Biol. Cybern. 70, 37–45 (1993)

Biological Cybernetics

© Springer-Verlag 1993

Representing three-dimensional objects by sets of activities of receptive fields

Shimon Edelman

Department of Applied Mathematics and Computer Science, The Weizmann Institute of Science, Rehovot 76100, Israel

Received: 22 January 1993/Accepted in revised form: 24 May 1993

Abstract. Idealized models of receptive fields (RFs) can be used as building blocks for the creation of powerful distributed computation systems. The present report concentrates on investigating the utility of collections of RFs in representing three-dimensional objects under changing viewing conditions. The main requirement in this task is that the pattern of activity of RFs vary as little as possible when the object and the camera move relative to each other. I propose a method for representing objects by RF activities, based on the observation that, in the case of rotation around a fixed axis, differences of activities of RFs that are properly situated with respect to that axis remain invariant. Results of computational experiments suggest that a representation scheme based on this algorithm for the choice of stable pairs of RFs would perform consistently better than a scheme involving random sets of RFs. The proposed scheme may be useful under object or camera rotation, both for ideal lambertian objects, and for real-world objects such as human faces.

1 Introduction

Many of the lower-level areas in the primate visual system are organized retinotopically, that is, as maps which preserve to a certain degree the topography of the retina. A unit that is a part of such a retinotopic map normally responds selectively to stimulation in a well-localized part of the visual field, referred to as its receptive field (RF). Different regions within the RF may contribute differently to the activity of the unit, according to the profile or the weighting function of the RF. The activity of the unit is frequently modeled by a (possibly nonlinear) function of the convolution of the activity distribution over the input area with the RF profile[1].

Idealized models of receptive fields can be used as building blocks for the creation of powerful distributed computation systems. For example, it has been pointed out recently that networks of units with gaussian RF profiles constitute a universal tool for approximating smooth functions in multidimensional spaces (Poggio 1990; Poggio and Girosi 1990). In another example, RFs with a plastic profile obeying a dynamic weight and threshold modification rule were shown to be capable of carrying out unsupervised dimensionality reduction (Intrator 1992; Intrator and Cooper 1992).

The present report concentrates on investigating the utility of collections of RFs in representing three-dimen sional objects. The main requirement in this task is that of invariance: the representation, expressed as a pattern of activity of RFs, should vary as little as possible when the object that is represented undergoes rigid transformations in space[2]. The RFs are assumed here to be of fixed gaussian radially symmetric or elongated profile, and their placement over the image of the object is assumed to be random.

Representing an object by a pattern of activities of RFs can be considered a form of feature extraction. The utility of the resulting feature vector in view of the possible transformations of the input has been considered by Amari, who showed that linear feature extraction by the computation of moments or of Fourier components commutes with the euclidean transformations of the object in three dimensions (Amari 1968; Amari 1978). This approach has recently led to the formulation of nonlinear equations that link the three-dimensional orientation of a moving planar textured surface patch to moment-like features measured on its two-dimensional image [Amari and Maruyama 1987; see also Nishihara and Poggio (1984) and Aloimonos and Shulman (1989)].

Producing a new image by convolving an input with a set of RFs of finite spatial support results in a subsampled and blurred version of the original image. Nevertheless, a collection of activities of RFs in general carries information that allows precise location of spatial details in the input (Snippe and Koenderink 1992).

[1] This requires that the time dimension be ignored, and the RF profile be assumed shift-invariant (Mallot et al. 1990)

[2] Note that the extensive body of results on invariance in computer vision [see, e.g., Mundy and Zisserman (1992) for a recent survey] is not directly relevant to the present work, which is based on the assumption of representation by receptive fields, and is mainly motivated by biological considerations

Computer simulations have shown that simple models based on representation by RFs are capable of replicating human-like hyperacuity-level performance in spatial discrimination tasks (Poggio et al. 1992) and, to some extent, in the higher-level task of recognizing three-dimensional objects (Edelman 1992). In the latter work, each three-dimensional object was represented by a few of its views, encoded as vectors of activities of a large number of RFs, randomly placed over the input image. New views of an object were recognized by interpolation among the representations of the stored views: no explicit reconstruction of the three-dimensional shape of the object was attempted.

Recent psychophysical evidence indicates that a similar strategy may be employed by the human visual system in some recognition tasks (Bülthoff and Edelman 1992). In particular, the performance of human subjects was found to depend on the stimulus attitude. This dependence was more moderate when shaded images of animal-like three-dimensional shapes, such as the one depicted in Fig. 1, were used as stimuli (Edelman 1992). For these stimuli, the recognition rate remained well above chance for a wide range of stimulus orientations relative to a familiar attitude. In comparison, the performance of the view interpolation model that represented individual views of the stimulus by collections of locally averaged intensity values dropped to chance at a misorientation of 60° with respect to the nearest stored view.

The task of a model that relies on view interpolation in replicating human performance can be facilitated by making representations of individual views invariant with respect to viewpoint. Toward that end, I propose to choose from a large set of RFs a subset that satisfies two constraints:

- *Direct use.* To be directly useful for memory-based recognition, simple functions defined on the chosen ensemble of RF activities are to vary as little as possible under reasonable transformations of the viewpoint.
- *Multilocality.* The individual RF profiles are to be of limited spatial extent, in line with the known functional architecture of the early stages of the primate visual system. At the same time, at the ensemble level, the RFs must represent nonlocal features of the input.

The existing approaches to representation by RFs rely on varieties of linear features such as moments or Fourier components, are designed to support the recovery of three-dimensional characteristics of the input, and are mostly applicable to isolated textured surface patches of simple shape (usually planar or quadric). In contrast, the new method proposed below is based on the observation that activities of RFs properly located over the image of an object covary when the observed object rotates around a fixed axis.

2 RF representation of objects undergoing rotation in depth

2.1 Lambertian shading

Consider the situation depicted in Fig. 2, which shows a rigid object undergoing arbitrary rotation in depth. Pick at random two patches, p_1 and p_2, on the object's surface, and let p'_1 and p'_2 be the corresponding patches after a small rotation around the specified axis. Assume that there is a distant point light source in the direction **L**, that the object's surface is lambertian, and that the mean albedo at p_1 and p_2 is, respectively, ρ_1 and ρ_2. Then the

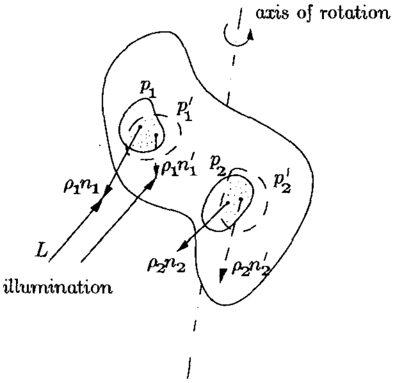


Fig. 2. Two patches, p_1 and p_2, with normals $\mathbf{n}_1$ and $\mathbf{n}_2$ on the surface of a rigid object undergoing a rotation around a fixed axis. The albedo coefficients at the two patches are ρ_1 and ρ_2. Following a finite rotation step, the patches are at p'_1 and p'_2

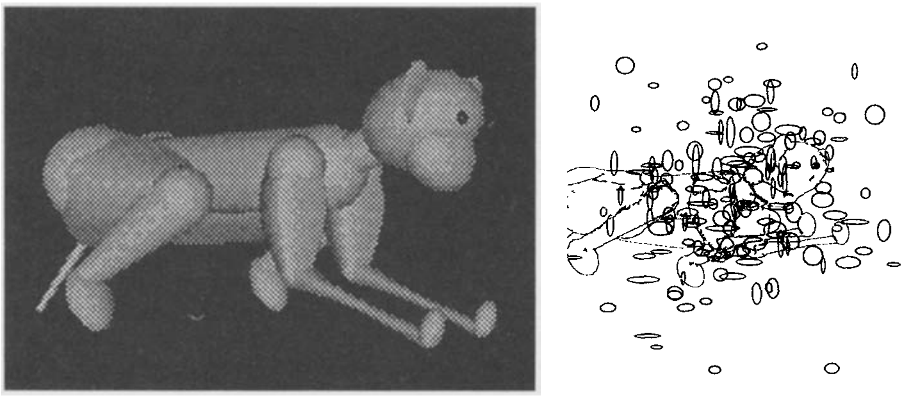


Fig. 1. *Left* An image of a three-dimensional object (from Edelman 1992); *right* representing the object shown at the left by a set of activities of 150 receptive fields (the image of the object superimposed on the map of receptive fields has been subjected to edge detection for clarity of presentation

intensities at the two patches before rotation are

$$I_1 = \mathbf{L}\cdot(\rho_1\mathbf{n}_1) \tag{1}$$

$$I_2 = \mathbf{L}\cdot(\rho_2\mathbf{n}_2) \tag{2}$$

where $\mathbf{n}_1$ and $\mathbf{n}_2$ are the surface normals at p_1 and p_2. Following the rotation, the intensities are

$$I'_1 = \mathbf{L}\cdot(\rho_1\mathbf{n}'_1) \tag{3}$$

$$I'_2 = \mathbf{L}\cdot(\rho_2\mathbf{n}'_2) \tag{4}$$

where I have used the assumption of a distant light source to equate $\mathbf{L}'$ with $\mathbf{L}$. Taking the difference between intensities of the two patches, one obtains

$$\begin{aligned}\Delta I &= I_2 - I_1\\ &= |\mathbf{L}|\,|\rho_2\mathbf{n}_2 - \rho_1\mathbf{n}_1|\cos\theta \end{aligned}\tag{5}$$

$$\begin{aligned}\Delta I' &= I'_2 - I'_1\\ &= |\mathbf{L}|\,|\rho_2\mathbf{n}'_2 - \rho_1\mathbf{n}'_1|\cos\theta' \end{aligned}\tag{6}$$

where θ (θ') is the angle between $\mathbf{L}$ and $\rho_2\mathbf{n}_2 - \rho_1\mathbf{n}_1$ before (after) the rotation. Because the object was assumed rigid, we have

$$|\Delta\mathbf{n}| = |\rho_2\mathbf{n}_2 - \rho_1\mathbf{n}_1| = |\Delta\mathbf{n}'| = |\rho_2\mathbf{n}'_2 - \rho_1\mathbf{n}'_1| \tag{7}$$

This means that the magnitude of the vector $\rho_2\mathbf{n}_2 - \rho_1\mathbf{n}_1$ that expresses the difference of orientation between patches p_1 and p_2 is invariant under the rotation. Thus, if the quantity ΔI changes following rotation (that is, if $\Delta I \neq \Delta I'$), this could be only due to a change in the *orientation* of the vector $\rho_2\mathbf{n}_2 - \rho_1\mathbf{n}_1$ with respect to the direction of the illumination L.

In the special case when the vector $\rho_2\mathbf{n}_2 - \rho_1\mathbf{n}_1$ is parallel to the axis of rotation, the angle θ will not change, and, consequently, the difference in intensity between the two patches, ΔI, will remain invariant under rotation. Consider now a set of locally averaged measurements of intensity such as the one provided by the set of receptive fields in Fig. 1. To obtain an invariant representation of an object by a subset of those measurements, one should pick pairs of RFs for which the difference in activity is stable over small rotations of the object. For any such pair of RFs, and for a fixed axis of rotation, ΔI will then remain stable[3]. A snapshot of activities of the chosen set of RF pairs can be used to represent the object. For a different object, another set of RF pairs will have to be picked.

To get a clearer idea of the conditions for the invariance of ΔI for two RFs, let us recall Euler's theorem stating that in a rigid rotation of a three-dimensional object in space, all points move around a common fixed axis[4]. The normals to a surface patch before and after rotation around a unit vector **a**, through an angle α, are therefore related by the Rodrigues formula (e.g., Kanatani 1990, p. 204). For clarity, and without loss of generality, the albedos $\rho_{1,2}$ are omitted from the expressions below:

$$\mathbf{n}'_1 = \mathbf{n}_1\cos\alpha + (\mathbf{a}\times\mathbf{n}_1)\sin\alpha + (1-\cos\alpha)(\mathbf{a}\cdot\mathbf{n}_1)\mathbf{a} \tag{8}$$

$$\mathbf{n}'_2 = \mathbf{n}_2\cos\alpha + (\mathbf{a}\times\mathbf{n}_2)\sin\alpha + (1-\cos\alpha)(\mathbf{a}\cdot\mathbf{n}_2)\mathbf{a} \tag{9}$$

By subtracting (8) from (9), one obtains the following expression relating the difference between the normals before and after rotation:

$$\begin{aligned}\Delta\mathbf{n}' = {}&\Delta\mathbf{n}\cos\alpha + (\mathbf{a}\times\Delta\mathbf{n})\sin\alpha\\ &+ (1-\cos\alpha)(\mathbf{a}\cdot\Delta\mathbf{n})\mathbf{a}\end{aligned}\tag{10}$$

From (10) it is obvious that the rotation leaves the orientation of $\Delta\mathbf{n}$ unchanged if $\mathbf{a}\times\Delta\mathbf{n} = 0$ (that is, if the difference between the normals is parallel to the axis of rotation), or, trivially, if $\Delta\mathbf{n} = 0$ (that is, if the two normals coincide)[5].

Recall now that we are really interested in the invariance of the product $\Delta I = \mathbf{L}\cdot\Delta\mathbf{n}$, and not of $\Delta\mathbf{n}$ on its own. This brings up another special case of interest: when the direction of illumination coincides with the axis of rotation ($\mathbf{L} = \mathbf{a}$), we have

$$\begin{aligned}\Delta I' &= \mathbf{L}\cdot\Delta\mathbf{n}'\\ &= \mathbf{L}\cdot\Delta\mathbf{n}\cos\alpha + \mathbf{L}\cdot(\mathbf{L}\times\Delta\mathbf{n})\sin\alpha\\ &\quad + \mathbf{L}\cdot(1-\cos\alpha)(\mathbf{L}\cdot\Delta\mathbf{n})\mathbf{L}\\ &= \mathbf{L}\cdot\Delta\mathbf{n}\\ &= \Delta I\end{aligned}\tag{11}$$

This situation occurs, e.g., when the illumination is by a uniformly lit hemispherical sky, and the object rotates around the vertical axis.

2.2 *Torrance–Sparrow shading*

If the lambertian approximation is not valid (e.g., if the reflectance depends on the emergent angle, and not

[3] This can be expressed as $\frac{d}{dt}\Delta I = 0$, bringing to mind the definition of optic flow in terms of the intensity gradient: $\frac{d}{dt}\nabla I = 0$ (Poggio et al. 1989). If the gradient of I can be approximated from a number of measurements of ΔI in a given neighborhood, the computational relationship between the issues of representation by RFs and the estimation of optic flow would be worth further exploration

[4] According to Chasles' theorem, any rigid motion in space can be decomposed into a combination of rotation around an axis and a translation along the same axis [see, e.g., Koenderink and van Doorn (1986)]. Here I assume that the translation component vanishes. This would happen, e.g., if the object is actively tracked. For a discussion of the advantages of tracking and other active modes of visual processing, see Aloimonos et al. (1988)

[5] As pointed out by a reviewer, the quantity $\Delta\mathbf{n}$ carries in it information related to the gaussian curvature of the surface. However, this information can only be extracted (1) if $\Delta\mathbf{n}$ is measured for a sufficient number of points around the locus of interest and (2) if inferences about each $\Delta\mathbf{n}$ can be made based on the measurement of the corresponding ΔI. As we have seen, the latter condition is only satisfied in special cases. Thus, there seems to be no straightforward way in which the proposed method can be extended to deal with the recovery of representations based on the gaussian map of the surface

only on the incident angle, at the surface), the conditions under which representations by activities of RFs are invariant become more restrictive. Consider the Torrance–Sparrow reflectance model, which includes, in addition to the diffuse or lambertian component, a specular component:

$$\begin{aligned} I &= I_{\text{diff}} + I_{\text{spec}} \\ &= c_1(\mathbf{L}\cdot\mathbf{n}) + c_2\frac{DGF}{(\mathbf{V}\cdot\mathbf{n})} \qquad (12) \end{aligned}$$

where D is the distribution function of the directions of microfacets on the surface, G is the amount by which facets shadow and mask each other, and F is the Fresnel reflection formula that gives the fraction of light incident on a facet that is reflected, as opposed to being absorbed (Torrance and Sparrow 1966). The coefficients c_1 and c_2 express the relative contributions of diffuse and specular reflection. The denominator of the specular term in (12), which is equal to the angle of slant of the surface with respect to the viewing direction $\mathbf{V}$, arises because the observer sees more of the surface area when the surface is slanted.

The facet distribution function D is a negative exponential in the angle between the surface normal and the so-called direction of maximum highlights $\mathbf{H} = (\mathbf{L} + \mathbf{V})/|(\mathbf{L} + \mathbf{V})|$:

$$D = e^{-s(\mathbf{H}\cdot\mathbf{n})} \qquad (13)$$

where s is a measure of shininess of the surface[6].

Because I now has two additive components, one of which depends exponentially on the product of $\mathbf{n}$ with $\mathbf{H}$, we can use neither the difference nor the ratio of Is at two patches to form an entity that would be invariant under object rotation. However, we can still look for pairs of patches with similar orientations ($\mathbf{n}_1 = \mathbf{n}_2$), which will yield ΔI equal to 0, and be assured that for such pairs $\Delta I'$ will also vanish.

Alternatively, if the specular component dominates the diffuse one in Eq. 12 (e.g., if $c_1 \ll c_2$), then the difference of the logarithms of Is at two patches will be

$$\begin{aligned} \log(I_1) - \log(I_2) &= \log\left(\frac{I_1}{I_2}\right) \\ &= \log\left(\frac{D_1}{D_2}\right) \\ &= \log\left[e^{-s(\mathbf{H}\cdot\mathbf{n}_1 - \mathbf{H}\cdot\mathbf{n}_2)}\right] \\ &= (-s\log e)(\mathbf{H}\cdot\Delta\mathbf{n}) \qquad (14) \end{aligned}$$

where I have assumed that G, and F at the two patches, are the same, and that $|\log(\mathbf{V}\cdot\mathbf{n}_1) - \log(\mathbf{V}\cdot\mathbf{n}_2)| \ll |(-s\log e)(\mathbf{H}\cdot\Delta\mathbf{n})|$[7]. In this case, the log intensity ratio behaves similarly to the ΔI of (11), and a sufficient condition for invariance across object rotation is again that $\Delta\mathbf{n}$ be parallel to the axis of rotation.

3 Representation of objects under fixed-aim camera rotation

Consider now an object at rest with respect to the illuminant, circumnavigated by a camera whose aim remains fixed at some point on the object. Obviously, if the object's surface is lambertian, the representation by RF activities will remain invariant (as long as the RFs continue to see the same patches of surface in the consecutive frames). Unlike in object rotation, in this case it does not even matter what is the relative orientation of the surface normals and the axis of rotation: the entire pattern of RF activities will be fixed.

Under Torrance–Sparrow shading, the same two cases as in object rotation (Sect. 2.2) can be discerned. If the diffuse component is nonnegligible, we will have to pick patches with $\Delta\mathbf{n} = 0$, for which the change in the vector $\mathbf{H} = k(\mathbf{L} + \mathbf{V})$, caused by rotation of the camera direction $\mathbf{V}$, will have no effect on ΔI. If the specular component predominates, then the log intensity ratios before and after camera rotation will be

$$\begin{aligned} R(I) &= \log(I_1/I_2) = -s\log e(\mathbf{H}\cdot\Delta\mathbf{n}) \\ R(I)' &= \log(I'_1/I'_2) = -s\log e(\mathbf{H}'\cdot\Delta\mathbf{n}) \qquad (15) \end{aligned}$$

Note that under camera rotation it is $\mathbf{H}$ that changes (because of the change in the viewer direction $\mathbf{V}$), and $\Delta\mathbf{n}$ stays constant. Thus, under camera rotation the log intensity ratio will not be invariant (unless it is equal to 1), and will change along with the relative orientation of the camera and the object.

4 Simulations

To assess the viability of RF-based representation, I have conducted four computational experiments that involved synthetically rendered images of computer-generated monkey-like and dog-like objects, as well as real human face images from a database taken under controlled orientation, illumination, and expression conditions (Moses Y, Edelman S, Ullman S, 1993). In all cases, the 8-bit gray-scale images were of size 512 × 352 pixels.

[6] In the original Torrance–Sparrow model, the exponent in (13) was squared. It has been observed, however, that the exact form of the dependence of D on $\mathbf{H}\cdot\mathbf{n}$ is not critical, as long as it satisfies some basic requirements, such as having a maximum at $\mathbf{H}\cdot\mathbf{n} = 0$, and a fast falloff rate for $\mathbf{H}\cdot\mathbf{n} > 0$ (Blinn 1988)

[7] It should be noted that the assumptions made in the preceding analysis are only of interest insofar as the proposed representation scheme withstands an empirical test on real-world images. If such a test succeeds, we will have learned something about the possible form of the reflectance function of the surfaces involved in the test

4.1 *Experiment 1: rotating synthetic objects, static camera*

The animal shapes whose images were used in the first experiment were created and rendered on a Silicon Graphics 4D35/TG workstation. The material reflectance model of the object surface included ambient and diffuse components of equal strength, and no specular component (thus, the lambertian model applies in this case). There was a single infinitely distant point of light, situated behind the simulated camera. To create the sequence of test images (Fig. 3), the objects were rotated around the vertical axis by increments of 15°.

The first two frames in a sequence were always used to identify the most stable pair of RFs (that is, the pair for which the difference ΔI changed the least from the first frame to the second one). The set of RFs from which the pairs were picked was chosen at random at the beginning of each trial. In most of the results reported below, the confidence limits for the various measurements were provided by computing the standard errors of the dependent variables over ten trials, in each of which the number of RFs was constant, but their placing on the image varied.

The utility of the most stable pair of RFs chosen as an invariant feature of the object was then assessed by computing several indices of invariance over the entire sequence of frames. The degree of invariance of ΔI for the most stable pair was compared with a similarly defined measure computed over ten randomly picked pairs of RFs. The measure of invariance (or, rather, of variability) I used was the standard deviation of ΔI, denoted below by *STD*. Plots of *STD* vs the number of randomly placed RFs, and vs the size of each RF, appear in Fig. 4. From these plots, it appears that, for size of $RF \approx 2$ pixels, using pairs of RFs chosen according to the proposed scheme has a clear advantage over using random pairs of RFs to represent the rotating object.

4.2 *Experiment 2: static human faces, rotating camera*

Experiment 2 involved images of three persons, taken from a sequence of viewpoints, under a constant frontal illumination. The four successive views were obtained by rotating the camera with its aim fixed at the center of the person's head. The rotation step was equal to 17°, so that the total rotation in depth between the first and the last view in the sequence was 51° (see Fig. 5; images produced under even larger rotations, available in the database, were not used because at large rotations self-occlusion limits the applicability of the method).

The results of the experiment with face image sequences (see Fig. 6) suggest that a representation scheme based on choosing stable pairs of RFs would perform consistently better than a scheme involving random sets of RFs. Together with the results of the previous experiment, this indicates that the proposed scheme may be useful under object or camera rotation, both for ideal lambertian objects, and for real-world objects such as human faces.

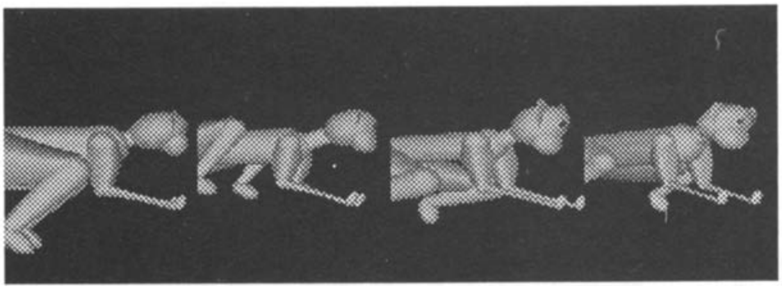


Fig. 3. A sequence of images of one of the two animal-like shapes used in experiment 1. The images were of size 512 × 352 pixels and are shown here reduced by a factor of two. The successive images were obtained by rotating the simulated camera in 15° steps around the vertical axis. In addition, the objects were allowed to undergo limited deformation, such as changes in the angles of the limbs and in the parameters that determined the shape of the body [see Edelman (1992) for details]

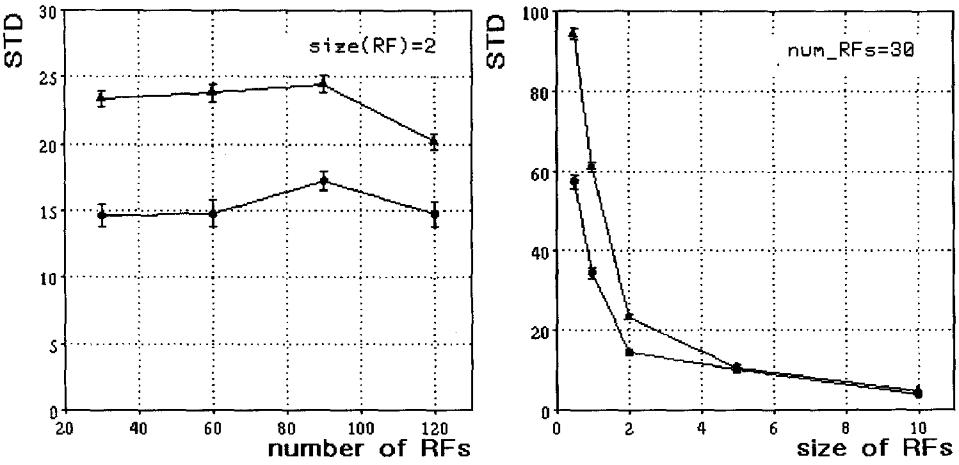


Fig. 4. Experiment 1: behavior of the standard deviation *STD* of ΔI, for random choices of RF pairs (*upper curves*) and for pairs that yield minimum change of ΔI between the first and the second frames in a sequence. The plots represent means computed over 10 runs per condition, using the monkey image sequence. The set of RFs was chosen randomly in the beginning of each run. *STD* is plotted vs the number of RFs, and vs the size of each RF (defined as the standard deviation of its gaussian profile, measured in pixels). Error bars denote standard error of the mean over the 10 runs

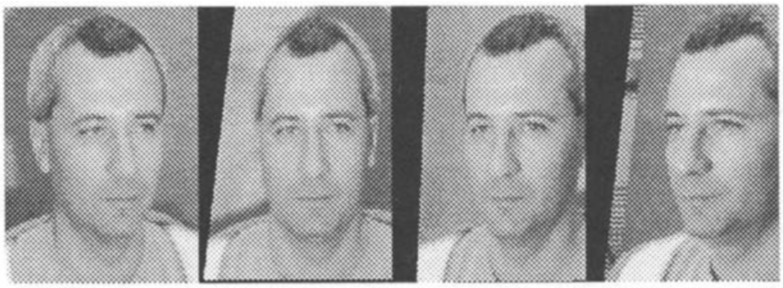


Fig. 5. A sequence of face images of one of the three persons used in experiment 2. The images were of size 512 × 352 pixels and are shown here reduced by a factor of two. All images in this figure were manually warped so that the locations of the eyes and the corners of the mouth remained the same as in the first image [an automatic scheme for such normalization is described in Edelman et al. (1992)].

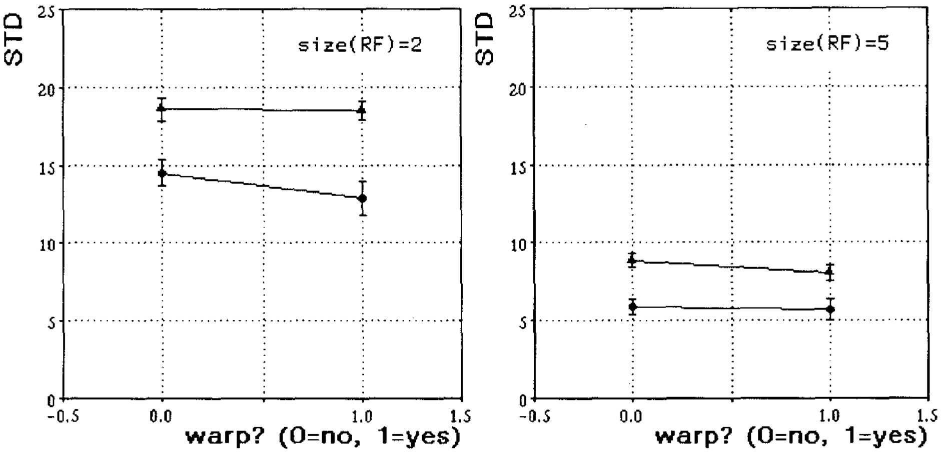


Fig. 6. Experiment 2: behavior of the standard deviation *STD* of ΔI for random choices of RF pairs (*upper curves*) and for pairs that yield minimum change of ΔI between the first and the second frames in a sequence. *STD* is plotted vs warp, a binary variable indicating whether the face images were normalized prior to processing, using the procedure explained in Fig. 5. The plots represent means computed over 10 runs per condition, one of the three available image sequences. The set of RFs was chosen randomly in the beginning of each run. Two sizes of RFs were tested, as indicated in the legends. Error bars denote standard error of the mean over the 10 runs

4.3 Experiment 3: stable RF pairs as a signature of the object

Objects that have patches of similar inclination with respect to the illuminant direction in the same retinal locations will be indistinguishable under representation by stable RF pairs. This scheme, therefore, should not be used on its own for object classes prone to such confusion. The next experiment was designed to estimate the degree to which the set of locations of the stable RF pairs could serve as a distinctive feature or a signature of a given face. The experiment consisted of ten trials, each of which started with the generation of a random set of 30 RFs (Fig. 7). Two most stable RF pairs were then picked for each of the three face image sequences. The six resulting pairs turned out to be distinct in eight of ten trials. In the other two trials, the same RF came out as the first choice for one individual, and as the second choice for another individual. Thus, even for objects such as faces, whose three-dimensional structure is fairly similar across individuals, the location of the stable RF pairs has the potential of serving as a distinctive feature in recognition.

4.4 Experiment 4: relative frequency of appearance of stable and unstable RF pairs

The last experiment addressed the question of the place of the *STD* characteristic of a stable pair of RFs in the distribution of the *STD* values for all pairs of RFs in a given population. Histograms of *STD* of all pairs of

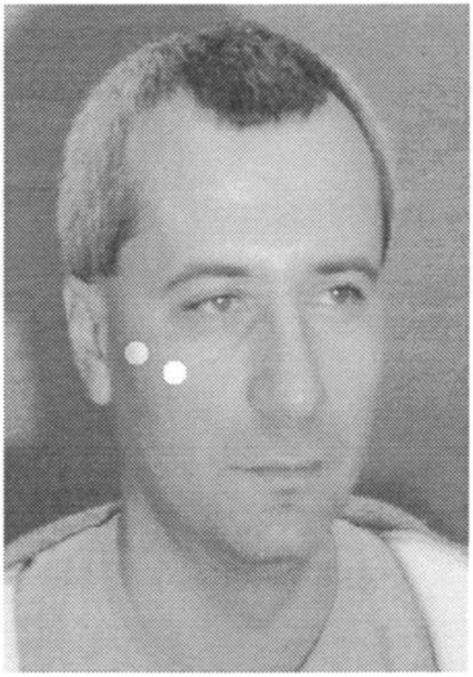
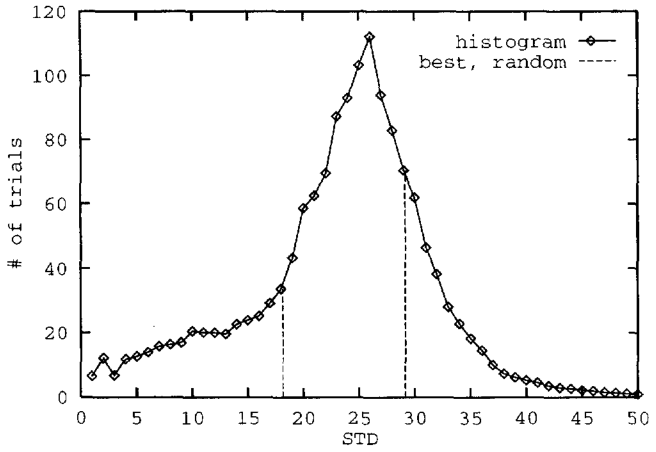


Fig. 7. An example of a stable pair of RFs found by processing a sequence of face images. The *STD* of this pair was 3.1, compared with 24.8 for an average of 10 randomly picked pairs of RFs for the same sequence of images. The total number of RFs in this example was 30. The RFs were circularly symmetric, gaussian, with σ uniformly distributed between 0.5 and 1.5 pixels



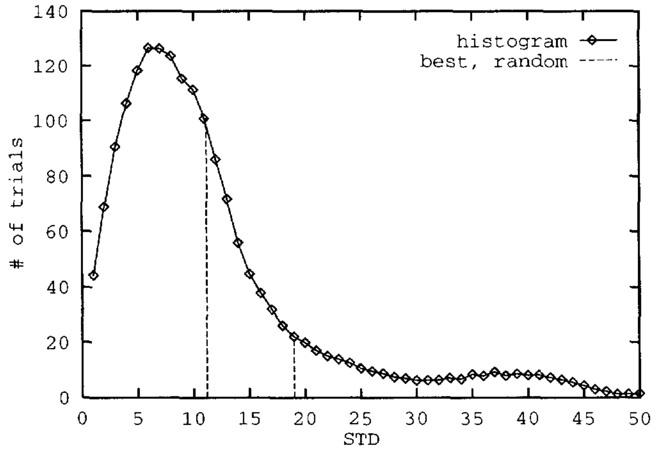


Fig. 8. *Top* Histogram of the variation of the difference between RF activities (*STD*), for all possible RF pairs. The histogram is averaged over 100 runs; each run involved a newly created set of 60 RFs. The *STD* values (also averaged over the 100 runs) for the stable pair of RFs and for a randomly chosen pair are shown by *vertical dashed lines* (the left and the right lines, respectively). Even though on the average the value of *STD* for the best RF pair is considerable, it is still much smaller than the mean *STD* for a random pair of RFs. In this experiment, the objects, which underwent rotation in depth, were animal-like shapes. *Bottom* A similar histogram, computed for the face images (camera rotation)

RFs for a population of size 60 appear in Fig. 8 for animal-like objects under object rotation (top) and for faces under camera rotation (bottom). The values of *STD* for a stable RF pair and for a randomly chosen pair are indicated by vertical lines in the histograms. This figure shows that for a sizable proportion of RF pairs the *STD* variation is significantly smaller than that of a randomly picked pair, and that stable pairs can be reliably identified by the proposed method.

5 Discussion

5.1 Ecological and computational considerations

The proposed method of picking stable pairs of RFs is applicable, in principle, to the construction of representations invariant with respect to any axis of rotation, as long as it is fixed in space, and the coverage of the input by the ensemble of RFs is dense enough. Restricting the space of allowed rotations would, however, increase the likelihood of a successful choice of RFs. This may be possible, if one considers ecological constraints on the transformations likely to be encountered in normal situations. Notably, at first glance, it appears that most rotations in depth are around the vertical axis[8], some are around the horizontal axis, and rather few are around oblique axes. It would be interesting to verify the validity of this assumption by computing appropriate statistics of natural scene sequences seen by an active observer subject to kinematic constraints imposed by primate anatomy.

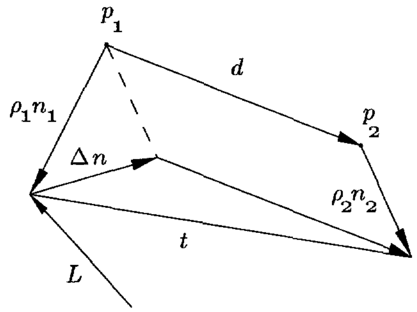


Fig. 9. A sketch of the vectors relevant to the determination of relative intensities at points p_1 and p_2. The invariance of the difference of intensities with respect to rotation around an axis **a** depends on the orientation of $\Delta\mathbf{n}$, not **d**, relative to **a**

The assumption of the predominance of rotations around, say, the vertical axis does not, unfortunately, help much in deciding which pairs of RFs should be wired together in advance of learning to represent objects. One may reason that the predominance of rotations around the vertical axis calls for connecting preferentially to each other those RFs that are situated along the retinal vertical meridians. This suggestion, however, is wrong, because what matters in determining whether ΔI for a pair of RFs will be invariant is the orientation of $\Delta\mathbf{n}$, and not of $\mathbf{d} = (p_1 p_2)$ alone, with respect to the illumination vector **L** (see Fig. 9). Nevertheless, psychophysical evidence on lateral masking/facilitation indicates that receptive fields situated along vertical and horizontal meridians may interact more strongly than those located in random locations relative to each other (Polat and Sagi 1992).

A possible computational rationale for this pattern of connections may have to do with the issue of nonuniform foreshortening of different patches of nonplanar objects under rotation in depth. Recall that a basic condition for the use of representation by RFs is that each RF be situated over roughly the same surface patch in the successive images of a rotating object. This problem,

[8] For the purpose of the present analysis, any translation of the observer with respect to a static scene – a very common natural motion – has the same effect as rotation in depth

which is an instance of the correspondence problem for motion (Ullman 1979), can be partially alleviated by tracking a point on the object's surface, say, one of the two RFs in a pair. In that case, however, the other RF may still get out of range of its corresponding patch, unless the line connecting the two patches (the vector **d** in Fig. 9) is parallel to the axis of rotation. Requiring that the two patches be foreshortened by similar amounts leads to the same constraint on the direction of **d**.

5.2 Biological considerations

In the above discussion, I have assumed that the individual RFs are excitatory and that pairwise differences between RF activities, realized via "lateral" connections, serve as a basis for the choice of a stable representation. In mammals, already at the level of the input to the primary visual cortex, the RFs possess an internal structure in which excitatory and inhibitory regions can be discerned (e.g., Hubel and Wiesel 1959). Such RFs can support the computations necessary for implementing the proposed representation scheme, provided that their excitatory and inhibitory subfields are spatially distinct, at least to some degree[9]. Indeed, in the so-called circularly symmetric RFs the center and the surround are frequently displaced with respect to each other (Kuffler 1953; Dawis et al. 1984). In the simple cells in striate cortex [e.g., in the bimodal cells described by Bishop et al. (1973)] as well as in the complex cells (Spitzer and Hochstein 1988) a spatial asymmetry between excitatory and inhibitory components can also be observed[10]. It should be noted that other computational roles of opponent RFs require that the excitatory and the inhibitory regions coincide rather than be apart from each other [e.g., as in the lightness computation model of (Hurlbert and Poggio 1988)]. Thus, the observed values of spatial asymmetry in opponent RFs may be the result of a compromise between conflicting requirements such as those of representational invariance with respect to rotation and of the ability to extract lightness.

5.3 Utility of the method

Even a partially invariant representation of three-dimensional objects could be of a considerable use to a visual system, especially if the latter relies on multiple stored views in the process of recognition. Systems based on multiple-view interpolation (e.g., Poggio and Edelman 1990) perform poorly on radically unfamiliar views of three-dimensional objects. As mentioned in the introduction, for some classes of stimuli a similarly poor performance is exhibited by human subjects (Bülthoff and Edelman 1992). In other cases humans successfully generalize recognition across large changes in object orientation relative to a single familiar view, while a view-interpolation model performs at a chance level for large misorientations (Edelman 1992). In these cases, endowing the representation of each familiar view of an object with partial invariance with respect to viewpoint could help the computational model approach human performance.

6 Conclusion

I have outlined a method that, given a three-dimensional object allowed to rotate in space, constructs its representation in terms of activities of a small number of receptive fields. This representation remains relatively stable over a range of object orientations and may be useful as an input to a memory-based recognition scheme, such as the one recently shown to replicate certain features of human performance in three-dimensional object recognition and classification tasks (Bülthoff and Edelman 1992; Edelman 1992). The proposed representation method may be extended computationally, to take advantage of the effects of tracking and size normalization, and may serve as a basis for three-dimensional object representation in computer vision systems. Future work will also consider a psychophysical investigation of this method and the possible interpretation it offers for recent neurobiological (Katz and Callaway 1992; Malach et al. 1992) and psychophysical (Polat and Sagi 1992) findings on lateral connections in the primary visual cortex of primates.

Acknowledgements. This work is supported by a grant from the Basic Research Foundation, administered by the Israel Academy of Arts and Sciences. I thank Shabtai Barash for general discussions and encouragement, Hedva Spitzer for useful references, and an anonymous referee for constructive and detailed suggestions.

References

Aloimonos JY, Shulman D (1989) Integration of visual modules: an extension of the Marr paradigm. Academic Press, Boston

Aloimonos JY, Weiss I, Bandopadhay A (1988) Active vision. Int J Comput Vision 2:333–356

Amari S (1968) Invariant structures of signal and feature spaces in pattern recognition problems. RAAG Mem 4:553–566

Amari S (1978) Feature spaces which admit and detect invariant signal transformations. In: Proc 4th Intl Conf on Pattern Recognition, Tokyo, pp 452–456

[9] Temporal aspects of biological RFs may also be relevant to the present analysis, as suggested in Sect. 2.1. The investigation of those is left for future work

[10] The existence of RFs with spatially distinct excitatory and inhibitory regions has been invoked by Harris and Gibson (1968) as an explanation of the McCollough effect (a contingency aftereffect, in which adaptation to a vertical red grating shown in alternation with a horizontal green grating causes similarly oriented black and white gratings to be perceived in colors complementary to the adapted ones; see McCollough 1965). In this connection, Harris (1980, p. 125) comments that, while purely local adaptation cannot account fully for the McCollough effect, one need not postulate the involvement of nonlocal mechanisms as specialized as edge detectors: simple "dipoles" or differences of RFs, of an unspecified main purpose, would do. The proposed method for object representation hints at a possible computational reason for the existence of such dipoles in the visual system

Amari S, Maruyama M (1987) A theory on the determination of 3D motion and 3D structure from features. Spatial Vision 2:151–168

Bishop PO, Coombs JS, Henry GH (1973) Receptive fields of simple cells in the cat striate cortex. J Physiol (Lond) 231:31–60

Blinn JF (1988), Models of light reflection for computer-synthesized pictures. In: Richards W (ed) Natural computation. MIT Press, Cambridge, Mass, pp 214–223

Bülthoff HH, Edelman S (1992) Psychophysical support for a 2-D view interpolation theory of object recognition. Proc Natl Acad Sci USA 89:60–64

Dawis S, Shapley R, Kaplan E, Tranchina D (1984) The receptive field organization of X-cells in the cat: spatiotemporal coupling and asymmetry. Vision Res 24:549–564

Edelman S (1992) Class similarity and viewpoint invariance in the recognition of 3D objects. CS-TR 92-17, Weizmann Institute of Science

Edelman S, Reisfeld D, Yeshurun Y (1992) Learning to recognize faces from examples. In Sandini G (ed) Proceedings of the 2nd European Conference on Computer Vision. (Lecture notes in computer science, vol 588) Springer, Berlin Heidelberg New York, pp 787–791

Harris CS (1980) Insight or out of sight? Two examples of perceptual plasticity in the human adult. In: Harris CS (ed) Visual Coding and Adaptability. Erlbaum, Hillsdale, NJ, pp 95–149

Harris CS, Gibson AR (1968) Is orientation-specific color adaptation in human vision due to edge detectors, afterimages, or "dipoles"? Science 162:1506–1507

Hubel DH, Wiesel TN (1959) Receptive fields of single neurons in the cat's striate cortex. J Physiol (Lond) 148:574–591

Hurlbert A, Poggio T (1988) Synthesizing a color algorithm from examples. Science 239:482–485

Intrator N (1992) Feature extraction using an unsupervised neural network. Neural Computation 4:98–107

Intrator N, Cooper LN (1992) Objective function formulation of the BCM theory of visual cortical plasticity: statistical connections, stability conditions. Neural Networks 5:3–17

Kanatani K (1990) Group-theoretical methods in image understanding. Springer, Berlin Heidelberg New York

Katz LC, Callaway EM (1992) Development of local circuits in mammalian visual cortex. Annu Rev Neurosci 15:31–56

Koenderink JJ, Doorn AJ van (1986) Optic flow. Vision Res 26: 161–180

Kuffler SW (1953) Discharge patterns and functional organization of mammalian retina. J Neurophysiol 16:37–68

Malach R, Amir Y, Bartfeld E, Grinvald A (1992) Biocytin injections, guided by optical imaging, reveal relationships between functional architecture and intrinsic connections in monkey visual cortex. Soc Neurosci Abstr 18:364

Mallot HA, Seelen W von, Giannakopoulos F (1990) Neural mapping and space-variant image processing. Neural Networks 3:16–25

McCollough C (1965) Color adaptation of edge detectors in the human visual system. Science 149:1115–1116

Moses Y, Edelman S, Ullman S (1993) Generalization across illumination and orientation changes in inverted and upright faces. CS-TR 93-14, Weizmann Institute of Science

Mundy JL, Zisserman A (eds) (1992) Geometric invariance in computer vision. MIT Press, Cambridge, Mass

Nishihara HK, Poggio T (1984) Stereo vision for robotics. In: Brady JM, Paul R (eds) Robotics research: the first international symposium. MIT Press, Cambridge, Mass, pp 489–505

Poggio T (1990) A theory of how the brain might work. Cold Spring Harb Symp Quant Biol 55:899–910

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

Poggio T, Girosi F (1990) Regularization algorithms for learning that are equivalent to multilayer networks. Science 247:978–982

Poggio T, Fahle M Edelman S (1992) Fast perceptual learning in visual hyperacuity. Science 256:1018–1021

Poggio T, Yang W, Torre V (1989) Optical flow: computational properties and net-works, biological and analog. In: Durbin R, Miall C, Mitchison G (eds) The computing neuron. Addison-Wesley, New York, pp 355–370

Polat U, Sagi D (1992) Lateral interactions between spatial filters: excitation and inhibition affected by spatial configuration. Perception 21 [Suppl 2]:92

Snippe HP, Koenderink JJ (1992) Discrimination thresholds for channel-coded systems. Biol Cybern 66:543–551

Spitzer H, Hochstein S (1988) Complex-cell receptive field models. Prog Neurobiol 31:285–309

Torrance KE, Sparrow EM (1966) Polarization, directional distribution, and off-specular peak phenomena in light reflected from roughened surfaces. J Opt Soc Am 56:916–925

Ullman S (1979) The interpretation of visual motion. MIT Press, Cambridge, Mass