Three processing characteristics of visual texture segmentation

TERRY CAELLI

Psychology Department, University of Alberta, Edmonton, Alberta T6G 2E9, Canada

Received 9 October 1984; accepted in revised form 22 March 1985

Abstract—Three mechanisms are outlined which are sufficient to determine texture segmentation or discrimination. They are: (1) convolution of detector profiles with the input image; (2) impletion, where the perceptual 'filling in' of the input surface occurs via a nonlinear filtering operation on each detector's output (3) grouping, where areas are segregated according to their differences in detector responses after impletion occurs. These mechanisms are compared with those proposed to occur in human visual texture discrimination.

INTRODUCTION

It is usually presumed that one fundamental facet of visual perception is that the retinal image is segregated according to differences in the characteristics of local luminance profiles. For the simple case where the image is conceived of as a real-valued intensity distribution function \( f(x, y) \) over the locally Euclidean (retinal) coordinate system \((x, y)\), such a segregation process reduces to that of grouping \([x, y]\) according to a subjective vector-valued function \(X(x, y)\) depicting the multidimensional perceptual response associated with each point. In this report \(X\) is envisaged to factor \([x, y]\) into equivalence classes corresponding to texture regions. That is, \(X\) would define the outputs of \(n\) detectors at each position, for example, as a result of convolution between \(f(x, y)\) and the detector profiles.

The lower bound for the number \((n\) above) of visual detectors operating in threshold detection or superthreshold discrimination tasks is still debated with respect to orientation, spatial frequency, phase and bandwidth parameters of the signal (for example, see Wilson, McFarlane and Phillips, 1983). However, upper bounds have been determined in the context of texture and natural image processing. Strong evidence has been found (Caelli, 1982; Caelli and Hübner, 1983) for an inability of observers to resolve orientation and spatial frequency into units narrower than \(\pm 5^\circ\) and \(\pm 1/8\) octave wide detector regions. These narrow-band filter units are consistent with Gabor detector profiles whose spatial frequency is such that one cycle covers a \(e^{-1}\) decay width of the modulating Gaussian profile (Watson, Barlow and Robson, 1983; Daugman, 1983), and are consistent with known receptive profiles in the visual cortex (see Pollen and Ronner, 1983).

However, the texture segregation problem is not completely solved by \(a\ priori\) knowledge of the elements which compose the detector response vector \(X(x, y)\). In fact, earlier proposals for texture discrimination by Julesz (1962) did not involve localized detectors. Rather, the (probability) densities of \(n\)th-order statistics (points, dipoles, third-order, etc.) were proposed to be perceptually computed via integration over restricted texture regions. Segregation occurred only if the statistical differences were large enough. For example, dipole (second-order) orientation and length distribution
differences yield texture discrimination as a function of the number of dipoles involved (Caelli and Julesz, 1979).

In a series of recent investigations it has become clear that texture discrimination cannot be totally accounted for in terms of such statistical differences. From previous results (Caelli and Julesz, 1978; Julesz and Bergen, 1983) specific local luminance profile differences are critical in determining discrimination and it is these local profile differences, averaged over textured regions, which determine segregation. These local profiles are termed 'textons' by Julesz (1981) and candidates such as lines, elongated 'blobs' (Gabor signals), crossings and line endings have been isolated as perceptually discriminating sources [possible bases for $X(x, y)$].

Though these results have initiated new investigations into the types of texture coding elements, they do not address the problem of how the segmentation process occurs, particularly when the textures are composed of spatially separated elements. The proposed solution involved three (including the initial detector code) operations which can be reflected in a dynamically parallel processing network and enables a quantitative state/space representation for each component. In order, these are:

1. Nonlinear convolution (NLC), where each detector's response is a logarithmic function of the convolution between the detector's profile and the input image.
2. Perceptual impletion where the 'filling in' of the textured surface is determined by the nonlinear (sigmoidal) enhancement of strong response and inhibition of weak response regions for each detector (a form of relaxation, Hummel and Zucker, 1983).
3. Grouping, where detector outputs between different positions in the image are compared to result in perceptual grouping. Here the image is segmented according to the correlation between detector outputs over positions within textured regions. This is accomplished by a local clustering process rather than by factoring the detector-by-position response matrix, since the latter seems biologically unlikely and contains a great deal of redundancy.

**THE MODEL**

**Level I activity: initial detector responses (NLC)**

The relationship between a given detector's response and an input signal, which seems most parsimonious with both neurophysiological and psychophysical results, is that of **nonlinear convolution**. For the input signal $f(x, y)$ and a detector $i$ centred at $(x, y)$: $D_i(a - x, b - y)$, the steady-state (spatial) response ($R_i$) is defined by:

$$R_i(x, y) = \phi \left[ \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(a, b)D_i(x - a, y - b)\,da\,db \right]$$

where $\phi$ represents the nonlinear response function [see equation (5)].

Although it is highly unlikely that the dimensionality of $D_i$ is synonymous with the total number of cells exhibiting receptive field profiles in the visual cortex, it is equally not clear as to the lower bound on the number and profile shapes of detectors actually utilized, or functional, in a given visual task. One further problem with the usual representations for receptive field profiles (for example, the Gabor signal) is that they are not independent in the sense that they have zero covariance (or dot product). This lack of orthogonality creates a problem when estimating the types of spatially distinct information that different detectors may process in an image. For these reasons we have