

Finding Boundaries in Images

JITENDRA MALIK

University of California at Berkeley

PIETRO PERONA

Università di Padova and Massachusetts Institute of Technology

Abstract

In computational vision, finding the boundaries of the regions in a image which correspond to different surfaces in the scene is usually approached as a problem of detecting brightness edges. In this paper, we argue that this is a limited view. Boundaries in images could be associated with differences on a number of visual attributes—brightness, color, texture, stereoscopic disparity and motion—all of which are utilized in human vision. Machine vision systems should do the same. We argue that convolution of the image with a bank of Gaussian derivative filters is a suitable common first stage for this task. We also present some new results on the problem of detecting and localizing brightness edges composed of step, peak and roof profiles.

1 Introduction

The objective of both human and machine vision is to start from some visual input, say a binocular pair of spatiotemporal image sequences $I_L(x, y, t)$, $I_R(x, y, t)$ of a given scene, and perform a set of computations which make possible the achievement of certain goals. These goals certainly include (a) the recognition of objects in the scene and (b) determination of spatial relationships in the scene adequate to support motor tasks like manipulation and locomotion.

Given that the world is piecewise smooth, it is reasonable to argue that computed descriptions of it should represent not only attributes of surfaces (like depth and orientation) but also explicitly represent their boundaries. Indeed, one could argue that much of the essential information is in fact contained in the boundary curves, as demonstrated by the vivid three-dimensional percept evoked by line drawings. These boundary contours would be very useful for the primary visual tasks: object recognition, locomotion (by defining the boundaries of free space), and manipulation.

A variety of cues are available in the visual input to support this task. Neighboring surfaces in general position differ in a number of visual attributes: luminance, color, texture, stereoscopic disparity and motion being perhaps the most significant. We begin by reviewing how biological systems utilize these cues to provide a segmentation of the visual input.

1.1 Boundary detection in biological vision

Psychophysicists have shown that humans can perceive boundary curves defined purely by differences in luminance, color [26], texture [13], stereoscopic disparity [12] and motion [3].

Further insight into the nature and representation of these boundary curves has come from a series of psychophysical experiments. We list some of the more significant observations:

1. Ramachandran, Rao and Vidyasagar[25] have shown that stereopsis can be obtained by fusing a luminance boundary in one eye with a disparate texture boundary or a chromatic boundary in the other eye.
2. Cavanagh, Arguin, von Grunau[7] show that apparent motion can be seen between two alternating stimuli, even if they are defined with respect to their background by different attributes. This was found to be true for all the combinations studied.
3. It is well known that line drawings convey a vivid three-dimensional percept because of the variety of cues (junctions etc) contained in them. Cavanagh [5] showed that this is also true for Necker cube outline figures defined by color and texture, and also (with some complications due to conflicting depth cues) for cubes defined by stereopsis and motion. However, he also found that the perception of shadows and subjective contours was limited to the luminance pathway.
4. By using a visual search task, Cavanagh, Arguin and Treisman [6] show that popout (i.e. the reaction time as a function of the number of distracters has zero slope) was found for oriented bars with boundaries defined by any of these attributes.
5. Boundaries associated with depth discontinuities are labeled to indicate which of the two regions bordering the boundary is nearer and hence is physically attached to the boundary curve [27].

From these converging lines of psychophysical evidence, one can hypothesize that boundaries can be computed independently from the neural representations of luminance, color, texture, stereoscopic disparity and motion. Also, it is reasonable to assume that there is a common representation for these boundaries. Otherwise the hardware for apparent motion would have to be replicated for all the combinations possible which seems to be a waste in view of how ecologically unlikely these combinations are. There would also need to be replication of the object recognition hardware.

1.2 Boundary detection in Machine Vision

Researchers in computational vision have, like psychophysicists, recognized the importance of different pathways such as those for stereopsis or motion. Early work includes that of of Horn on shape-from-shading, Marr and Poggio on stereopsis and Ullman on motion. However, finding boundaries in images has largely been approached as a problem of detecting brightness

edges which provide the primary input for the different shape-from-X modules. The dominant framework was most clearly expounded by Marr[18] who argued for the initial computation of a primal sketch—essentially a brightness edge map followed by postprocessing with some local grouping operations. The different modules use the primal sketch as input to compute a common representation of depth and/or surface orientation in viewer centered coordinates known as the 2.5D sketch.

Marr argued that detection of brightness edges (at multiple scales) was the right first step because they are associated with physical discontinuities. If there existed perfect edge detection mechanisms, this argument would carry a lot of weight. As any one who has run edge-detection programs on real images knows, zero-crossings and their counterparts in other models are often purely due to noise and many real discontinuities are missed. Edge detection at this early stage violates the principle of least commitment; this is particularly so if the texture or stereopsis or motion module is only going to have access to the edges and not to the underlying image. We believe instead that these modules should make use of a much more complete representation of the image—in the next section we argue that the result of convolving the image with a bank of Gaussian derivative filters is much more suitable for this purpose.

Our alternative to the Marr framework is to have a first stage of convolving the image with a bank of filters and to use this representation as the basis for the parallel processing of brightness, color, texture, stereopsis and motion. In each of these pathways, significant discontinuities (boundary curves) are detected. From signal-to-noise ratio considerations, it is clear that there would be an advantage in having the different modules cooperate in detecting boundaries—this idea has been explored further in a Markov Random Field framework by Poggio, Gamble and Little [23]. There is a final common representation—a boundary contour sketch—which could be utilized for visual tasks like recognition.

In the next section, we elaborate on our approach.

2 Local analysis of image patches by filtering

Our approach is loosely inspired by the current understanding of processing in the early stages of the primate visual system. A recent survey may be found in DeValois and DeValois[8]. Of particular interest is the notion of a ‘channel’—a family of quasi-linear cells corresponding to the same shape of receptive field, but different locations in cortex (and hence different locations of the visual field). They can be modeled as computing the convolution of the retinal image I with f_i the point spread function of some linear filter. There are of course many different channels corresponding to the different choices of f_i which will reflect the orientation and size (equivalently, spatial frequency) preferences of the neuron.

To a first approximation, we can classify them into three categories:

1. Cells with radially symmetric receptive fields. The usual choice of f_i is a Difference of Gaussians (DOG) with the two Gaussians having different values of σ . The receptive fields of these cells are similar to those in retinal ganglion cells. Alternatively, these receptive fields can also be modeled as the Laplacian of Gaussian.
2. Oriented oddsymmetric cells whose receptive fields can be modeled as rotated copies of a vertical oddsymmetric receptive field. A suitable point spread function for such a receptive field is $(f(x, y) = G'_{\sigma_1}(x)G_{\sigma_2}(y))$ Note that when $\sigma_1 = \sigma_2$, this point spread function corresponds to directional derivative of a Gaussian which Canny [4] has shown is quite close to being an ‘optimal’ edge detector.
3. Oriented evensymmetric cells whose receptive fields can be modeled as rotated copies of a vertical evensymmetric receptive field. A suitable point spread function for such a receptive field is $(f(x, y) = G''_{\sigma_1}(x)G_{\sigma_2}(y))$ Note that when $\sigma_1 = \sigma_2$, this point spread function corresponds to second directional derivative of a Gaussian which Canny [4] has shown is quite close to being an ‘optimal’ bar detector.

The use of Gaussian derivatives (or equivalently, differences of offset Gaussians) for modeling receptive fields of simple cells is due to Young [28]. Other models have been proposed. Our preference for Gaussian derivatives is based on their computational simplicity and their natural interpretation as ‘blurred derivatives’[16, 15].

Suppose that the image is convolved with a bank of linear filters where the f_i are either (1) (2) (3). We will refer to the collection of response images $I * f_i$ as the *hypercolumn transform* of the image.

Why is this useful from a computational point of view?

- The vector of filter outputs $I * f_i(x_0, y_0)$ characterizes the image *patch* centered at x_0, y_0 by a set of values at a *point*. This is similar to characterizing an analytic function by its derivatives at a point—one can use a Taylor series approximation to find the values of the function at neighboring points. As Koenderink and Van Doorn[16] point out, this is more than an analogy, because of the commutativity of the operations of differentiation and convolution, the receptive fields described above are in fact computing ‘blurred derivatives’. We recommend the Koenderink papers[16, 15] for a discussion of other advantages of such a representation.
- Since filters at multiple scales are used in this characterization, the hypercolumn transform provides a natural setting for multiscale analysis. Premature decisions selecting the scale of analysis are not made. Coarse to fine strategies may be employed.
- This approach tries to extract maximum mileage from simple, local and parallel computations, making VLSI implementations feasible.

We now discuss how the hypercolumn transform can be utilized in the different pathways.

- *Brightness*. In computational vision, it is customary to model brightness edges as step edges and to detect them

by marking locations corresponding to the maxima of the outputs of odd-symmetric filters (e.g. Canny [4]) at appropriate scales. However, it should be noted that step edges are an inadequate model for the discontinuities in the image that result from the projection of depth or orientation discontinuities in physical scene. Mutual illumination and specularities are quite common and their effects are particularly significant in the neighborhood of convex or concave object edges. In addition, there will typically be a shading gradient on the image regions bordering the edge. As a consequence of these effects, real image edges are not step functions but more typically a combination of steps, peak and roof profiles. In section 3, we will outline how the hypercolumn transform approach can be modified to detect and localize correctly these composite edges. More details may be found in [22].

- *Texture*. As the hypercolumn transform provides a good local descriptor of image patches, the boundary between differently textured regions may be found by detecting curves across which there is a significant gradient in one or more of the components of the hypercolumn transform. For an elaboration of this approach, see Malik and Perona[17].
- *Stereopsis*. In stereopsis, the primary problem is that of determining the corresponding points in the left and right view. The responses of filters at a given point in an image form a vector that characterizes the local region of the image. Matching can then be based on finding points in the two views for which these vectors are maximally similar. Kass [14] and Jones and Malik [11] have demonstrated this approach on images of both synthetic and natural scenes. More recent work by Jones and Malik has shown that this approach can utilize orientation and spatial frequency disparity between the two views to extract information about surface slant and tilt, instead of regarding it as an annoying source of noise as in the traditional edge-based methods. The detection of disparity discontinuities can also be done in this framework.

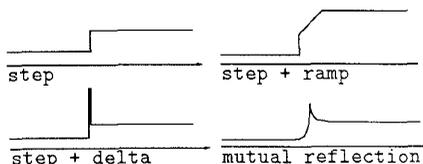


Figure 1: Some examples of edges

- *Motion*. While the problem of computing optical flow has largely been approached by differential methods or by solving the correspondence problem across frames, it can also be studied in the framework of spatiotemporal filtering[1] in a natural generalization of the hypercolumn transform.

3 Brightness boundaries

The problem of detecting and localizing discontinuities in greyscale intensity images has traditionally been approached as one of finding step edges. This is true both for the classical linear filtering

approaches as well as the more recent approaches based on surface reconstruction.

Because of mutual illumination and specularity effects[10, 9], step edges are an inadequate model for the discontinuities in the image that result from the projection of depth or orientation discontinuities in physical scene. One must instead model them as composite edges—a combination of step, peak and roof profiles (Figure 1).

Most local edge detection methods are based on some decision making stage following a linear filtering stage. Typically one looks for maxima in the filtered image perpendicular to the orientation of the edge. Such an approach (e.g. Canny [4]) results in a systematic error in localization whenever there is a composite edge([24](page 9), or [2](Fig. 2. 1)). In section 3.1, we prove that this problem is not specific to the Gaussian derivative filters used by Canny, but is present whatever the linear filter used. For any such filter there is a systematic localization error for composite edges. We outline an alternative approach that does not suffer from this problem.

In section 3.2, we describe the 2D extension of the approach. At each point, the locally dominant orientations θ_i which correspond to the local maxima (over θ) are determined. Allowing for multiple orientations enables junctions to be correctly localized without any rounding. Experimental results are presented.

3.1 Dealing with composite edges

We want to detect and localize edges which we choose to model as arbitrary combinations of lines, steps and roofs. For specificity and simplicity, we assume that the composite edge is $I = c_1\delta + c_2\delta^{(-1)}$, though similar considerations apply for other composite edges.

(A word about notation: we will write $f^{(-1)}(x)$ for $\int_{-\infty}^x f(t)dt$, and $f^{(-n)}(x) = (f^{(-n+1)})^{(-1)}(x)$. So $\delta^{(-1)}$ will be the step function and $\delta^{(-2)}$ a ramp.)

First we establish a proposition which shows that edge localization by looking at peaks in the responses of a fixed, finite family of linear filters leads to systematic errors.

Proposition 1 *For any fixed finite family of filters $\{f_1, f_2, \dots, f_k\}$ there exists an image $I = c_1\delta + c_2\delta^{(-1)}$ for which none of the filter responses have a maximum at $x = 0$*

Proof. Edges are declared at the maxima of the response $I * f(x) = c_1f(x) + c_2f^{(-1)}(x)$. To ensure correct localization, there should be a maximum at $x = 0$ for any combination of c_1, c_2 . For a filter f_i , its response has a maximum at $x = 0$ only if $(I * f_i)'(0) = 0$. Now $(I * f_i)' = c_1f_i' + c_2f_i$, implying that the vector $[c_1 c_2]^T$ is orthogonal to $[f_i'(0) f_i(0)]^T$. To establish the proposition, one has only to pick a composite edge for which the vector $[c_1 c_2]^T$ is not orthogonal to any of the vectors in the fixed, finite family of the k 2D vectors $[f_i'(0) f_i(0)]^T, i = 1, \dots, k$.

In other words, if we had available to us the outputs of k different filters with a clever strategy which would enable us to pick the 'right' filter f_i whose response should be used to localize the edge, we would still be unable to guarantee zero localization error.

Somehow the problem seems to be that for any particular linear filter we are able to construct a composite edge for which the filter is not matched. This suggests an alternative view—construct a parametrized filter which is a linear combination of an even filter f_e (matched to $\delta(x)$) and an odd filter f_o (matched

to $\delta^{(-1)}$) and try to ‘adapt’ it to the particular composite edge in the image by picking the parameter value that maximizes the filter response at each point.

Call $f_\alpha(x) = \cos \alpha f_e(x) + \sin \alpha f_o(x)$ the filter, $I = c_1 \delta + c_2 \delta^{(-1)}$ the image, and $U(\alpha, x) = (I * f_\alpha)(x)$ the response. We want to choose α such that at each point x the response is maximized. Define $V(x) = \max_\alpha U(\alpha, x)$ and call $\alpha(x)$ the maximizing parameter (i.e. $V(x) = U(\alpha(x), x)$). Notice that $\alpha(x)$ must satisfy the equation $\frac{\partial}{\partial \alpha} U(\alpha(x), x) = 0$.

We would like the ‘maximal’ response $V(x)$ to have a maximum in zero, corresponding to the location of the edge: $V'(0) = (U_\alpha \alpha_x + U_x)(\alpha(0), 0) = 0$. Since $U_\alpha(\alpha(x), x) = 0$ then it must be $U_x(\alpha(0), 0) = 0$. Making use of the fact that $f_o(0) = f_e^{(-1)}(0) = 0$ we get the following system of equations:

$$U_x(\alpha(0), 0) = c_1 \sin \alpha f'_e(0) + c_2 \cos \alpha f'_o(0) = 0 \quad (1)$$

$$U_\alpha(\alpha(0), 0) = -c_1 \sin \alpha f_e(0) + c_2 \cos \alpha f_o^{(-1)}(0) = 0 \quad (2)$$

The maximizing value of α , $\alpha(0)$, can be obtained from Equation 2. Substituting this into Equation 1 gives the following condition:

$$f'_e(0) = -f_o^{(-1)}(0) f'_o(0) \quad (3)$$

If this condition is satisfied, the mixed edge $c_1 \delta + c_2 \delta^{(-1)}$ will be localized exactly by the maximum of $V(x)$ defined above.

An alternative approach yields the same condition. Define the vector of filters $F(x) = [f_e(x), f_o(x)]^T$. We localize features by looking for local maxima in the norm of the (vector) response to this filter of I . The squared norm of the response, $|I * F|^2$ is

$$W(x) = \{c_1 \delta + c_2 \delta^{(-1)} * f_e\}^2 + \{c_1 \delta + c_2 \delta^{(-1)} * f_o\}^2 \quad (4)$$

Equating the derivative of this expression with respect to x at the origin to 0 gives the condition

$$c_1 c_2 f_e^2(0) - c_1 c_2 f'_o(0) f_o^{(-1)}(0) = 0 \quad (5)$$

which is the same as Equation 3

Thus, we have the possibility of getting arbitrarily precise localization of composite edges simply by looking for peaks in the response to a *quadratic* filter, i.e. in $\sum (I * f_i)^2$.

This is similar in form to the approach used by Morrone, Owens et al. Morrone et al [19] by a series of psychophysical experiments demonstrated that the human visual system detects features at points of strong *phase congruency*—these could be edges (spectral components have 0 phase), narrow bars (spectral components have 90 phase) or points on trapezoids where ramps meet plateaus (spectral components have 45 or 135 phase). To detect points of phase congruency, Morrone and Owens [20] find maxima of a local energy function $E(x) = F^2(x) + H^2(x)$ where $F(x)$ is the result of a convolution $I * f(x)$, and $H(x)$ is its Hilbert transform (equivalently I could be convolved with the Hilbert transform of f). Morrone and Owens show good empirical results for a particular choice of f .

While their reasoning was in the Fourier domain and aimed at detecting phase congruency based on a psychophysical definition of a feature, we arrive at a similar formulation purely motivated by a computational criterion of localizing composite edges exactly.

From our formulation it follows that there is nothing particularly sacred about the use of Hilbert filter pairs as done by Morrone, Owens et al. In fact, if the composite edge consists

of, say a bar and a step edge at quite different scales, one should probably use f_e and f_o tuned to different widths (scales) and thus not Hilbert pairs. To make a proper choice of these filters, one should instead bring to bear the criteria of having a good signal-to-noise ratio, low stochastic localization error etc. analogous to the approach used by Canny for linear filters.

We have formulated performance criteria for evaluating quadratic filtering based edge-detectors. Details may be found in [22].

3.2 Detecting edges in two dimensions

To detect edges in 2D, we use a Gaussian window to compute the 2D extension of the filter $F(x, y) = f(x)G_{\sigma_2}(y)$. We use two kernels, even and odd; rotated copies F_θ^e, F_θ^o , of the filter are used to compute the square modulus $W(x, y, \theta) = (F_\theta^e * I)^2 + (F_\theta^o * I)^2$, where I is the image. In practice one cannot afford to compute convolutions of the image with filters at an infinity of orientations. It turns out that it is possible to approximate kernels $F(x, y, \theta)$ with arbitrary precision using linear combinations of a finite number of functions. This technique is based on the singular value decomposition of the linear operator associated to the kernel and is described in detail in [21]. What is important to remark here is that it is possible to reason on a continuum of orientations.

At edge points the filter output ‘energy’ W will have a maximum at the orientation θ_e parallel to the edge. Fix θ_e and consider $W(x, y, \theta_e)$. Along a line orthogonal to the edge the problem reduces to the 1D case: there will be an energy maximum at the edge. Edges can be found by marking as ‘edge points’ all the points $p = (x, y, \theta)$ that satisfy:

$$\frac{\partial}{\partial \theta} W(p) = 0, \quad \frac{\partial}{\partial \mathbf{v}_\theta} W(p) = 0 \quad (6)$$

where \mathbf{v}_θ is the unit vector orthogonal to the orientation associated to θ .

Details of the search for the edge points may be found in [22].

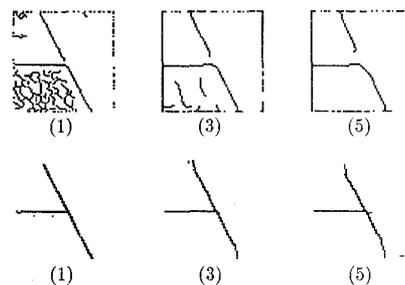


Figure 2: T-junction example. Top: Canny detector with $\sigma = 1, 3, 5$. Bottom: our 2D detector, same $\sigma_1, \sigma_2 : \sigma_1$ ratio 3:1

We have tested the algorithm on both natural and synthetic data. We used a quadratic filter with $f_e = G''_{\sigma_1}$, $f_o = (G''_{\sigma_1})_H$, and the 2D extension computed using a windowing function G_{σ_2} with $\sigma_2 : \sigma_1 = 2$ or 3. In Figure 2 we compare the edges obtained by our edge detector with those found by the Canny edge-detector. While the false positives are largely a matter of threshold selection (in this case), note how the junction is broken up and rounded by the Canny edge detector. Figure 3 shows a comparison for a more complex image.

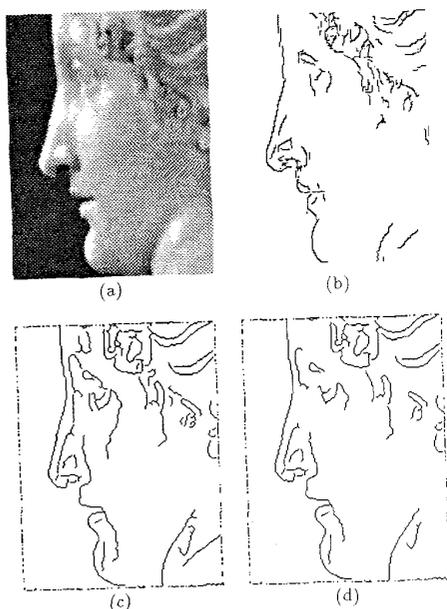


Figure 3: Comparison of the Canny detector and our 2D detector. (a) Original (Paolina Borghese, Canova circa 1800). (b) our detector, $\sigma_1 = 1$, $\sigma_2 : \sigma_1$ ratio 2:1. (c-d) Canny detector with $\sigma = 1$, and threshold (150,250), and (200,400) respectively.

Acknowledgements

This research was supported by NSF Presidential Young Investigator Award No. IRI-8957274 to J.Malik. P.Perona conducted part of this research while at the International Computer Science Institute, Berkeley.

References

- [1] E. Adelson and J. Bergen. Spatiotemporal energy models for the perception of motion. *Journal of the Optical Society of America*, 2(2):284–299, 1985.
- [2] A. Blake and A. Zisserman. *Visual reconstruction*. MIT press, 1987.
- [3] O. Braddick. A short-range process in apparent motion. *Vision Research*, 14:519–528, 1974.
- [4] J. Canny. A computational approach to edge detection. *IEEE trans. PAMI*, 8:679–698, 1986.
- [5] P. Cavanagh. Reconstructing the third dimension: Interactions between color, texture, motion, binocular disparity, and shape. *Computer Vision, Graphics and Image Processing*, 37:171–195, 1987.
- [6] P. Cavanagh, M. Arguin, and A. Treisman. Effect of surface medium on visual search for orientation and size features. *Journal of Experimental Psychology: Human Perception and Performance*, 16(3):479–491, 1990.
- [7] P. Cavanagh, M. Arguin, and M. von Grunau. Interattribute apparent motion. *Vision Research*. pages 1197–1203. 1989.
- [8] R. DeValois and K. DeValois. *Spatial Vision*. Oxford University Press, 1988.
- [9] D. Forsyth and A. Zisserman. Mutual illumination. In *Proceedings of the IEEE CVPR*, pages 466–473, 1989.
- [10] B. Horn. Image intensity understanding. *Artificial intelligence*, 8(2):201–231, 1977.
- [11] D. Jones and J. Malik. Computational stereopsis—beyond zero-crossings. *Invest. Ophthalmol. Vis. Sci. (Supplement)*, 31(4):529, 1990.
- [12] B. Julesz. *Foundations of Cyclopean Perception*. University of Chicago Press, 1971.
- [13] B. Julesz. Textons, the elements of texture perception and their interactions. *Nature*, 290:91–97, 1981.
- [14] M. Kass. Computing visual correspondence. In *Proceedings: Image Understanding Workshop*, pages 54–60, McLean, Virginia, June 1983. Science Applications, Inc.
- [15] J. Koenderink. Operational significance of receptive field assemblies. *Biological Cybernetics*, 58:163–171, 1988.
- [16] J. Koenderink and A. van Doorn. Representation of local geometry in the visual system. *Biological Cybernetics*, 55:367–375, 1987.
- [17] J. Malik and P. Perona. Preattentive texture discrimination with early vision mechanisms. *Journal of the Optical Society of America - A*, 7(5):923–932, 1990.
- [18] D. Marr. *Vision*. W.H.Freeman & Co., 1982.
- [19] M.C.Morrone, D. Burr, J. Ross, and R. Owens. Mach bands depend on spatial phase. *Nature*, (324):250–253, 1986.
- [20] M. Morrone and R. Owens. Feature detection from local energy. *Pattern Recognition Letters*, 6:303–313, 1987.
- [21] P. Perona. Finite representation of deformable functions. Technical Report 90-034, International Computer Science Institute, 1947 Center st., Berkeley CA 94704. 1990.
- [22] P. Perona and J. Malik. Detecting and localizing edges composed of steps, peaks and roofs. In *Proceedings of the International Conference on Computer Vision*. Osaka, 1990.
- [23] T. Poggio, E. Gamble, and J. Little. Parallel integration of vision modules. *Science*, 242:436–440, 1988.
- [24] J. Ponce and M. Brady. Towards a surface primal sketch. Technical Report 824, MIT Artificial Intelligence Laboratory, 1985.
- [25] V. Ramachandran, V. Rao, and T. Vidyasagar. Apparent movement with subjective contours. *Vision Research*. 13:1399–1401. 1973.
- [26] R.L.Gregory. Vision with isoluminant colour contrast: 1. a projection technique and observations. *Perception*, 6:113–119, 1977.
- [27] S. Shimojo, G. Silverman, and K. Nakayama. Occlusion and the solution to the aperture problem for motion. *Vision Research*, 29(5):619–626, 1989.
- [28] R. Young. The gaussian derivative theory of spatial vision: Analysis of cortical cell receptive field line-weighting profiles. Technical Report GMR-4920, General Motors Research, 1985.