Level 2

Early Vision

Carlo Tomasi, Stanford University, California, USA

CONTENTS

Early vision algorithms in humans and computers process images from the eye and from electronic video cameras respectively. They infer the shape, appearance, and motion of objects in the world. Conventionally, the lack of semantic interpretation distinguishes between 'early' and higher levels of vision.

INTRODUCTION

0071.001 The crystalline lens in the human eye focuses the entering light onto the array of receptors in the retina, forming an image of the world. This retinal image encodes the color and brightness of the light that surfaces in the world reflect from light sources into the eye. The retinal image changes over time as objects or light sources move relative to the observer. The human early vision system analyses these changing patterns of color and brightness to determine the position, shape, motion, and appearance of objects in the world. Conventionally, vision is said to be 'early' when it implies little or no semantic interpretation of the scene. Early vision therefore excludes higher cognitive aspects like object recognition or event interpretation.

0071.002 Computer vision systems make similar inferences from the images produced by electronic video cameras. The basic computational elements and the overall architecture of human and computer early vision systems differ greatly. However, the abstract nature of the computations they both perform does not depend on the mechanisms of their implementation in man or machine.

0071.003 The first step in vision is the formation of images, either in a camera or in the eye. Thereafter, images are analyzed and summarized in terms of edges, colors and textures, in order to provide a description of images that is more compact and depends to a lesser extent on changes of lighting or viewpoint. When changes of an image over time are considered, the motion of points in the field of view provides valuable information about the world. Image motion results from both observer motion and the movements and deformations of objects in the field of view. Its analysis allows distinguishing foreground from background, reconstructing the geometry of the three-dimensional world, and computing the motion of the observer within the environment. Additional sources of information about the world's geometry are: stereoscopic vision, which employs two cameras or eyes; the variations in the shading of visible surfaces; and the analysis of how edges meet one another in simple scenes. Although effortless to humans, early vision is the very difficult task of forming a stable representation of the world from the variable images seen by a moving observer.

IMAGE FORMATION

0071.004 In the 'pinhole camera' model of perspective projection, the rays of light passing through a point O in space intersect an image plane, which records the intensity and color of each ray. The point O represents the optical center of the eye or camera, and the image plane stands for the retina or the camera sensor. Let a point P on a visible surface in the world have coordinates (X, Y, Z) in a Cartesian reference system with origin at O and Z-axis orthogonal to the image plane. If the focal distance (the distance from O to the image plane) is f, the image p of P has coordinates $x = fX/Z$ and $y = fY/Z$. These coordinates are measured in a Cartesian image reference system, whose origin is the image point nearest to O.

0071.005 The pinhole camera model captures the essential property of image formation: each point on the image plane corresponds to a line in the world,

called the projection ray of that point. This model does not account for secondary properties of real lenses, such as their imperfections, their limited ability to form sharp images, or the dependency of image brightness on lens size.

0071.006 To model the finite number of sensing elements in real vision systems, the image coordinates x and y are discretized into integer pixel values within a finite range: $i = q(x/s)$ and $j = q(y/s)$ $(-n \leq i, j \leq n)$, where $q(a)$ is the integer nearest to a, s is the size of a sensing element, and n is a positive integer. The function q captures the discrete nature of images, and the bound n on image coordinates accounts for the finite field of view of a real vision system. This model does not account for possible overlap between adjacent sensing elements, or for sensors not on a square grid.

0071.007 In the human eye, two types of photoreceptors at pixel (i, j) encode the color and brightness of incoming light. The 'rods' are highly sensitive to all wavelengths in the visible spectrum, but cannot distinguish colors. The 'cones' are less sensitive, but exist in three types, responsive to different, (but overlapping) bands of the visible spectrum. In electronic color cameras, red, green, and blue filters are superimposed on three brightness sensors at each pixel. Black-and-white cameras have one sensor per pixel.

CONVOLUTION WITH LINEAR FILTERS

0071.008 Convolution is a ubiquitous operation in early vision. Intuitively, it amounts to additively blending the pixel values of small image neighborhoods to form a new pixel value. For instance, the blurring of a lens can be described as a weighted average of neighboring pixels in an ideal, sharp image. In this case, the blurred image is the convolution of the sharp image with an operator that averages pixel values together. At a somewhat higher level, an edge can be detected by comparing neighboring image pixels with an edge template. This is, again, a convolution, between the input image and the template.

0071.009 Another example of convolution is image smoothing. To reduce the effects of noise in images, it is often useful to replace each pixel in an image with a weighted average of the intensity values that surround it. This averaging operation can be described by saying that the image is convolved with an operator, or 'kernel', that contains the weights to use for the average. As a simple example, if we compute the average of a pixel and its eight immediate neighbors, the kernel is a 3-by-3 matrix all of whose elements are equal to 1/9. These nine numbers are multiplied by the nine pixels in question, and the products are added together to yield the output average value. This procedure is repeated everywhere in the image.

Formally, let $L(i, j)$ be a function of pixel coordinates (i, j), and let $I(i, j)$ be the output, say, of a rod or cone. The convolution J of I with L is defined as 0071.010

$$J(i,j) = \sum_a \sum_b I(i-a, j-b)L(a,b) \quad (1)$$

where the sums are performed over the domain of definition of the 'filter kernel' L. Thus, the output J at (i, j) is a linear combination of the values of the input I in some neighborhood of (i, j) and the values of the filter kernel L provide the combination coefficients. Similarly, single or triple summations appear in the definitions of convolutions of functions of one or three variables with filter kernels of one or three variables respectively.

Convolution with a bell-shaped kernel smoothes the input. A well-known example of a smoothing kernel is the isotropic Gaussian function 0071.011

$$G(i,j) = k\mathrm{e}^{-(i^2+j^2)/2} \quad (2)$$

and its one- and three-dimensional equivalents. 'Isotropic' here means that the function G is rotationally symmetric, that is, it has the same shape in all directions. For instance, an out-of-focus lens blurs in approximately the same way in all directions, and its output is often approximated by the convolution of an isotropic Gaussian with the input image.

Convolution with the derivative of G with respect to one of its arguments approximates the partial derivative of the input with respect to the corresponding variable. This operation is very useful for edge detection, described next. 0071.012

EDGE DETECTION

Edges are curves in the image across which image brightness or color changes abruptly. They are caused by shadow boundaries, contours of objects, or changes in surface orientation or reflectance. Standard algorithms for edge detection convolve the brightness $I(i, j)$ of the input image separately with the two partial derivatives G_i and G_j of the Gaussian kernel G to approximate the spatial gradient (I_x, I_y) of I. Some algorithms (Canny, 1986) then define edges to be ridges in the magnitude of the gradient. Others (Marr and Hildreth, 1980) convolve again I_x with G_i and I_y with G_j and add the resulting images together to obtain the Laplacian of 0071.013

image brightness. Edges are then zero crossings of the Laplacian.

0071.014 Noise in the sensing elements produces random fluctuations of perceived brightness, which can cause spurious edges to be detected. A single threshold on the gradient magnitude cannot be used to suppress these edges: if the threshold is too high, good edges are removed as well, and if it is too low, spurious edges persist. Some algorithms (Canny, 1986) require edges to contain some elements above a high threshold, but are then extended to all edge elements that are connected to the former and are above a lower threshold. Other algorithms link edge elements into curves, and preserve only those curves that are longer than a given length.

OPTICAL FLOW AND MOTION PERCEPTION

0071.015 As a point in the world moves relative to the observer, so does its image. When specified for every visible point, this image motion is called the motion field. The true motion field cannot be determined unambiguously from measurements in a very small image neighborhood. For example, if the neighborhood straddles a vertical edge and the edge moves in any direction, one only sees the horizontal component of motion. The edge may also be moving up or down, but this motion cannot be seen in the small aperture of this neighborhood. This inability to observe the motion field directly is called, somewhat awkwardly, the aperture problem.

0071.016 The *optical flow* is defined to be the smallest image motion that is consistent with local image measurements. In the example above, the optical flow along the edge is $(u, 0)$. This example shows that the optical flow is not always the same as the motion field: the latter is the true, projected motion; the former is the apparent motion.

0071.017 In computer vision, approximate motion fields are computed by combining values of the optical flow over several neighboring pixels, assumed to share the same motion field (Lucas and Kanade, 1981), or by imposing smoothness constraints on the field itself (Horn and Schunck, 1981). The human visual cortex computes the motion field by comparing the outputs of filters tuned to different orientations in space-time, that is, to different sets of spatial and temporal frequencies and directions of motion. Accuracies of 5% for velocities between 2 and 15 degrees per second have been reported for humans (McKee and Welch, 1985).

0071.018 If only the observer is moving, the motion field at as few as five points is sufficient in principle to compute both the translation and the rotation of the observer, except for an overall scale factor, as well as the distance of the five points in the world from the observer (Thompson, 1959). However, this computation is very sensitive to inaccuracies in the field measurements, and reliable results require more motion field values. This computation is called 'structure from motion'. Accuracies as good as one degree of visual angle for observer translation have been reported both in psychophysics and in computer vision, where algorithms have been proposed also for reconstruction from more than two images.

COLOR AND TEXTURE PERCEPTION

Color

0071.019 The three types of cone in the human eye are sensitive to three different bands in the visible spectrum of light, between 370 and 730 nanometres. Sensitivities peak at about 440, 540 and 560 nanometres for the three types, with much overlap in particular between the latter two bands. The distribution of the approximately 5 million cones in each eye varies from about 160 000 per square millimetre in the fovea to about 20 000 per square millimetre at the periphery, with the short-wavelength receptors being about 10 times sparser than those of the other types, consistently with the greater amount of blurring that the crystalline lens introduces at shorter wavelengths. The density in the fovea corresponds to a separation between cones of about half a minute of visual angle.

0071.020 The spectrum of light impinging on a set of cones is the product of the spectrum of the light source and the reflectance of the visible surfaces. Yet if the color of the light source is changed, humans perceive surface colors as if the change in the light source were only about half as much as the actual change. This ability of the human visual system to compensate for changes in the color of the light source is called color constancy. To achieve color constancy, the visual system must estimate the color of the light source at least approximately. Some theories (Land, 1986) propose that the visual system selects a color for the light source that corrects the average color perceived by all the cones to grey. Other theories (D'Zmura and Iverson, 1993) propose that the visual system detects specular reflections in the image, and takes them to reflect the color of the light source unchanged.

0071.021 Both psychophysical and physiological evidence suggest that colors are perceptually organized into pairs of 'opponent' colors in the visual cortex. This scheme for color encoding posits three mechanisms, each responding to a pair of sensations considered to be opposite to each other: light and dark, red and green, and blue and yellow.

Texture

0071.022 An image region has visual texture when the distribution of its brightness values exhibits periodicity, either deterministic or stochastic. For instance, tiles on a floor are deterministically periodic, and leaves on a tree are stochastically periodic. When a texture on a complex surface is projected to an image, the deformations caused by foreshortening and by the changing normal to the surface modulate the periodic components of the texture.

0071.023 Psychophysics and psychology indicate that humans can recognize materials and objects based on visual texture (texture classification), discriminate image regions with different textural properties (texture discrimination), and infer the shape and slant of a surface in the world from the modulation of its texture (shape from texture). Statistical representations (Caelli and Julesz, 1978) describe visual textures by the first- and second-order distributions of image brightness values in small regions of the image. Brightness histograms have been used in computer vision to capture the complete first-order empirical distribution. Summaries of the second-order distribution have included co-occurrence matrices, the fractal dimension of the spatial distribution of brightness values, and the conditional densities of an underlying Markov random field model.

0071.024 Structural representations describe visual textures as the repetition of a basic pictorial element, or *texton*, according to some placement rule. This rule can take the form of the description of a grid on which textons are arranged, or it can be a formal grammar, either deterministic or stochastic, that generates the grid points. For instance, the texton of a tiled floor could be a single tile, and the grammar is the description of a regular grid of squares.

0071.025 Current models of human texture analysis favor filter-based representations of texture (Bergen and Adelson, 1988). These are derived from the responses of a bank of linear filters tuned to different sizes and possibly orientations of the image intensity patterns. The integral of the magnitude of these responses over a small image neighborhood measures the energy contents of that neighborhood at the sizes and orientations that characterize the filters. A final stage of computation groups neighborhoods that have similar energy responses.

0071.026 If the texture on a surface in the world is uniform, the different distances and slants of different surface patches relative to the observer produce gradual variations of the corresponding image texture, as described by the equations of perspective projection. Shape from texture solves these equations for either distance or slant, and infers the three-dimensional shape of the surface.

REGION ANALYSIS AND SEGMENTATION

0071.027 The technique of segmentation partitions an image into regions such that different parts of the same region are similar to each other in some sense. For example, gray-level segmentation may require that the greatest difference between two pixels in the same region be less than a fixed threshold. A good segmentation would then have regions that are as large as possible given this constraint. Similarly, segmentation can be based on color or texture features. The results of segmentation differ from those of edge detection, mainly because edges are generally open curves while regions are bounded by closed contours.

0071.028 In computer vision, images are often segmented in the hope that the resulting regions belong to different objects, or to different parts of an object. This is, however, rarely the case, because the varied colors of objects, shadows, shading, and variations in lighting produce large variations within objects , while at the same time similar colors in adjacent objects are not uncommon. Even so, describing an image as a collection of disjoint regions can offer advantages for later stages of processing, at least in terms of computational complexity.

0071.029 Several segmentation methods are based on repeated splitting or merging, and sometimes on a combination of both. In a splitting method, for instance, the whole image may initially be considered as a single region, to be split, say, in half if the region as a whole is not homogeneous enough. The same procedure is then applied recursively to the resulting regions, until all regions are sufficiently homogeneous. Merging methods proceed in the other direction, starting with each pixel being considered as its own region, to be merged with neighbors as long as the resulting region is homogeneous.

0071.030 'Stochastic relaxation' has also been used for segmentation. The class of images of interest is modelled as a Markov random field, which specifies the probability that a pixel has a certain value given the values of its neighbors. Given a particular image, relaxation iteratively adjusts pixel values to maximize the likelihood of the image having being drawn from the class in question (Geman and Geman, 1984). For segmentation, the underlying Markov random field assigns highest probability to noisy piecewise-constant images. As a result, relaxation turns the input image into a piecewise-constant one, whose discontinuities are the segmentation boundaries.

BINOCULAR STEREOPSIS

0071.031 'Binocular stereopsis' computes distances (called 'depths') to the visible surfaces by comparing the images of the world seen by two eyes, in humans, or by two cameras, in computer vision. This computation assumes knowledge of the relative position and orientation of the eyes or cameras. First, a stereoptic correspondence module finds pairs of points in the two images that are projections of the same point in the world. Then, the depth for each pair is computed by triangulation.

Stereoptic Correspondence

0071.032 Two matching points in the two images of a stereoptic pair are likely to look similar to each other. However, this is not always the case. For instance, a point on a glossy surface may have the color of the surface when viewed from one eye, but the color of the light source in the other eye because of a specular reflection. Furthermore similar points do not necessarily match. On a blank wall, for instance, many points look the same. Thus, similarity of appearance, the main criterion for matching two points, is neither necessary nor sufficient for a match. Correspondence is also complicated by the fact that points visible in one image may not be visible in the other because of an intervening occluding object, so that not every image point need have a match.

0071.033 Yet the human visual system can establish correspondences effortlessly and on the basis of minimal information. Random-dot stereograms like the one in figure 1 serve to demonstrate this ability, and show that no prior recognition is necessary for stereoptic correspondences to be established (Julesz, 1960).

0071.034 Knowledge of the relative position and orientation of the two eyes, or cameras, restricts matches for any given point in one image to be on a known line in the other. This is because the two optical centres and any one point P in the world define a plane, the so-called epipolar plane of P. This plane intersects the two image planes at two lines called the epipolar lines, which pass through the two images of P. Hence the 'epipolar constraint': the match for any point on one epipolar line must be on the corresponding epipolar line. The angle formed by the projection rays of two matching points is called the disparity.

0071.035 One way to establish correspondences is to compare small image patches along corresponding epipolar lines in the two images. Sums of squared differences can quantify the comparison between a small image patch P_L centered at (i_L, j_L) in the left image I_L and a patch P_R centered at (i_R, j_R) in the right image I_R:

$$s = \sum_a \sum_b [I_L(i_L + a,\ j_L + b) - I_R(i_R + a, j_R + b)]^2 \quad (3)$$

where the summation indices range over the patches.

0071.036 Even for image pairs as ambiguous as those in figure 1, and even with imperfect image measurements, a small value of s indicates a likely match, as long as the patches being compared are not too small. When these local comparisons fail to determine matches unambiguously, more global criteria must be invoked. For instance, matches may be required to correspond to smooth, or at least continuous, surfaces. In computer vision, these more global requirements have been enforced by the use of various techniques, including stochastic relaxation, dynamic programming, and network flow optimization methods.

Triangulation

0071.037 The depth of a point in the world is easily computed from a pair of corresponding points. Let α be the angle that the left projection ray of point P forms with the optical center of the left camera. Let δ be the disparity, and let the baseline, that is, the distance between the two optical centers, be b. Then, a simple geometric construction shows that the distance between the left optical centre and point P is

$$\rho = b(\cos\alpha + \sin\alpha \cot\delta) \quad (4)$$

SHAPE FROM SHADING

0071.038 The perceived brightness of a surface varies with the orientation of the surface relative to illumination and viewing directions, among other factors. As a consequence, the shading on a uniformly colored surface conveys some information about the shape of the surface itself. The locations of concavities and convexities are examples of qualitative information that can be gathered in this way, but the shape of the surface can be reconstructed even quantitatively if its reflectance map is known.

0071.039 The reflectance map of a surface expresses surface brightness as a function of surface orientation for a particular distribution of the light sources. If the depth of a surface is represented as a function $z(x,y)$ of the image coordinates x and y, the two partial derivatives p and q of z encode the orientation of the surface, since the surface is orthogonal to the vector $(-p, -q, 1)$ everywhere. Therefore, the reflectance map can be expressed as a function $R(p,q)$ that gives the brightness of a surface patch with normal proportional to $(-p, -q, 1)$ in viewer coordinates. If the image at (x, y) has brightness $I(x,y)$, one thus obtains the equation

$$R(p, q) = I(x, y) \tag{5}$$

0071.040 The two functions $R(p, q)$ and $I(x, y)$ are known (the former by assumption, the latter by measurement), and the unknown depth $z(x, y)$ appears through its partial derivatives p and q. The equation above is therefore a partial differential equation, which can be solved for z by numerical means.

0071.041 Because the reflectance map combines information about light and surface, it is often hard to determine *a priori*. However, it is conceivable that one could learn approximate reflectance maps for surfaces of various materials and, say, known position of the sun in the sky.

LINE LABELLING IN POLYHEDRAL SCENE ANALYSIS

0071.042 The remarkable human ability to understand line drawings inspired early computer vision researchers to separate the image interpretation task into two stages. In the first stage, edge detection transforms an image into a line drawing, which the second stage then interprets.

0071.043 In a world of polyhedral objects with no surface markings, a line segment in the drawing can only represent the convex or concave edge between two visible faces, or an edge that occludes some surface at a greater depth. Two or more line segments meet at junctions corresponding to vertices in the world. If the number of lines meeting at a junction is bounded (typically by a maximum of three), junctions can be classified into a finite taxonomy depending on the number and angles of the meeting lines. For instance, two lines form a V junction. In Y junctions, three lines meet at angles that are all smaller than 180 degrees in the image; and in W junctions one of the three angles exceeds 180 degrees.

0071.044 A line drawing can then be interpreted by assigning labels – convex, concave, or occluding – to all line segments, making sure that no impossible junctions result. For instance, a concave edge cannot meet a convex edge at a V junction; and three occluding edges cannot meet at a W or Y junction. Remarkably, these rules usually restrict the interpretation of a line drawing to one or very few possibilities, assuming there are no accidental alignments of features. Huffman (1971) developed an elegant algorithm for finding a possible labelling of a line drawing, and started a very active area of investigation.

0071.045 However, attempts to extend these results to complex images failed for several fundamental reasons. Firstly, objects in the world are not always polyhedral. Secondly, an edge detector also detects surface markings, shadow contours, and other curves, for which simple consistency rules cannot be given. Thirdly, edges computed from images are usually broken, and junctions are hard to pinpoint. Because of these difficulties, this line of research has been abandoned.

SIZE AND POSITION INVARIANCE

0071.046 As an objects moves relative to the viewer, the size and position of its projection on the image change. However, objects are recognized in spite of these changes, and also in spite of changes of illumination or viewing angle. The methods for three-dimensional reconstruction described in the previous sections form one basis for explaining this invariance of recognition performance under changing stimuli. According to this explanation, the visual system would compute invariant representations of the image, such as three-dimensional object models in world coordinates.

0071.047 An alternative explanation seems to be more consistent with physiological evidence, and posits instead the existence of mechanisms that compensate for variations in size, position, and other factors (Ito *et al.*, 1995). Almost half of the neurons in the anterior part of monkey inferotemporal cortex seem to respond to the same object in the

field of view under large changes in object size and position, although different cells respond within different ranges of variation. On the other hand, other neurons in the same area respond only when the object is presented in a narrow range of sizes and positions. Invariance in the former type of neuron may be achieved by convergence of multiple cells of the latter type.

0071.048 These two explanations of perceptual invariance lead to entirely different theories of how the world is represented in the visual system. The proposal based on compensation is more consistent with pictorial representations of objects, in which images are transformed and aligned to achieve constancy, than the miniature world of the reconstruction-based approach. But the evidence is still insufficient to allow us to conclude definitely either way.

CONCLUSION

0071.049 Some of the problems of early vision are problems of image processing or statistical estimation. For instance, edge detection amounts to template matching in the presence of noise; and geometric reconstruction estimates three-dimensional quantities from noisy measurements of image motion. However, vision is inherently a process of inference, and early vision is no exception: several assumptions must be made in order to reconstruct aspects of the three-dimensional world from its two-dimensional projections on eye or camera, and finding assumptions that are adequate in most situations is still an challenge to computer vision researchers. Vision is a form of cognition, and the study of early vision may be one of the best approaches towards understanding intelligence.

References

Bergen JR and Adelson EH (1988) Early vision and texture perception. *Nature* **333**: 363–364.

Caelli T and Julesz B (1978) On perceptual analyzers underlying visual texture discrimination: Part I. *Biological Cybernetics* **28**(3): 167–175.

Canny J (1986) A computational approach to edge detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **8**(6): 679–698.

D'Zmura M and Iverson G (1993) Color constancy. I. Basic theory of two-stage linear recovery of spectral descriptions for lights and surfaces. *Journal of the Optical Society of America (A)* **10**: 2148–2165.

Geman S and Geman D (1984) Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **6**(6): 721–741.

Horn BKP and Schunck BG (1981) Determining optical flow. *Artificial Intelligence* **17**: 185–203.

Huffman DA (1971) Impossible objects as nonsense sentences. *Machine Intelligence* **6**: 295–323.

Ito M, Tamura H, Fujita I and Tanaka K (1995) Size and position invariance of neuronal responses in monkey inferotemporal cortex. *Journal of Neurophysiology* **73**: 218–226.

Julesz B (1960) Binocular depth perception of computer-generated patterns. *Bell System Technical Journal* **39**: 1125–1162.

Land EH (1986) Recent advances in retinex theory. *Vision Research* **26**: 7–22.

Lucas BD and Kanade T (1981) An iterative image registration technique with an application to stereo vision. *Proceedings of the Seventh International Joint Conference on Artificial Intelligence*, pp. 674–679.

Marr D and Hildreth E (1980) Theory of edge detection. *Proceedings of the Royal Society of London (B)* **207**: 187–217.

McKee SP and Welch L (1985) Sequential recruitment in the discrimination of velocity. *Journal of the Optical Society of America (A)* **2**: 243–251.

Thompson EH (1959) A rational algebraic formulation of the problem of relative orientation. *Photogrammetric Record* **3**(14): 152–159.

Further Reading

Gibson JJ (1950) *The Perception of the Visual World*. Boston, MA: Houghton Mifflin.

Hering E (1878/1920) *Handbuch der gesammter Augenheilkunde*, part 1, chap. XII. Berlin: Springer. [Originally published as: *Grundzüge der Lehre vom Lichtsinn*.]

Horn BKP (1986) *Robot Vision*. Cambridge, MA: MIT Press.

Kanatani K (1993) *Geometric Computation for Machine Vision*. Oxford: Clarendon.

Longuet-Higgins HC (1981) A computer algorithm for reconstructing a scene from two projections. *Nature* **293**: 133–135.

Marr D and Poggio T (1976) Cooperative computation of stereo disparity. *Science* **194**: 283–287.

Pollard SB, Mayhew JEW and Frisby JP (1985) PMF: a stereo correspondence algorithm using a disparity gradient constraint. *Perception* **14**: 449–470.

Svaetichin G (1956) Spectral response curves from single cones. *Acta Physiologica Scandinavica* **134**: 17–46.

Zucker SW (1976) Toward a model of texture. *Computer Graphics and Image Processing* **5**: 190–202.

Glossary

Field of view The cone of directions that are visible through the lens of a camera, eye, or other optical system.

Fovea A small area, near the center of the retina, which is packed with cones and affords acute vision.

Gradient A vector that points along the direction of steepest ascent of a function, and measures the rate of change along that direction.

Photoreceptor One of the receptors for light stimuli in the eye.

Pinhole camera A camera made of a dark box with a small hole at the centre of one side and a screen on the opposite side.

Pixel Any of the small, usually rectangular, discrete elements that together constitute an image.

Texton A figural element which, when periodically repeated on a surface, constitutes a visual texture.

Visual cortex A layer of gray matter in the occipital lobe of the primate brain, responsible for integrating visual information from the eyes.

Keywords: (Check)

eye; camera; perception; sensing; computer vision

0071f001 **Figure 1.** The images in this random-dot stereogram are identical, except that a central square in the right image has been shifted slightly to the right, and the resulting gap filled with random dots. When viewed with eyes crossed so as to fuse the two images into one, a square floating in front of a plane should appear after a few seconds

1. References, (Lucas and Kanade, 1981): Please supply names of editors, place of publication, and publisher.