 to x_2 (cf. Ratliff and Sirovich, 1978). The study of invariances has a long history in psychology (Gibson, 1979) and in computational vision (Mundy and Zisserman, 1992), but it is of limited interest in the present context. This is because, on the one hand, it is difficult to make the metameries serve a useful purpose, and, on the other hand, they can be removed simply by making the number of linearly independent RFs equal to the dimensionality of the input space. The matrix L is then invertible and there are no strict "invariances."

A more interesting case arises when the matrix L, while being nonsingular, changes the *relative* similarity of pairs of images in a manner that causes, for example, the representation of an image A, originally more similar to C than to B, to be more similar to the representation of B. This effect is best understood by visualizing the equi-similarity contours in the original high-dimensional space X (see Figure 1). The representation can then be characterized by

[1] This analysis neglects the essentially nonlinear process of light adaptation in the retina but is valid for images in which the mean illumination level is roughly constant.

[2] We assume the Euclidean metric in R^n; some of the simulations will also consider the city-block metric.

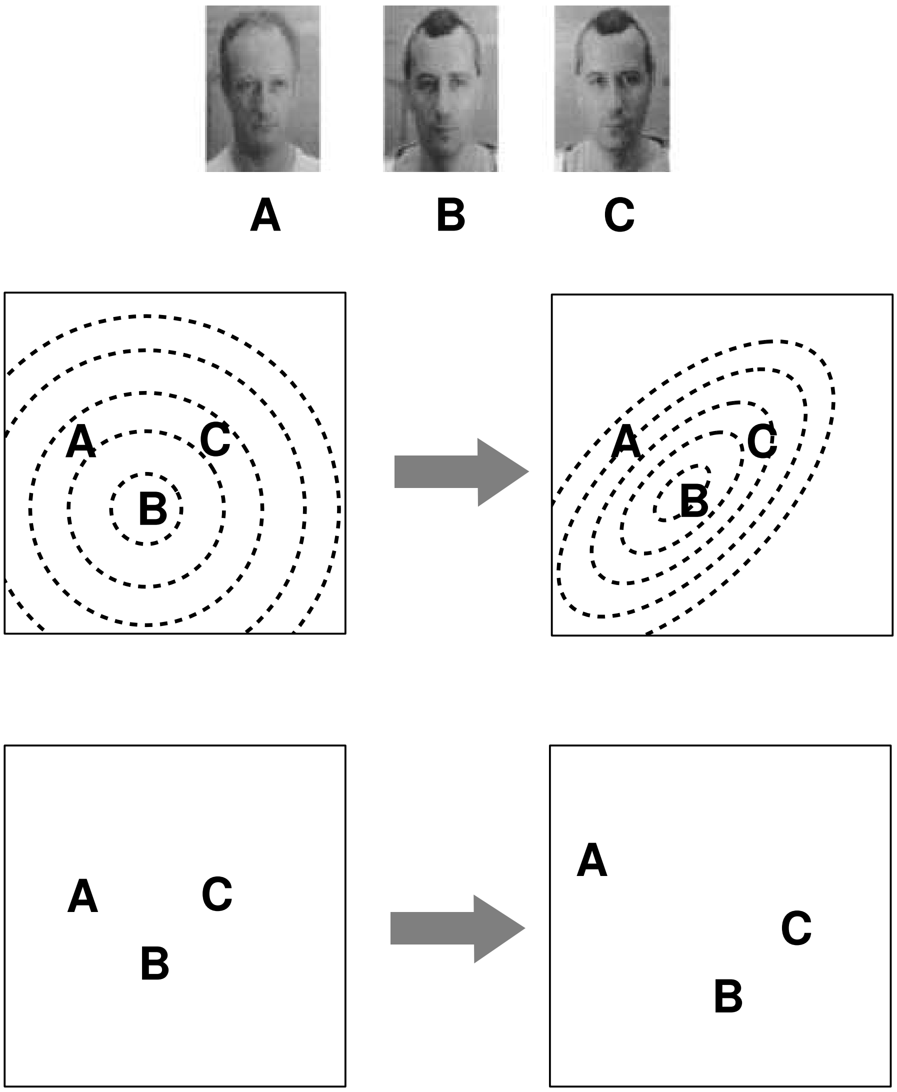


Figure 1: In this work, images are considered as points in a high-dimensional space (of which only two dimensions are shown in this illustration). We explore the similarity measure induced on this space by different representations (see section 2). Consider the three face images in the top row: a simple (e.g., pixel-based) similarity measure would yield the same similarity between B and C as between B and A, whereas we would like B to be closer to C, which is another image of the same object, than to A (middle left). This can be achieved by transforming the pixel space, in which the equi-similarity contours around a reference image are hyperspheres, into a space of receptive field activities, in which the similarity measure would improve the clustering of different views of the same 3D object (middle right). The transformation of similarities can be visualized by a change of coordinates making the similarity between representations monotonically related to the distance between the corresponding points in the plot (bottom row).

specifying the class of all images whose distance to image A is smaller than some constant ϵ. It is easy to show that the equi-similarity contours are hyperellipsoids centered on A, whose axes of symmetry are the eigenvectors of the matrix L^*L. The elongation of the ellipsoid in the direction of the eigenvector v_i is $\epsilon/\sqrt{\lambda_i}$, where λ_i is the corresponding eigenvalue with respect to the operator L^*L (see appendix A).

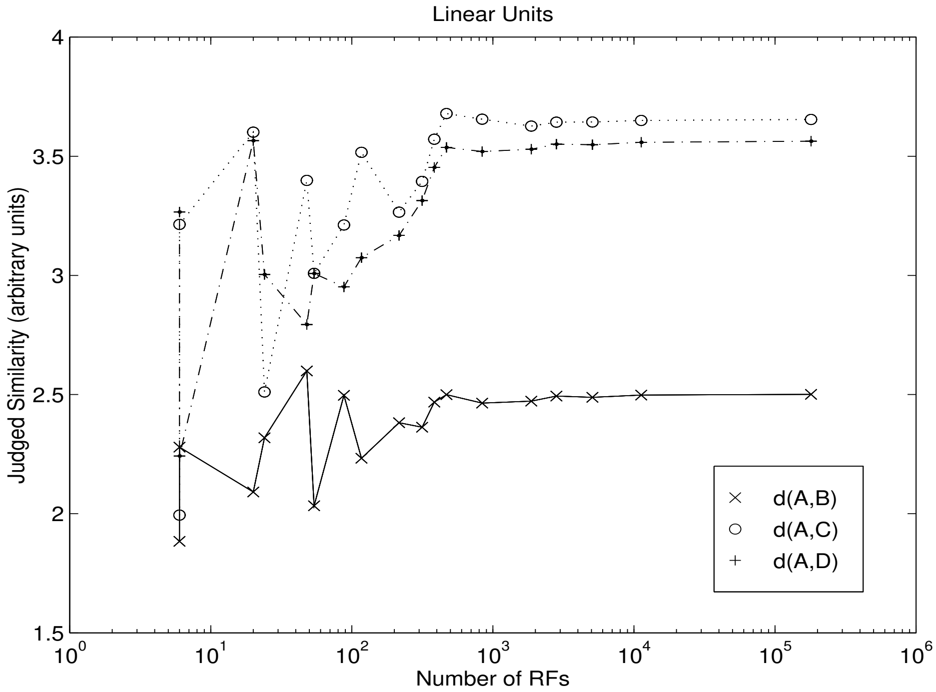
This simple observation has two significant implications. First, linear representations that induce no invariances can still have a dramatic influence on the relative similarity of images. Second, the relative similarity of images cannot be predicted by considering the individual RFs, but rather depends on the *ensemble* of RFs through the operator L^*L. Let us compare, for example, a representation in which the RFs are individual pixels to one in which the RFs form an orthonormal basis of sine and cosine gratings. While in the first case the RFs are localized in space, and in the second one they are spread out in space and localized in frequency, both representations induce the same similarity measure, because the operator L^*L in both cases is the identity matrix.

The preceding discussion suggests that to characterize the similarity measure induced by a biological representation one would have to determine the details of all the receptive fields, and then compute the eigenvectors of a huge matrix. However, if the system is shift-invariant (i.e., if all the RFs are shifted copies of a given profile, RF^0), the eigenvectors of L^*L can be found analytically: they are the sine and cosine gratings with eigenvalues given by the Fourier transform of RF^0.

The analytic solution for the eigenvectors of L^*L still holds when there are multiple RFs at each location of the sampling grid (pixel), or, more importantly, when only a random subset of the pixels is contained in the sample. The latter is a result of the saturation property of RF-based representations in which the individual RFs come from a parametric family. The saturation property for a family of receptive fields $RF(\theta; x, y)$ parameterized by θ can be stated as follows. Consider a sequence of representations $\{L_n\}$, such that the receptive fields of the n'th representation are obtained by taking n samples from the parametric family $RF(\theta; x, y)$. Then the similarity measure $d_n/\sqrt{n}$ induced by L_n converges to a limit d_∞ which is the similarity induced by including all the receptive fields in the family.[3] In fact, $d_n/\sqrt{n} \rightarrow d_\infty$ can be seen as a numerical integral, and standard error bounds for numerical integration can be used to estimate its convergence properties. In general, the rate of convergence is higher when the RF properties vary smoothly as a function of the parameter θ (Davis and Rabinowitz, 1975).

For example, if the receptive field family is $RF(\theta_x, \theta_y; x, y) = g(x - \theta_x, y - \theta_y)$ and the Fourier transform G of g is band-limited by G_{max}, then $d_n = d_\infty$ for all placings of n receptive fields on a uniform rectangular grid in which the distance between centers of the receptive fields is smaller than $1/2Gmax$. Likewise, if the family of RFs includes also all rotations of g, not all of these need to be used to determine the similarity: the representation will saturate when the number of orientations is larger than $2A_{max}$ where A_{max} is the angular band limit of G

[3]The division by $\sqrt{n}$ makes no difference, of course, in the relative similarity of images.



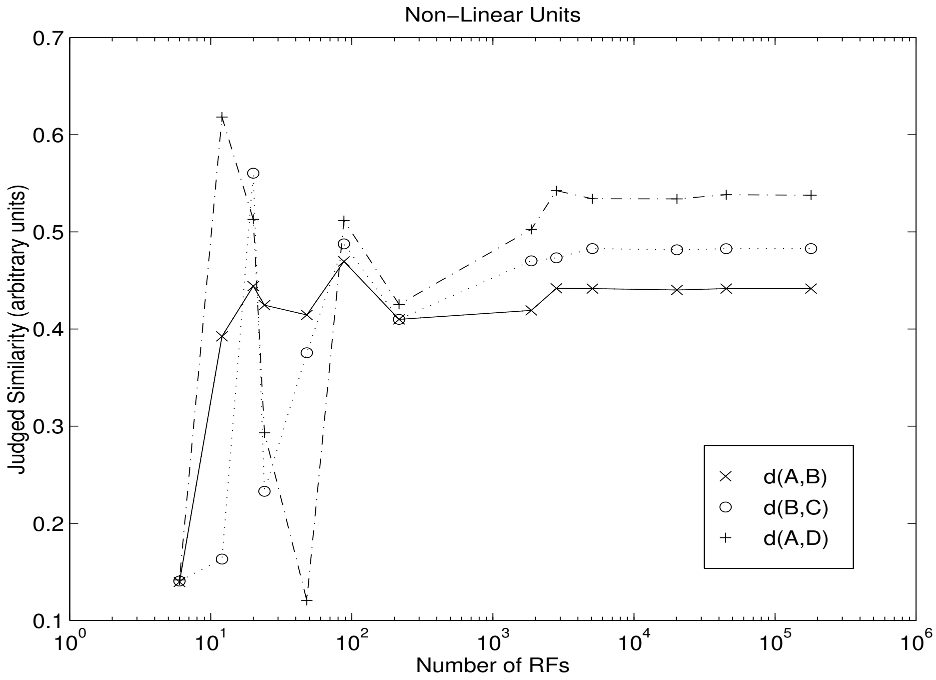


Figure 2: When the representation is created by taking n samples of a parametric family of receptive fields (e.g., shifted versions of a basic receptive field), then the similarity measure tends to saturate with increasing n. This is true both for linear units (top) and generalized linear units (bottom). Each graph shows the induced similarity of one image to three others (plotted by separate curves), as a function of the number of receptive fields. Note the logarithmic scale on the abscissa.

(Freeman and Adelson, 1991). Note that the critical sampling rates in orientation and in space are independent of the properties of the images to be compared.[4] For example, if the RFs are rotated derivatives of Gaussians, then the number of RFs needed for the representation to saturate is inversely proportional to the standard deviation of the Gaussian, and only three orientations are needed: $A_{max} = 1$ (Freeman and Adelson, 1991).

A second example is when the RFs in d_n are randomly placed over the image. If the random placement is uniform (i.e. the probability of placing an RF at each location is the same), then the representation still converges to d_∞. If the random placement is biased, then d_n converges but to a different limit.

2.3 Nonlinear representations

If linear representations can have such a strong effect on the similarity, why are nonlinear units needed? One reason is that without nonlinearity there is no advantage to divergent representations (in which the number of units is larger than the dimensionality of the input). It can be shown that for any divergent linear representation there exists a non-divergent representation which induces the same similarity measure. Specifically, if all units are linear, then a representation based on multiple oriented units, is equivalent to a representation with circularly symmetric units. In other words, without nonlinearity there is no advantage to using oriented units. Likewise, a linear representation with multiple scales is equivalent to a single-scale representation.

Some of the tools used to analyze linear representations are no longer useful for generalized linear representations. First, the equi-similarity contours are no longer ellipsoids: their shape depends on the "reference" image on which they are centered, as well as on the similarity threshold ϵ. Thus, even shift-invariant nonlinear representations cannot be analyzed simply by looking at the Fourier Transform of the RF. Second, the dissimilarity of two images can not be simply calculated by calculating the norm of the representation operator applied to the the difference image. In the generalized linear case, the dissimilarity can be thought of as a weighted norm on the represenation of the difference image, where the weights depend on the two images being compared.

The saturation property, on the other hand, still holds for nonlinear representations, and a bound on the number of RFs needed to obtain saturation can be estimated independently of the two images being compared, via the Taylor expansion of the nonlinearity.

[4]The value $1/2G_{max}$ is, of course, the Nyquist sampling rate of the convolved image. However, unlike in the classical sampling theorems, we do not assume here that the system explicitly reconstructs the convolved image. In fact, the representations saturate even when the receptive fields are randomly placed over the image, in which case reconstruction is impossible.

3 Computational experiments

In the previous section, we showed how to compute the similarity measure induced by a given representation. However, this computation does not tell us how useful the induced similarity measures are. To determine that, we need a definition of the goodness of a similarity measure. As suggested in the introduction, we base this definition on the utility of the representation for recognizing 3D objects. Specifically, we require that a view of a 3D object be more similar to other views of the same object (obtained under different viewing positions and illumination conditions) than to views of another 3D object. We proceed now to examine how well do the similarity measures induced by early visual representations satisfy the object-based criterion.

The main experiment described below (section 3.2) involved images of natural 3D objects (human faces), processed by RFs resembling those found in the mammalian visual pathway. Three other experiments have also been conducted. The first of these (section 3.3) considered the effects of changing the representation (using different RF profiles) while holding the 3D objects fixed. The second control experiment (section 3.4) involved the same RFs as in the main experiment, applied to a different set of 3D objects (randomized fractal surfaces). The last experiment (section 3.5) addressed some of the questions regarding the properties of face images, raised by the preceding experiments.

3.1 Methodology

We assessed the utility of a representation by observing the clustering of the points belonging to the different input classes in a 2D space obtained by multidimensional scaling, and by testing the generalization performance of a radial basis function classifier fed with points from that space. Because it is impossible to visualize directly the locations of image representations and the associated equi-similarity contours in a high-dimensional space, the dimensionality had to be reduced, preferably by a method that preserves as much as possible the relative distances between the different points. We found multidimensional scaling (MDS)[5] suitable for this purpose. The radial basis function (RBF) network we used to estimate the classification performance supported by the different representations is described in appendix B.

3.2 Experiment 1: remapping of similarity in the visual pathway

Purpose. This experiment tested the basic assumption of the present work, stated in section 1, namely, that at the successive stages of the visual pathway the representation of images is modified in such a manner that views of the same 3D object tend to cluster more and more tightly. This experiment involved images of human faces, and used RFs resembling those found in the early stages of mammalian vision.

[5]MDS is a method that, given a table of inter-point distances, finds an embedding of the points in a space of prescribed dimensionality which yields the smallest discrepancy between the true distance table, and the table obtained by computing the distances between the points in the resulting space (Kruskal and Wish, 1978).

Figure 3: One class of face images used in the simulations. Images included changes in illumination, viewing position, and facial expression, as well as image-plane translations and white noise.

Figure 4: All four individuals whose images were used in the simulations. Points belonging to the four images classes are marked in Figures 5 and 6 by the symbols shown near each image.

Images. The data set included $512 \times 352 \times 8$ images of faces of four individuals. The different images of each face varied in viewing position, illumination direction, and facial expression. The images also included pictures that were corrupted by white noise, as well as pictures translated by 6 pixels in the image plane simulating small errors in the normalization procedure compensating for simple geometric transformations of the input, as employed in (Edelman et al., 1992). The various images of one face in the data set are shown in Figures 3 and 4. This image set is a subset of the face image database described in (Moses and Ullman, 1992).

Receptive fields. The simulated visual pathway in this experiment was composed of four stages: (1) pixel RFs simulating the photoreceptor-level representation; (2) circularly symmetric center-surround RFs simulating the lateral geniculate body (LGN) representation; (3) oriented Difference of Gaussians or DOG RFs simulating the primary visual area (V1), and (4) DOG RFs coupled with a sigmoidal nonlinearity of the form $f(t) = \tanh(at - b)$, with fixed a, b. Units

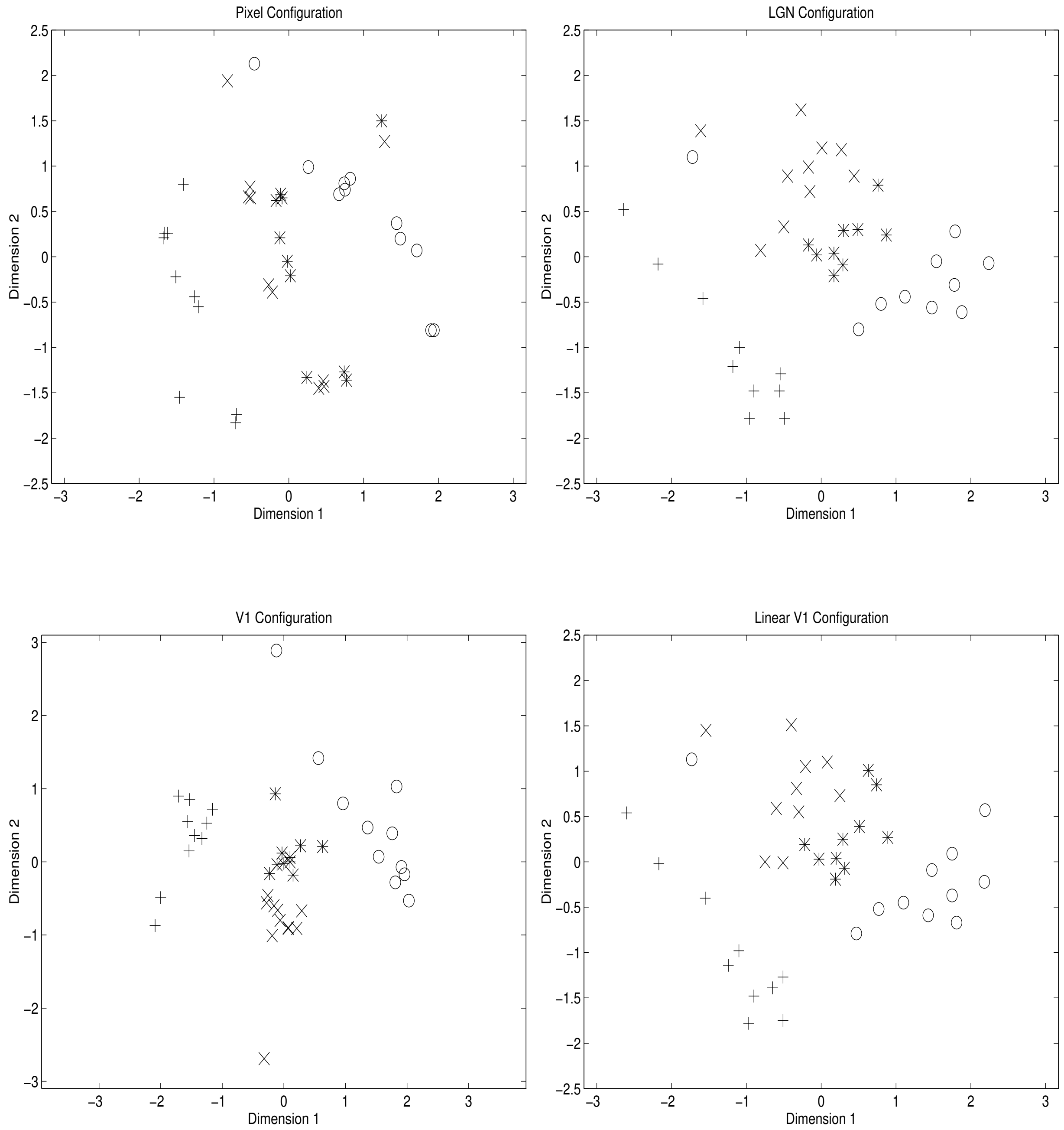


Figure 5: The relative similarity among the face images induced by models of (clockwise from top left) retina, LGN, linear V1, nonlinear V1, using the Euclidean metric. Images of the same face are marked by the same symbol (see Figure 4). The clustering at the retina is clearly inferior to the clustering in the LGN and V1. Note the near equivalence of the LGN and linear V1, as predicted by the theory. The outlier point in V1 corresponds to an image taken with a relatively large deviation in viewing position (17°).

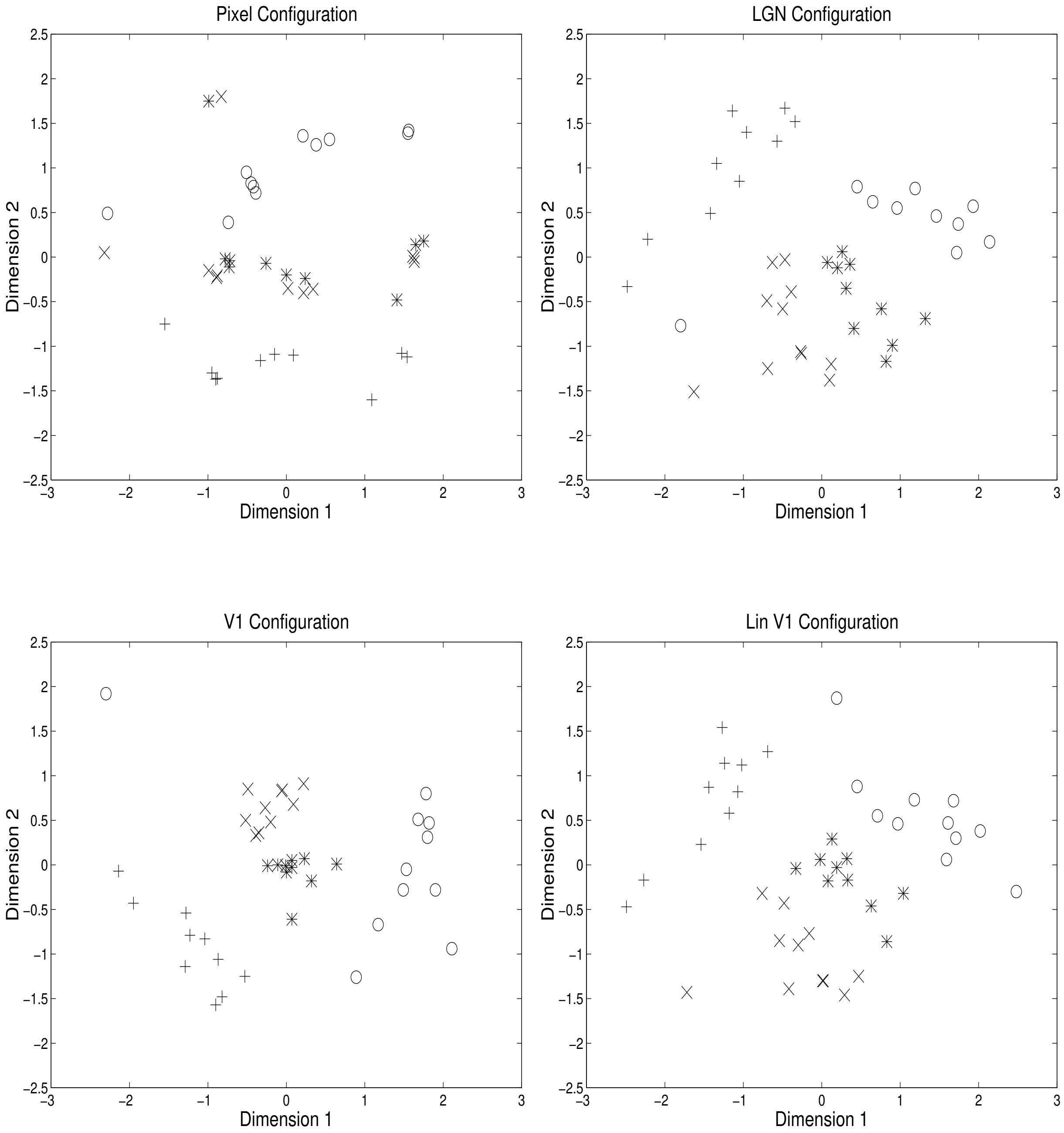


Figure 6: The relative similarity among the face images induced by models of (clockwise from top left) retina, LGN, linear V1, nonlinear V1, using the city-block metric. Images of the same face are marked by the same symbol (see Figure 4). Results are similar to those obtained under the Euclidean metric.

were centered on a rectangular grid with a spacing of 6 pixels; at each such center there were four units, oriented at 0, 45, 90 and 135 degrees, respectively.

Results. The results of the MDS analysis for the Euclidean and the city-block metrics appear in Figures 5 and 6. The results for both metrics are similar. In particular, the clustering in the retinal or pixel representation space is clearly inferior to the clustering in the LGN and V1. Note the near equivalence of the LGN and linear V1, as predicted by the theory (section 2.2). The clusters in the nonlinear V1 representation are slightly better than in the linear model. The outliers in the V1 clusters correspond to images of faces taken from a viewpoint that differed by a relatively large amount (17°) from the viewpoints used in making the other images.

Results of the RBF classification are shown in Figure 7. With only one example, the performance of the retinal representation was clearly inferior to that of the other representations. The performance of the LGN representation is nearly identical to that of the linear cortex model. The nonlinear V1 model shows a small but significant additional improvement: the number of classification trials (summed over all iterations) was $105,000$, and a difference of 4% in the classification rate means an additional $4,200$ correct classifications. With two examples the retinal representation performance improved, but was still inferior to the other ones (the number of classification trials here was $42,000$.)

Discussion. We found the high performance of the RBF network trained on a single example of each class somewhat surprising. An examination of the misclassified images showed that the model often failed to generalize over the 17° shift in viewing position, but generalized successfully for smaller changes. The performance clearly improved with an addition of another training example. Note that even the retinal representation performed relatively well in this condition. This finding is suggestive in light of the question raised by Adini et al. (1993) regarding the number of images that need to be stored to achieve high performance in a pixel-based multiple-view face recognition scheme.

3.3 Experiment 2: the importance of RF shape

Purpose. In principle, it is possible that in experiment 1 most of the advantage of the V1-like representation over the raw pixels was due to factors other than the joint action of the center-surround RFs (e.g., it could arise simply from the shift from pixels to spatially extended RFs, or to RFs which show no response to a constant field). Experiment 2 thus aimed at determining the contribution of the particular form of the RFs found in the visual pathway (namely, the center-surround DOGs and oriented RFs) to the utility of the resulting representation.

Images. The same image set as in the previous experiment was used here.

Receptive fields. We compared the performance of four different kinds of RFs:

- Pixel RFs, applied to images normalized to have zero mean luminance (this is equivalent to RFs with an inhibitory lobe equal in size to the entire image). These were included to rule out the claim that any zero-mean RF profile would do.

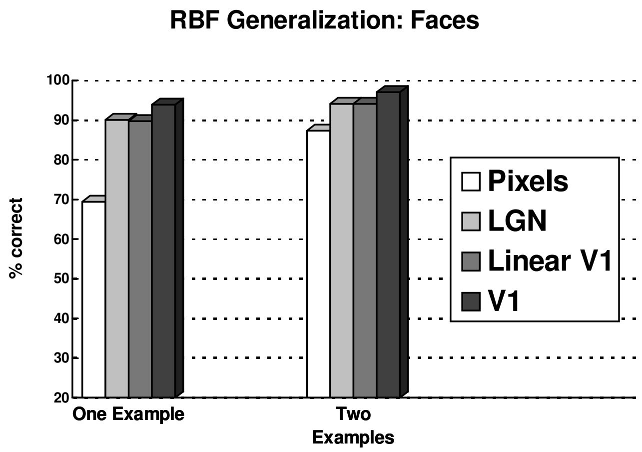


Figure 7: Performance of the RBF classifier for the different representations. With only one example (left), the performance with the retinal representation was clearly inferior to that of the other representations. Performance of the LGN representation is nearly identical to that of the linear cortex model. The nonlinear V1 model shows a small but significant additional improvement. With two examples (right) the retinal representation performance improves, but is still inferior.

- Random RFs, with zero-mean white noise profile, and size roughly equal to that of the LGN RFs in experiment 1. These were included to demonstrate that not any spatially extended RFs would do.
- RFs set to the principal components of the face ensemble. Principal components of face images, or eigenfaces, proved to be useful in computer vision applications of face recognition (Turk and Pentland, 1991).
- Eigenface RFs, excluding the first three principal components. The exclusion of the first several eigenfaces has been found to improve recognition in some cases (Pentland et al., 1993; O'Toole et al., 1993).

Results. The MDS-derived configurations from experiment 2 are shown in Figure 8. One may observe that luminance normalization did not improve clustering.[6] The white-noise RFs actually degraded the clustering, while RFs based on the first 11 principal components yielded a configuration that was almost identical to the pixel configuration up to a rigid transformation (under which the MDS procedure is invariant). Finally, RFs based on principal components number 4 through 14 yielded a three-cluster structure, in which clusters corresponding to two of the original face categories coalesced.

The RBF classification results (shown in Figure 9) are as expected from the MDS configurations. None of the representations improved performance compared to the original pixel representation, while the use of random RFs resulted in worse performance.

Discussion. The results of experiment 2 indicate that the particular shape of the RF profile found in the visual pathway may be important for obtaining the required pattern of similarities among object views: a number of alternative RF types did not perform nearly as well as the "natural" ones. In the next experiment, we tested whether the natural RFs would retain their advantage when applied to images of unnatural objects.

3.4 Experiment 3: natural vs. not so natural objects

Purpose. A number of recent works succeeded to derive some of the properties of the RFs found in the visual pathway from the statistics of natural scenes (Field, 1987; Atick and Redlich, 1992; Hancock et al., 1992). As the present study is concerned with the issue of faithful representation of similarities between objects, a related question suggested itself here: would the advantage of the V1-type center-surround RFs over pixels hold for images of shapes that only resemble natural objects in some of their statistical properties.

[6] It should be noted that the apparent linear separability in the 2D plot, also present in the pixel 2D plot, does not necessarily indicate linear separability in the high-dimensional representation space before the action of the MDS routine.

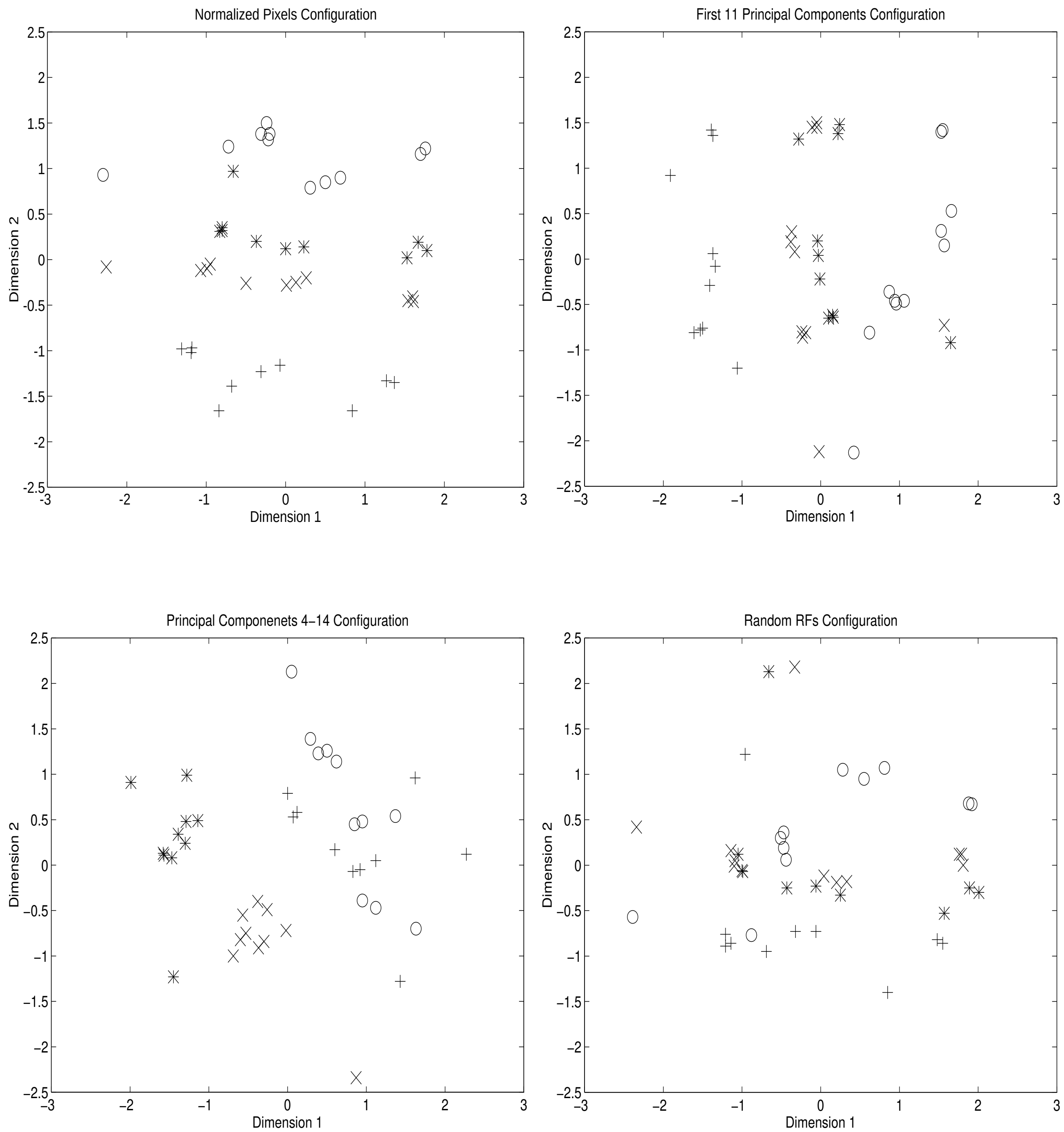


Figure 8: The relative similarity among the face images induced by the four control representations described in section 3.3. None of the representations improve the clustering while the random-profile RFs actually impair the performance.

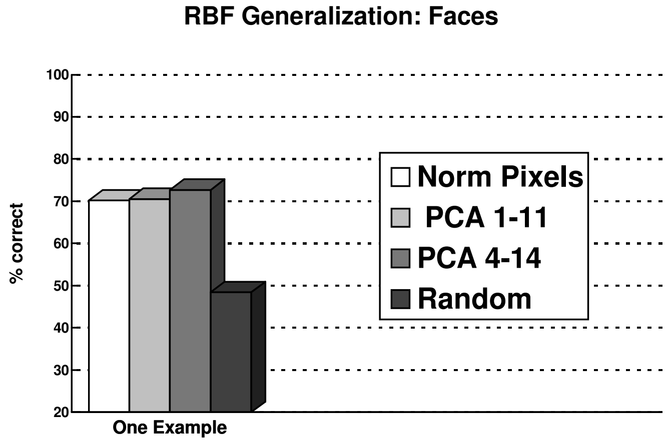


Figure 9: Performance of the RBF classifier on the four different types of representations in experiment 2. As expected from the MDS configurations, the RF sets derived from the principal components did not perform better than the raw pixels, while random-profile RFs caused the performance to deteriorate.

Images. The image set consisted of four surfaces imaged under different viewing conditions: change in illumination direction and slight changes in viewing position. The surfaces were random fractal surfaces which were constructed so as to have the same fractal dimension as the surfaces of the faces in the original image set (see appendix C for details). These surfaces have roughly the same second order statistics as the faces, but their higher order statistics are different. The surfaces are shown in Figures 10 and 11.

Receptive fields. The same RF profiles and nonlinearity parameters were used here as in experiment 1.

Results. RBF classification was performed using the same protocol as before. The results appear in Figure 12. None of the RF sets that performed so well in experiment 1 generalized successfully across changes in illumination given the fractal surfaces tested here.

Discussion. The results of experiment 3 show that, insofar as the RFs in the visual pathway are adapted to natural scenes, the fractal images in Figure 11 are indeed unnatural. This finding is compatible with the discussion in (Field, 1994), where it is claimed that to look natural, images must exhibit certain properties of phase correlation across spatial scales, in addition to the amplitude correlations implied by their fractal nature. In this connection, we note that within the similarity representation framework proposed in the present paper, the

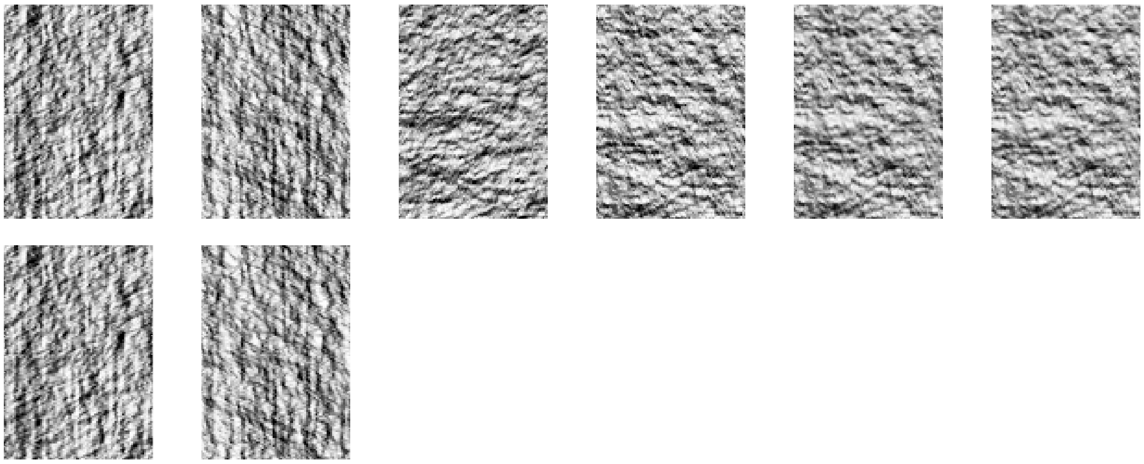


Figure 10: One class of fractal images used in the simulations. Images included changes in illumination and in viewing position. The surfaces were constructed to have the same fractal dimension as the faces in Figure 4.

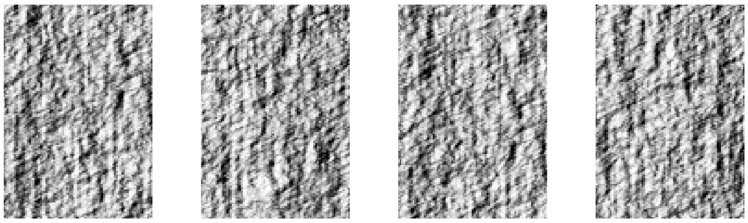


Figure 11: All four fractal surfaces whose images were used in the simulations.

properties of *image differences* are more important than the properties of individual images. Consequently, in the next section, we attempt to gain an understanding of the advantage of natural images by spatial frequency analysis of image differences.

3.5 Experiment 4: image differences and their sources

According to the similarity representation hypothesis, the successive representations in the visual pathway reduce the differences among images of the same object, compared to differences among images that belong to distinct objects. To appreciate the potential of this approach, one must consider the image difference information available to the system initially, i.e., in the pixel space. In this section, we compare two kinds of image differences: those obtained by subtracting images of the same object under different illuminants, with those obtained by subtracting images of different objects under the same illuminant. If the two kinds of image differences exhibit distinct properties, subsequent processing could capitalize on the distinction to attain better clustering of images of the same object in the representation space.

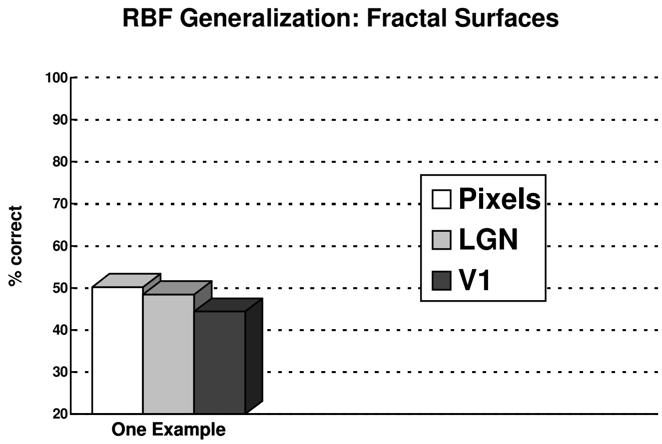


Figure 12: Performance of the RBF classifier on the fractal surfaces. Unlike in the case of face images, all RF representations here failed to generalize across changes in illumination.

$D1$ $D2$

Figure 13: Two image differences: $D1$, on the left, is a difference between two images of the same face under different illuminant directions. $D2$, on the right is the difference of two images of two different faces under the same illumination conditions.

3.5.1 Spatial frequency analysis of image differences: face images

Consider Figure 13, which shows differences between two pairs of face images. On the left, $D1$ is the difference between two images of the same face under different illuminant directions. On the right, $D2$ is the difference between images of two different faces under the same illumination. Note that $D1$, which is a difference of two face images, looks like an image of a face. Indeed, it has been shown that in the absence of self-shadows, the difference between two images of a Lambertian surface under different illuminant directions yields an image of the same surface under a third illuminant direction (Shashua, 1992; Moses, 1993). A simple consequence of this result is that no generalized linear representation can be strictly invariant to changes in illuminant direction. Still, representations that are not invariant but show reduced sensitivity to illuminant direction are possible. This examplifies the distinction made in section 2 between strict invariance and change in relative similarity.

If we restrict ourselves to linear shift-invariant representations, then the characterization of the two kinds of image differences is best done in the Fourier domain. Figure 14 shows the power spectra of the two differences. Note that the spectrum of $D1$ decays faster than the spectrum of $D2$. Thus, a representation that is more sensitive to high frequency differences than to low frequency ones, would emphasize the difference encoded by $D1$ relative to the difference encoded by $D2$. Indeed, the difference (dissimilarity) patterns found in the LGN and linear $V1$ representations in experiment 1 do precisely that. These representations are most sensitive to image differences within the spatial frequency band defined by the support of the Fourier transform of the receptive field. In this band, $D1$ has 11.1% of its energy, while $D2$ has there 27.8% of its energy. Consequently, subjecting these two differences to band-pass filtering contracts the within-category distance ($D1$) three times more than than the across-category distance ($D2$).

The reduced high-frequency content of difference images arising from illumination changes is not surprising, because these are simply new images of the same face, and, as such, are expected to be relatively smooth. One way to formalize this intuition is by fitting a function of the form f^{α} to the power spectrum of the difference images. Kube and Pentland (1988) have shown that the exponent α, which is related to the fractal dimension of the surface, is an invariant of the imaging process. Thus, $D1$ would have the same exponent as the depth map of the face, whereas $D2$ would have the same exponent as the difference between depth maps of two faces, and hence would be expected to have more energy at the high frequencies, as seen in Figure 14.

3.5.2 Spatial frequency analysis of image differences: non-face images

The preceding argument assumed implicitly that the face surfaces that give rise to the images are relatively smooth. The failure of this assumption in the case of the random, fractal surfaces is illustrated by considering the image differences shown in Figure 15. As in Figure 13, $D1$, on the left, is a difference between two images of the same fractal surface under different illuminant

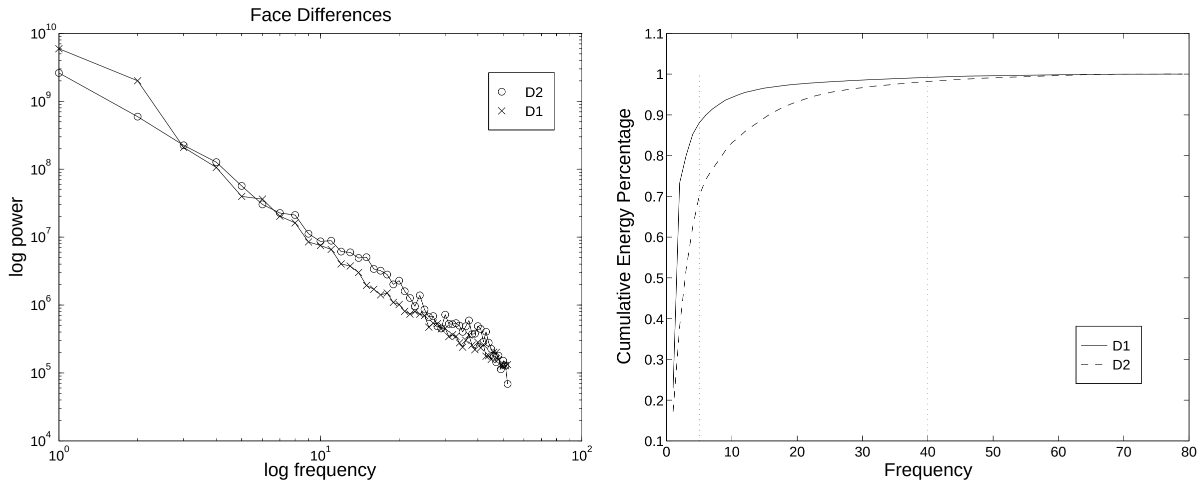


Figure 14: *Left:* the power spectrum of the two differences in Figure 13, plotted in log-log scales. The spectrum of $D1$ (denoted by $\times$'s) decays faster than the spectrum of $D2$ (denoted by o's), because it is an image of a smoother surface. The slope of the regression is -2.90 for $D1$, and -2.86 for $D2$. *Right:* although the fractal dimensions in the two cases are rather similar, the percentage of the energy of $D1$ in the pass band (indicated by the vertical lines) of the filter corresponding to the center-surround RFs in the simulation of area V1 of experiment 1 is one third that of $D2$, as shown in the right figure.

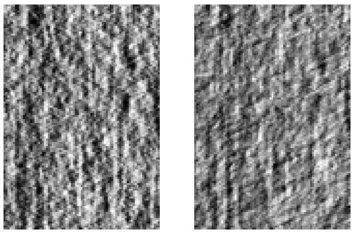
directions, while $D2$, on the right, is the difference of images of two different fractal surfaces under the same illumination conditions. As before, $D1$ is merely an image of a surface viewed under a third illumination condition, while $D2$ is not. However, $D1$ is no longer significantly smoother than $D2$, as can be seen in Figure 16. The energies of $D1$ and $D2$ in the pass band of the LGN/V1 RFs are nearly identical. Note that changing the size of the RF (and consequently the spatial frequency bandwidth) would not improve performance in this case.[7]

4 General discussion

4.1 Faithful representation of shape similarity as a goal of sensory coding

According to the similarity hypothesis advanced in the present work, an important goal of early visual processing is a remapping of the input into a space where proximal (represented) similarity covaries with the distal (true) similarity between objects (Edelman, 1993). Theoretical results concerning the representational properties of the first few stages of the visual pathway, and simulations of the action of the RF assemblies found there on image similarity, support

[7] A comparison of the power spectra in Figures 16 and 14 shows that the fractal dimension of the fractal images is not equal to that of the face images. Apparently some of the assumptions of Kube and Pentland (Lambertian surface, uniform albedo) are not satisfied in this case.



$D1$ $D2$

Figure 15: Two image differences: $D1$, on the left, is a difference between two images of the same fractal surface under different illuminant directions. $D2$, on the right is the difference of two images of two different fractal surfaces under the same illumination conditions.

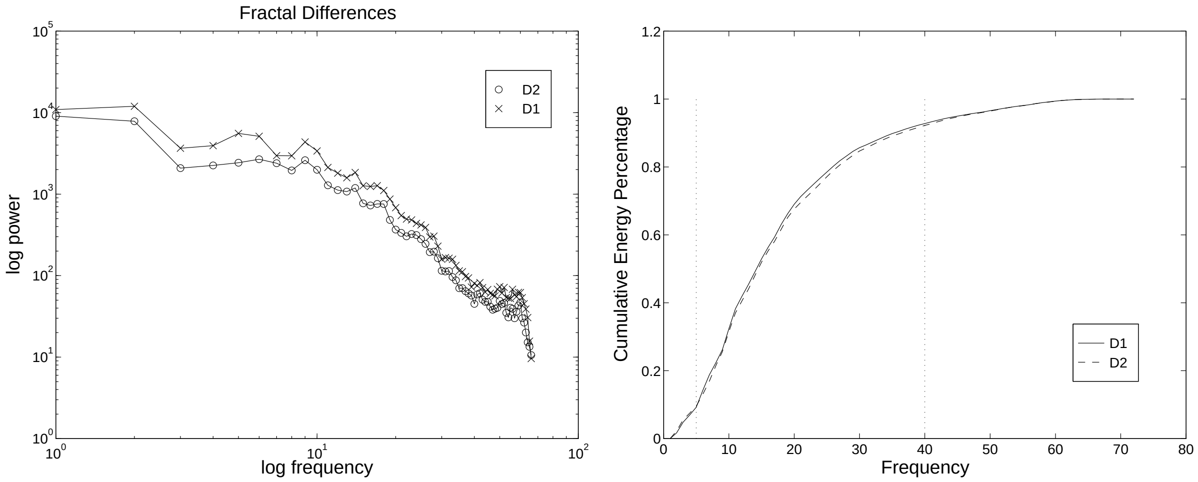


Figure 16: *Left:* the power spectrum of the two differences in Figure 15, plotted in log-log scales. The spectrum of $D1$ (denoted by $\times$'s) does not decay significantly faster than that of of $D2$ (denoted by o's). The slope of the regression is -1.817 for $D1$ and -1.792 for $D2$. *Right:* the percentage of the energy of $D1$ in the pass band of the LGN/V1 RFs is almost identical to that of $D2$, as shown in the right figure.

this hypothesis. Specifically, we showed in section 2 that simple linear representations such as those found early in the visual pathway may encode the similarity between images in a manner that is significantly different from the representation of the same images at the pixel (or photoreceptor) level. Furthermore the computational experiments described in section 3 indicate that the successive transformations of similarities lead to a representation that is more useful for the recognition of natural objects such as faces. These results explain in part the success of machine vision systems that employed representations derived from activities of linear receptive fields (Edelman et al., 1992; Jones and Malik, 1992).

The present contribution to the understanding of the sensory coding does not include a prediction concerning the shape of receptive field that is optimal under the similarity hypotheses. Such a prediction requires a precise formulation of the optimal similarity measure for vision, beyond the general statement that the representations of different views of the same 3D object should be closer than those of views of different objects. Consider the cloud of points in the high-dimensional space spanned by the RF activities, formed by representations of different views of the same 3D object. Views of another object in this space would form a second cloud of points; if the shapes of these two clouds and their disposition were known, it would be possible to define an optimal family of receptive fields, whose induced similarity contours would conform to the shapes of the object clouds. However, the shapes of the clouds, along with the derived RFs, depend on the properties of the objects, resulting in RFs that can only be considered optimal for the given object class (Intrator and Gold, 1993). Both the determination of the typical shape of object clouds in RF space, and the search for an RF profile that would perform equally well in representing a wide variety of object classes, are at present open research issues.

4.2 Comparison to other hypotheses regarding sensory coding

The similarity hypothesis may be compared with the other recently advanced hypotheses concerning the purpose of early visual transformations, e.g., efficiency of information representation (Atick, 1992), or production of a sparse distributed code (see Field, 1994, for an extensive review). In this connection, we offer two observations. First, it is important to realize that a representation that is optimal under one criterion may be far from optimal under another one. Consider, for example, representations obtained through dimensionality reducing techniques such as the Karhunen-Loeve Transform. The KLT is optimal in the sense that it finds a linear low-dimensional projection of the data in a manner that preserves as much as possible the similarities between images in the original space. Such a representation happens to be far from optimal for classification of 3D objects, because the image-based similarities that are preserved usually do not correspond to objective similarities. Indeed, results of section 3.4 indicate that the KLT representation performs as badly as the raw pixels, and significantly worse than the simulated LGN/V1 representation.

Our second observation is that differences between the optimization criteria employed by various studies of sensory coding do not necessarily mean that their findings lead to a conflict. Note that the similarity hypothesis regarding the goal of sensory coding differs from most of those reviewed, e.g., in (Field, 1994) by being based on an assumption as to what may constitute the ultimate goal of vision (namely, object classification), and not just the goals of its intermediate stages. Thus, it is not unthinkable that a given code optimizes simultaneously a number of criteria (say, sparseness and suitability for object classification), if these do not contradict each other.

Acknowledgments

Part of this research was completed when YW was at the Tel Aviv University. We thank Yael Moses for discussing with us some of her unpublished results, and for making available to us her database of face images, and E. Adelson for comments on a previous version of this manuscript. The support of the Grodetzky Center for the Study of Higher Brain Function at the Weizmann Institute of Science is gratefully acknowledged. YW is supported by a training grant from the National Institute of General Medical Sciences. SE is an incumbent of the Sir Charles Clore Career Development Chair at the Weizmann Institute.

References

Adini, Y., Moses, Y., and Ullman, S. (1993). Face recognition: the problem of compensating for changes in illumination direction. CS-TR 21, Weizmann Institute of Science.

Atick, J. J. (1992). Could information theory provide an ecological theory of sensory processing? *Network*, 3:213–251.

Atick, J. J. and Redlich, A. N. (1992). What does the retina know about natural scenes? *Neural Computation*, 4:196–210.

Daugman, J. G. (1988). An information-theoretic view of analog representation in the striate cortex. In Schwartz, E. L., editor, *Computational Neuroscience*, pages 403–423. MIT Press.

Davis, P. J. and Rabinowitz, P. (1975). *Methods of numerical integration.* Academic Press, New York.

Desimone, R. and Ungerleider, L. (1989). Neural mechanisms of visual processing in monkeys. In Boler, F. and Grafman, J., editors, *Handbook of Neuropsychology*, volume 2, pages 267–299. Elsevier, Amsterdam.

Edelman, S. (1993). Representation, Similarity, and the Chorus of Prototypes. CS-TR 93-10, Weizmann Institute of Science. to appear in Minds and Machines, 1994.

Edelman, S. and Poggio, T. (1992). Bringing the Grandmother back into the picture: a memory-based view of object recognition. *Int. J. Pattern Recog. Artif. Intell.*, 6:37–62.

Edelman, S., Reisfeld, D., and Yeshurun, Y. (1992). Learning to recognize faces from examples. In Sandini, G., editor, *Proc. 2nd European Conf. on Computer Vision, Lecture Notes in Computer Science*, volume 588, pages 787–791. Springer Verlag.

Field, D. J. (1987). Relations between the statistics of natural images and the response properties of cortical cells. *Journal of the Optical Society of America*, A 4:2379–2394.

Field, D. J. (1994). What is the goal of sensory coding? *Neural Computation*, 6:559–601.

Freeman, W. T. and Adelson, E. (1991). The design and use of steerable filters. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13:891–906.

Gibson, J. J. (1979). *The ecological approach to visual perception.* Houghton Mifflin, Boston, MA.

Hancock, P. J. B., Baddeley, R. J., and Smith, L. S. (1992). The principal components of natural images. *Network*, 3:61–70.

Intrator, N. and Gold, J. (1993). Three-dimensional object recognition in gray-level images: the usefulness of distinguishing features. *Neural Computation*, 5:61–74.

Intrator, N., Gold, J. I., Bülthoff, H. H., and Edelman, S. (1992). Three-dimensional object recognition using an unsupervised neural network: understanding the distinguishing features. In Moody, J., Hanson, S. J., and Lippman, R. L., editors, *Neural Information Processing Systems*, volume 4, pages 460–467. Morgan Kaufmann, San Mateo, CA.

Jones, D. G. and Malik, J. (1992). A computational framework for determining stereo correspondence from a set of linear spatial filters. In Sandini, G., editor, *Proc. ECCV-92*, pages 395–410, Berlin. Springer.

Kruskal, J. B. and Wish, M. (1978). *Multidimensional Scaling.* Sage Piblications, Beverly Hills, CA.

Kube, P. and Pentland, A. (1988). On the imaging of fractal surfaces. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 10:704–707.

Marr, D. (1982). *Vision.* W. H. Freeman, San Francisco, CA.

Moses, Y. (1993). *Computational approaches in face recognition.* PhD thesis, Feinberg Graduate School of the Weizmann Institute of Science.

Moses, Y. and Ullman, S. (1992). Limitations of non model-based recognition schemes. In Sandini, G., editor, *Proc. 2nd European Conf. on Computer Vision, Lecture Notes in Computer Science*, volume 588, pages 820–828. Springer Verlag.

Mundy, J. L. and Zisserman, A., editors (1992). *Geometric invariance in computer vision.* MIT Press, Cambridge, MA.

O'Toole, A., Abdi, H., Deffenbacher, K., and Valentin, D. (1993). Low-dimensional representation of faces in higher dimensions of the face space. *Journal of the Optical Society of America*, 10:405–410.

Pentland, A., Starner, T., Etcoff, N., Masoiu, A., Oliyide, O., and Turk, M. (1993). Experiments with eigenfaces. In *Looking at People Workshop, IJCAI 93.*

Poggio, T. and Edelman, S. (1990). A network that learns to recognize three-dimensional objects. *Nature*, 343:263–266.

Ratliff, F. and Sirovich, L. (1978). Equivalence classes of visual stimuli. *Vision Research*, 18:845–851.

Shashua, A. (1992). Illumination and view position in 3D visual recognition. In Moody, J., Hanson, S. J., and Lippman, R. L., editors, *Neural Information Processing Systems*, volume 4, pages 404–411, San Mateo, CA. Morgan Kaufmann.

Turk, M. and Pentland, A. (1991). Eigenfaces for recognition. *J. of Cognitive Neuroscience*, 3:71–86.

Ullman, S. and Basri, R. (1991). Recognition by linear combinations of models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13:992–1005.

A Proofs

We prove here the results stated in section 2. Most of the proofs are elementary and are given here for completeness.

Proposition 1. The equi-similarity contour for the Euclidean metric is an ellipsoid whose principal axes are eigenvectors of L^*L with elongation in direction v_i equal to ϵ/λ_i where λ_i is the corresponding eigenvalue .

Proof: This follows from the following, more general statement.

The class of all images whose similarity to a pattern m is less than or equal to ϵ is obtained from the ϵ-ball of the metric d in R^n centered on $R(m)$ through rotation, rescaling, and an orthogonal map to the image space.

Proof: The set $\{x \mid d[Lx, Lm] < \epsilon\}$ is by definition the ϵ-ball of the metric d in R^n, centered on m. To project this set into the image plane we apply the pseudoinverse of the mapping L to the ϵ-ball. The singular value decomposition of L is $L = V\Sigma W^*$ where the columns of V are the eigenvectors of LL^*, the columns of W are the eigenvectors of, L^*L and Σ is a diagonal matrix whose nonzero elements are the eigenvalues of LL^*. The pseudoinverse of L is given by $W\Sigma^+V^*$ where Σ^+ is still a diagonal matrix. Thus, applying the pseudoinverse to to the ϵ-ball means applying V^*, then Σ^+, and then W. Since V, W are orthogonal operators, and Σ^+ is a diagonal operator, this is equivalent to a rotation, rescaling and a final orthogonal map to the image space. If the operator L is invertible, this characterizes fully the generalization class. Otherwise, the generalization class should also be closed under addition of vectors in $KerL$.

Proposition 2. For a generalized linear representation, the similarity of two images x_1, x_2 is a weighted norm of $L(x_1 - x_2)$ where the weights depend on the two images being compared.

Proof: We assume that the metric d is derived from a norm in R^n and $d(Rx_1, Rx_2) = \|Rx^1 - Rx^2\|$ where $Rx = FLx$. Let $v = Rx^1 - Rx^2$, $y^1 = Lx^1$, $y^2 = Lx^2$. Then $v_i = f(y_i^1) - f(y_i^2) =$

$\dot{f}(\eta_i)(y_i^1 - y_i^2)$ where η_i is between y_i^1 and y_i^2. Thus $d(Rx_1, Rx_2) = \|DLx_1 - DLx_2\|$ where D is a diagonal operator whose values depend on x_1, x_2. Or: $d(Rx_1, Rx_2) = \|L(x_1 - x_2)\|_W$ where W depends on x_1, x_2.

Proposition 3. Let V be a matrix whose columns are orthonormal non-degenerate eigenvectors of LL^* : $LL^* = VDV^*$, where D is a diagonal matrix whose nonzero elements are positive. Then $W = L^*VD^{-1/2}$ is a matrix whose columns are orthonormal eigenvectors of L^*L with the same eigenvalues: $L^*L = WDW^*$ and $W^*W = I$

Proof:

$$WDW^* = L^*VD^{-1/2}DD^{-1/2}V^*L = L^*L \tag{2}$$

$$W^*W = D^{-1/2}V^*LL^*VD^{-1/2} = D^{-1/2}V^*VDV^*VD^{-1/2} = I \tag{3}$$

Proposition 4. If L is a convolution operator, then the eigenvectors of L^*L are sine and cosine gratings, and the eigenvalues are equal to the magnitude of the Fourier transform of the receptive field.

Proof: Since L is a convolution, then it is equivalent to multiplication in the Fourier domain, hence: $L = F^*DF$ where F is the Fourier transform, and F^{-1} its inverse. D is a diagonal operator whose nonzero elements are the (complex-valued) values of the Fourier transform of the receptive field. Therefore:

$$L^*L = F^{-1}D^*FF^{-1}DF = F^{-1}D^*DF \tag{4}$$

Thus L^*Lx is equivalent to multiplying the Fourier transform of x by the magnitude of the Fourier transform of the receptive field. Hence the complex exponentials are eigenvectors of L^*L, and since the receptive field is a real-valued function, the sine and cosine gratings are eigenvectors as well.

Proposition 5. If L is a multiple-convolution representation scheme, i.e., the receptive fields are translations of a family $RF(s,\theta)$, then L^*L is still diagonalized by sine and cosine gratings with the eigenvalue of the grating of spatial frequency ω_x, ω_y given by:

$$\lambda_i^2 = \sum_{s,\theta} |\hat{RF}(s,\theta;\omega_x,\omega_y)|^2 \tag{5}$$

Proof: We assume for simplicity that the family of receptive fields includes only two receptive fields RF^1,RF^2 (the proof for more receptive fields is analogous). Then L is a direct sum of two operators $L = L_1 + L_2$ with

$$L_1^*L_2 = L_2^*L_1 = 0 \tag{6}$$

The operators L_1, L_2 are convolutions with receptive fields RF^1 and RF^2 respectively. From equation 6 we obtain:

$$L^*L = L_1^*L_1 + L_2^*L_2 \tag{7}$$

Because L_1, and L_2 are convolution operators, by proposition 4 above the sine and cosine gratings are eigenvectors of both $L_1^*L_1$ and $L_2^*L_2$ and therefore they are also eigenvectors of L^*L with eigenvalues which are the sum of the eigenvalues with respect to $L_1^*L_1$ and $L_2^*L_2$.

Proposition 6. Let L be as above. Then L is equivalent to a representation whose receptive fields are shifted versions of one receptive field $\tilde{RF}$ whose Fourier transform is defined in equation 5. If the family of receptive fields includes all rotations of a single receptive field, then $\tilde{RF}$ is rotationally invariant.

Proof: Since the similarity measure induced by a representation is completely characterized by its principal features and sensitivities, the representation L is equivalent to convolution with a single $\tilde{RF}$. Since the Fourier transform of a rotated receptive field is a rotation of the Fourier transform of the receptive field, in the case where the family of receptive fields include all rotations of a single receptive field, the Fourier transform of $\tilde{F}$ is a sum of rotations of a single Fourier transform, and is therefore rotationally invariant. Since the Fourier transform is rotationally invariant, then the receptive field itself is also rotationally invariant.

Proposition 7. Let L_n be a sequence of subsampled convolution representations, i.e., the n receptive fields of L_n are translations of one receptive field, with the centers of the units placed on a uniform rectangular grid. Let L_∞ be the continuous cyclic convolution operator, and d_n denote the similarity measure induced by L_n. If the receptive field function is band-limited by f_{max} then $d_n/\sqrt{n} = d_\infty$ for all n such that the distance between centers of the receptive fields are less than $1/2f_{max}$.

Proof: We prove for the case of a one-dimensional signal. The extension to a two-dimensional image is straightforward.

Lemma: (Davis and Rabinowitz, 1975, p.110) Let $f(x)$ be periodic with period p:

$$f(x) = \sum_{n=0}^{\infty} \alpha_n \cos(\frac{2\pi nx}{p}) + \beta_n \sin(\frac{2\pi nx}{p}) \tag{8}$$

Define the integration error as

$$E_N(f) = \frac{p}{N} \sum_{k=0}^{N-1} f(\frac{k}{N}p) - \int_0^p f(x)dx \tag{9}$$

Then:

$$E_n(f) = p \sum_{k=1}^{\infty} \alpha_{kn} = p(\alpha_n + \alpha_{2n} + \ldots) \tag{10}$$

Now, denote by $h(t)$ the result of the convolution operator. $h(t)$ is periodic with a period equal to the dimension of the picture p. Since $h(t)$ is band-limited $\alpha_n = \beta_n = 0$ for all

$n \geq pf_{max}$. Therefore the coefficients of $h(t)^2$ obey $\alpha_n = \beta_n = 0$ for all $n \geq 2pf_{max}$. Hence $d_n^2/n - d_\infty^2 = E_n(h^2) = 0$.

Proposition 8. Let the n receptive fields of L_n be translations and rotations of one receptive field. Let $L_\infty(x, y, \alpha)$ be the response of a unit whose RF is centered on (x, y) and with orientation α. If the receptive field function $g(x, y)$ is band-limited by f_{max} and angularly band-limited by t_{max} then $d_n/\sqrt{n} = d_\infty$ for all n such that the distance between centers of the receptive fields are less than $p/2f_{max}$ and the distance between orientations of the receptive field at each location is less than π/t_{max}

Proof: The fact that $g(x, y)$ is angularly bandlimited implies that $L_\infty(x, y, \alpha)$ is bandlimited as a function of α for any fixed x, y. Consequently, if we denote by $\hat{L}(u, v, t)$ the 3D Fourier transform of L_∞, then $\hat{L}$ vanishes for $|u| > f_{max}, |v| > f_{max}, |t| > t_{max}$. By the Lemma in the previous proposition, d_n is exact if the sampling in space is finer than $p/2f_{max}$ and the sampling in orientation is finer than π/t_{max}

Proposition 9. Define a sequence $\{L_n\}$ of linear RF representations where the individual receptive fields constitute n independent random samples from a family of receptive fields $RF(\theta; \cdot)$ according to a probability density function $f(\theta)$. Define L_∞ $X \mapsto R$ by $(L_\infty x)(\theta) = < RF(\theta; \cdot), x >$. Let $d_n(x, y) = \|L_n x - L_n y\|$ where the norm is the Euclidean norm in R^n. Then:

$$\lim d_n^2(x, y)/n = \int f(\theta) \left[(L_\infty x)(\theta) - (L_\infty y)(\theta) \right]^2 d\theta \tag{11}$$

with probability one.

Proof: Let

$$h(\theta) = [(L_\infty x)(\theta) - (L_\infty y)(\theta)]^2] \tag{12}$$

and let H be a random variable, whose value is $h(\theta)$ for each random choice of θ. Then d_n^2/n is the average of n samples of H, and by the law of large numbers it converges, with probability one, to the expectation of H. This expectation is the right hand side of equation 11.

Proposition 10. Let L be a divergent linear representation with Euclidean metric, then there exists a non-divergent representation with the same induced similarity measure.

Proof: The similarity measure is completely characterized by the operator L^*L. Since L^*L is positive semidefinite, we can define $M = \sqrt{L^*L} = P^*\sqrt{\Sigma}P$ where P is an orthogonal operator which diagonalizes L^*L, and Σ is a diagonal operator. Obviously, $M^*M = L^*L$ and hence the two representations are equivalent. Since L^*L is a $D \times D$ matrix, so is M, and therefore the number of receptive fields in M is equal to the dimensionality of the input space.

B The RBF classifier

To illustrate the utility of the various representations for recognition, we computed the performance of an RBF classifier serving as a simple model of recognition by interpolation among multiple views (Poggio and Edelman, 1990). The RBF network consisted of three layers: (1) an input layer, which was fed with the activities of the RFs os the representation in question, (2) a hidden layer, in which each unit computed the distance between its "center" or prototypical view and the input image, and (3) an output layer. The hidden units computed distance according to the following expression:

$$h_i(x) = e^{-\|x-c_i\|^2/2\sigma_i^2} \tag{13}$$

The output layer consisted of four units that computed linear combinations of the activities of the hidden units, and that were trained to produce a unary representation of the image class (i.e., the required output vector for the i'th individual face was $[o_1, o_2, \ldots, o_k, \ldots]^T$ with $o_k = \delta_{ik}$, using the Kronecker δ notation). For a detailed description of the RBF network see (Edelman and Poggio, 1992). The simulation protocol was as follows:

- Iterate:
  - Choose random training images;
  - Incorporate training images as centers. Set σ_i to the average distance to example i;
  - Compute the weights using the Widrow-Hoff rule;
  - Test on the full data set.

The training was done either on one or on two images from each class.

C The construction of the fractal images

We constructed the four fractal surfaces by combining the power spectrum of a face image with four random phase spectra. This gave us the Fourier spectra of four surfaces. We then inverted the Fourier spectra and considered the results as depth maps of surfaces. The surfaces were then rendered using standard computer graphics techniques.

If the fractal dimension is an invariant of the imaging process (Kube and Pentland, 1988), this procedure should yield surfaces with the same fractal dimension as the surface of a face.