Texture-Based Processing in Early Vision and a Proposed Role for Coarse-Scale Segmentation

Michael Hucka [a] and *Stephen Kaplan* [a,b]
[a]Department of Electrical Engineering and Computer Science
[b]Department of Psychology
The University of Michigan, Ann Arbor, MI 48109
hucka@umich.edu,skap@umich.edu

Abstract

Humans and other natural systems are remarkably adept at extracting spatial information from vision. To better understand this process, it would be useful to know how the visual system can make an initial estimate of where things are in a scene and how they are oriented. Texture is one source of information that the visual system can use for this purpose. It can be used both for segmenting the visual input and for estimating spatial orientations within segmented regions; moreover, each of these two processes can be performed starting with the same mechanisms, namely spatiotemporally-tuned cells in the visual cortex. But little attention has been given to the problem of integrating the two processes into a single system. In this paper, we discuss texture-based visual processing and review recent work in computer vision that offers insights into how a visual system could solve this problem. We then argue that a beneficial extension to these approaches would be to incorporate an initial coarse-scale segmentation step. We offer supporting evidence from psychophysics that the human visual system does in fact perform such a rough segmentation early in vision.

1 Introduction

Imagine yourself running through rough terrain, perhaps fleeing a predator, or perhaps chasing after prey. Your visual system does not have time to scrutinize the countless trees, rocks, and other objects you pass by. What you need most is enough spatial information to avoid obstacles, to orient yourself, to pick a path. In this situation, even a rough sketch of the spatial layout of the environment can provide crucial information.

Extracting spatial layout is a fundamental capability of biological vision systems. In situations where time and information are limited, an initial, approximate description of the layout of surfaces in the environment can be highly useful; it can equally serve as a starting point for more detailed spatial and object analysis in less constrained situations. But how does the visual system extract layout? What kinds of visual information and associated neural information-processing mechanisms are useful for this purpose?

Texture is one well-known source of information that the visual system can use (Gibson, 1950). Motion, stereopsis, and other sources of information are powerful cues for spatial perception and important when available. However, patterns of texture can be obtained even from brief glimpses, during which a scene will appear static and motion cues are unavailable, and at distances and visual angles at which the effectiveness of stereopsis is limited. Texture can be used for several purposes, notably segmenting the input and estimating spatial properties such as surface orientation. But the question of how these processes can be integrated into a single framework has received little attention.

In this paper, we discuss texture-based visual processing and review research from computer vision that offers insights into how a visual system could solve this integration problem. Drawing on evidence from the human visual system, we then argue that an initial coarse-scale segmentation step would be a beneficial extension to these approaches.

2 Variations in Texture as Useful Sources of Information

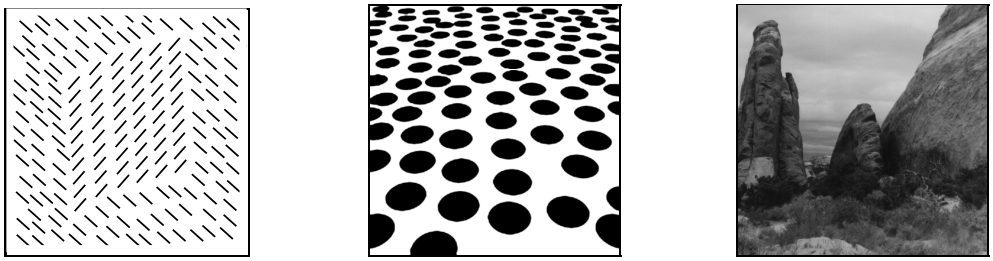
The features that comprise the "grains" of a visible texture are small, locally homogeneous regions surrounded by variations in dimensions such as luminance, orientation, or color. It is inherently a multi-scale phenomenon: texture arises over regions that are much larger than the basic repetitive pattern of elements composing it. Further, regional pattern structure is more important than individual, local details. Indeed, there is evidence strongly suggesting that the visual system does not pay much attention to the shapes of individual texture elements (Todd and Akerstrom, 1987; Sakai and Finkel, 1995). Rather than analyzing textures on the basis of element shapes (Julesz, 1986)—something difficult at best for irregular natural textures—it has proven effective to use spatial-frequency content in local neighborhoods (Graham, 1991; Sagi, 1991; Sakai and Finkel, 1995).

The Role of Cortical Cells. The simple and complex cells (Hubel and Wiesel, 1959) of the primary visual cortex of the brain have interesting properties in this regard. Cortical simple cells have receptive fields divided into distinct excitatory and inhibitory subregions, within which light is summed approximately linearly. These neurons can be modeled as linear spatiotemporal filters combined with additional mechanisms, such as rectification, that produce the final response (Heeger, 1992). Conversely, complex cells do not have distinct subregions within their receptive field; they can be modeled as pooling the outputs of multiple, overlapping, spatiotemporal filters that have properties similar to those of simple cells (Heeger, 1992).

A complex cell is not particularly sensitive to the position of a stimulus within its receptive field (De Valois and De Valois, 1990). Its response can be thought of as representing the *amount* of a particular type of spatial structure, such as texture, within its tuning range and its region of spatial pooling. These properties suggest that the role played by complex cells in spatial vision is unlikely to be contour processing, for which the ability to localize line and edge segments is important. Instead, a complex cell may be better suited to signaling the presence of *regional* repetitive variations in luminance and other stimulus dimensions. Such a response would be a useful starting point for texture analysis (De Valois and De Valois, 1990).

Segmentation Based on Texture. Spatial-frequency-sensitive mechanisms resembling complex cells are in fact commonly used in models of texture-based segmentation, both in psychology and computer vision. (For some reviews, see Sagi, 1991.) A commonly-used framework consists of first computing the responses of linear spatiotemporal filters at different spatial frequencies and orientations at every location of an input. These responses are combined non-linearly so as to simulate the processing performed by cortical complex cells. After normalizing the responses, segmentation can be performed on the resulting set of values either by searching for rapid changes (which are taken to indicate boundaries between different texture regions), or by clustering regions with similar values.

These models have shown that it is possible to segment a visual input based on textural information and spatial filtering mechanisms. But to date, nearly all models of texture segmentation have been designed and tested on inputs consisting of patches of textures viewed frontally (e.g., Figure 1a). Unfortunately, the natural world rarely consists of flat surfaces oriented frontally to the viewer. Components of real scenes often contain texture variations *within* regions, caused by the projections of curved or slanted surfaces. This is a problem for most models of texture-based segmentation; they are designed with the assumption that separate regions are each characterized by *homogeneous* texture. Consequently, they can be mislead by the systematic distortions of texture within regions that characterize slanted or curved surfaces (Krumm, 1993).



(a) (b) (c)

Figure 1: (a) Example of the type of stimulus often used with models of texture-based segmentation. (b) A slanted surface covered with dots, illustrating gradients of compression, foreshortening, scaling and density. (c) A natural scene, composed of many surfaces with different orientations.

Estimation of Surface Orientation Based on Texture. Variations in texture within regions of a natural scene can convey useful information. The brightness patterns on a surface become distorted when projected in perspective onto the retinas of the eyes. This distortion is a function of the shape of the surface, its distance from the observer, and its orientation relative to the observer. The spatial-frequency content of each small patch in the image is systematically related to texture distortion (Figure 1b). For example, one effect is *compression*; it leads to flattening of texture elements in the image, which in turn leads to changes in the dominant spatial frequency and orientation content. Element sizes are reduced by the flattening effect, leading to the presence of higher spatial frequencies in regions containing smaller visible elements. Element orientations tend towards a direction perpendicular to the direction of the texture gradient as the elements are increasingly flattened; thus, less compressed regions contain a greater range of orientations, whereas more compressed regions tend toward one predominant orientation. Other texture distortion effects also occur.

The relationships underlying these effects are fundamental and available to any system capable of perceiving them. Researchers in computer vision have explored the precise nature of the geometrical relationships between a textured surface and the resulting projective distortions (Gårding, 1995; Krumm, 1993; Malik and Rosenholtz, 1994). Gårding (1995) and Malik and Rosenholtz (1994) together recently developed a particularly rigorous and elegant mathematical formulation. They used the tools of differential geometry to express the relationships between different types of texture distortions in the image and the shape and orientation parameters of the physical surface. They modeled texture distortion as an affine (i.e., linear) transformation between neighboring image patches. They also demonstrated the approach by implementing a computer vision system that uses differences in the spatial-frequency content of neighboring image patches to estimate texture distortion.

A limitation of most current models is that they are designed with the assumption that the input contains a single surface. This is the case for most psychophysical (e.g., Todd and Akerstrom, 1987), neurobiological (e.g., Sakai and Finkel, 1995), and computational (e.g., Malik and Rosenholtz, 1994) approaches. Realistic scenes, however, contain multiple textured regions, and most current models cannot cope with such inputs because they assume that texture inhomogeneities are related only to changes in the shape of a single surface. A few researchers have attempted to address this limitation. A recent example is work by Black and Rosenholtz (1995), who extended Malik and Rosenholtz's (1994) approach to handle inputs containing multiple textured regions. However, their system is still incomplete; it is not designed to segregate the different regions in an input, and further, relies on an external agency to tell it the number of surfaces expected in the image.

3 The Challenge of Combining Both Processes in a Common Framework

Most surfaces in the natural world are visibly textured, and so a viable vision system must be capable of segmenting inputs containing textured regions. Moreover, given that most surfaces in the world are neither flat nor viewed frontally, it is also necessary for the system to be capable of coping with textures that are systematically distorted due to 3-D effects. At the same time, for a visual system attempting to estimate the spatial layout of surfaces in a scene, it would be advantageous to exploit texture distortions as a way of estimating surface orientations, rather than throwing away the information contained in the distortions.

Indeed, recent psychophysical work in humans by Kingdom and Keeble (1996) supports the view that common neural mechanisms underlie the initial stages of texture-based segmentation and texture-based estimation of surface properties. Kingdom and Keeble used stimuli composed of a dense array of small, oriented elements (specifically, Gabor patches). The arrays contained either abrupt variations in element orientations, or smooth variations in element orientations across regions. Kingdom and Keeble examined their subjects' ability to detect the patterns in brief presentations. They were able to account for their subjects' performance using a common model for both types of stimuli. This suggests that the detection of abrupt variations in textures (useful in texture-based segmentation) and smooth variations in textures (useful for shape analysis) is subserved by a common visual mechanism. This mechanism appears to act on the outputs of spatiotemporal filters in the early visual cortex.

Thus it is likely that, in the human visual system, the initial stages of texture-based segmentation and shape estimation are combined in a common framework that employs spatial filtering. But *how* does the visual system integrate these two processes?

This question has seen little attention in psychophysical or neurobiological research on vision. As discussed above, most current models of texture-based segmentation assume that textured regions within the visual input are free of systematic gradients that are not part of the boundaries between texture regions. Conversely, most models of texture-based shape estimation assume inputs consisting of only a single surface. These are incompatible assumptions. To integrate both processes into a common architecture requires a way of segmenting inputs containing possibly distorted textures, as well as a way of performing shape estimation within multiple visual regions.

A clue to how this problem may be solved comes from recent work in computer vision. Krumm (1993; Krumm and Shafer, 1994) addressed the problem directly and developed a system designed to perform texture-based segmentation by explicitly accounting for surface orientation from texture distortion. Krumm's approach to extracting surface orientation is based on geometrical principles similar to those of Malik and Rosenholtz (1994). His system uses spatial-frequency changes between local patches of an image to estimate surface slant in each neighborhood. This is used to undo the effects of surface slant, and to estimate what the spatial-frequency content of an image patch would be if the corresponding surface patch were viewed frontally. The system then uses bottom-up clustering to segment the input on the basis of these "frontalized" image patches, agglomerating patches that appear to lie on the same surface.

In the absence of evidence about how biological vision systems actually accomplish this task, we theorize that it is plausible that the brain follows the *general idea* behind Krumm's approach: namely, to perform texture-based segmentation on inputs containing multiple regions of distorted textures, the visual system must compensate for textural distortions *during* the segmentation process. As Krumm has pointed out, a segmentation process that does not take into account texture distortions will be mislead by 3-D effects in realistic visual inputs.

Krumm's system does have certain limitations. One is that the system is not able to determine automatically the final number of regions in the input during segmentation; a human must chose

the final, overall segmentation from a set of candidates that differ in the number of regions found. Krumm acknowledges this limitation and suggests it could be resolved within the existing bottom-up processing organization.

We propose that the human visual system solves this problem by using a multi-level approach rather than a strictly bottom-up segmentation. As we discuss in the next section, there is evidence that the visual system extracts coarse global organization before finer-scale visual structure. Such a rough segmentation could be used to derive an estimate of the number of regions in the input; it could also guide processing at finer spatial scales.

4 The Influence of the Global Pattern

Seminal work by Navon (1977) suggests that, all other things being equal, the visual system processes whole pattern information before information about pattern substructures. Navon used hierarchically structured stimuli which he flashed for short durations to human subjects. He examined the relative speed of processing and interference effects when subjects were directed to attend either to the global structure or the local structure of the stimuli. Subjects' reaction times turned out to be faster for the global pattern, and further, Navon found asymmetric interference: when the global pattern conflicted with the local one, subjects' identification speed for the local pattern was slowed. Navon termed these results "global precedence."

Hughes et al. (1984) provided evidence that global precedence involves the early visual system rather than being a post-perceptual effect such as decision processing. They further suggested that the spatiotemporal filters in early vision play a role in the phenomenon. A number of other results also support the idea that spatial-frequency selective mechanisms are involved (e.g., Shulman et al., 1986). Compelling experiments were performed by Hughes et al. (1990), who investigated whether a global advantage occurs when using stimuli that are nearly devoid of low spatial-frequency content. Hughes et al. performed experiments in which they presented subjects with hierarchically-structured geometric patterns using a procedure similar to Navon's (1977). They found that patterns that normally yielded global precedence, when filtered to remove low spatial frequencies, failed to yield a global advantage and instead led to local precedence. In other words, without coarse-grained spatial information in the visual input, the global pattern no longer seemed to dominate subjects' perceptions. Hughes et al. therefore suggested that under certain conditions, "global spatial structure might be directly encoded by low-frequency mechanisms rather than by hierarchical feature processing" (Hughes et al., 1990).

Different experiments carried out by Schyns and Oliva (1994) also argue for a predominantly coarse-to-fine process in vision. Schyns and Oliva constructed hybrid images composed of the low spatial-frequency content of one scene superimposed on the high spatial-frequency content of another scene. They used these hybrids in a matching task, in which they flashed the hybrid images for short durations to human subjects and examined the relationship between exposure duration and type of image. The results implied that at short exposure durations, the low-frequency components (the coarse spatial structure) of the hybrid images had a greater impact on perceptual processing.

These sets of results argue not just for the dominance of rough, overall spatial structure in initial visual processing; they specifically implicate low spatial-frequency content. It is interesting to note that these findings parallel another set of results: humans can react to low spatial-frequency stimuli more rapidly than high spatial-frequency stimuli (Parker and Dutch, 1987).

5 Summary

Global precedence and related phenomena support the view that visual processing in humans usually proceeds in a multiscale, coarse-to-fine manner, with the visual system extracting a coarse de-

scription of the arrangement of major visible elements before finer details. The interaction between global and local structure found in global precedence indicates that the global structure can influence the visual processing at finer spatial scales. At low spatial frequencies, texture will be blurred out and the pattern of overall luminance contrast will dominate. This coarse-scale segmentation could serve to guide processes operating in parallel at increasingly finer scales, where texture distortions resulting from slanted or curved surfaces become important and the segmentation process cannot rely on textural homogeneity to define a region. The coarse-scale segmentation could also be used to estimate the number of different surfaces in the input, which would make it a useful extension to computer vision systems such as those of Krumm (1993) and Black and Rosenholtz (1995). All of these processes can take place in a common framework that begins with the responses of orientation- and spatial-frequency-sensitive complex cells in the early visual cortex.

References

Black, M. J. & Rosenholtz, R. (1995). Robust estimation of multiple surface shapes from occluded textures. *IEEE International Symposium on Computer Vision, Miami, Florida*.

De Valois, R. L. & De Valois, K. K. (1990). *Spatial Vision*. New York: Oxford University Press.

Gårding, J. (1995). Surface orientation and curvature from differential texture distortion. *Proceedings of the Fifth International Conference on Computer Vision*.

Gibson, J. J. (1950). The perception of visual surfaces. *The American Journal of Psychology*, 58(3), 367–384.

Graham, N. (1991). Complex channels, early local nonlinearities, and normalization in texture segregation. In M. S. Landy & J. A. Movshon (eds), *Computational Models of Visual Processing*.

Heeger, D. J. (1992). Normalization of cell responses in cat striate cortex. *Visual Neuroscience*, 9, 181–197.

Hubel, D. H. and Wiesel, T. N. (1959). Receptive fields of single neurons in the cat's striate cortex. *Journal of Physiology*, 148, 574–591.

Hughes, H. C., Fendrich, R. and Reuter-Lorenz, P. A. (1990). Global versus local processing in the absence of low spatial frequencies. *Journal of Cognitive Neuroscience*, 2(3), 272–282.

Hughes, H. C., Layton, W. M., Baird, J. C. and Lester, L. S. (1984). Global precedence in visual pattern recognition. *Perception & Psychophysics*, 35(4), 361–371.

Julesz, B. (1986). Texton gradients: The texton theory revisited. *Biological Cybernetics*, 54, 254–251.

Kingdom, F. A. A. and Keeble, D. R. T. (1996). A linear systems approach to the detetion of both abrupt and smooth spatial variations in orientation-defined textures. *Vision Research*, 36(3), 409–420.

Krumm, J. (1993). *Space Frequency Shape Inference and Segmentation of 3D Surfaces*, PhD thesis, Robotics Institute, Carnegie Mellon University, Pittsburgh, Pa.

Krumm, J. and Shafer, S. A. (1994). Segmenting textured 3D surfaces using the space/frequency representation. *Spatial Vision*, 8(2), 281–308.

Malik, J. and Rosenholtz, R. (1994). Recovering surface curvature and orientation from texture distortion. *ECCV '94: Third European Conference on Computer Vision, Stockholm. (Proceedings)*

Navon, D. (1977). Forest before trees: The precedence of global features in visual perception. *Cognitive Psychology*, 9, 353–383.

Parker, D. M. and Dutch, S. (1987). Perceptual latency and spatial frequency. *Vision Research*, 27(8).

Sagi, D. (1991). Spatial filters in texture segmentation tasks. In B. Blum (Ed.), *Channels in the Visual Nervous System: Neurophysiology, Psychophysics and Models*. London: Freund Publishing House, Ltd.

Sakai, K. and Finkel, L. H. (1995). Characterization of the spatial-frequency spectrum in the perception of shape from texture. *Journal of the Optical Society of America A*, 12(6), 1208–1224.

Schyns, P. G. and Oliva, A. (1994). From blobs to boundary edges: Evidence for time- and spatial-scale-dependent scene recognition. *Psychological Science*, 5(4), 195–200.

Shulman, G. G., Sullivan, M. A., Gish, K. and Sakoda, W. J. (1986). The role of spatial frequency channels in the perception of local and global structure. *Perception*, 15, 259–279.

Todd, J. T. and Akerstrom, R. A. (1987). Perception of three-dimensional form from patterns of optical texture. *Journal of Experimental Psychology: Human Perception and Performance*, 13(2), 242–255.