Monocular Depth by NonLinear Diffusion

Mariella Dimiccoli †, Jean-Michel Morel ‡ and Philippe Salembier †

† Technical University of Catalonia
Dept. of Signal Theory and Com.
Jordi Girona 1-3, Barcelona, Spain
mariella,philippe@gps.tsc.upc.edu

‡ Superior Normal School of Cachan
Dept. of Applied Mathematics
Pr. Wilson 61, Cachan, France
morel@cmla.ens-cachan.fr

Abstract

Following the phenomenological approach of gestaltists, sparse monocular depth cues such as T- and X-junctions and the local convexity are crucial to identify the shape and depth relationships of depicted objects. According to Kanizsa, mechanisms called amodal and modal completion permit to transform these local relative depth cues into a global depth reconstruction. In this paper, we propose a mathematical and computational translation of gestalt depth perception theory, from the detection of local depth cues to their synthesis into a consistent global depth perception. The detection of local depth cues is built on the response of a line segment detector (LSD), which works in a linear time relative to the image size without any parameter tuning. The depth synthesis process is based on the use of a nonlinear iterative filter which is asymptotically equivalent to the Perona-Malik partial differential equation (PDE). Experimental results are shown on several real images and demonstrate that this simple approach can account a variety of phenomena such as visual completion, transparency and self-occlusion.

1. Introduction

To infer the shape and the distance from the viewpoint of depicted objects, our visual system is influenced by several factors, commonly referred to as pictorial depth cues because of their use by artists to convey a greater sense of depth in a flat medium. The whole issue of how these factors are grouped together by the visual system to convey an unique, stable depth perception is what Kanizsa [14] called the more general "enigma of perception". Gestalt theory was a first scientific attempt to address this fundamental issue. Gestaltists consider human perception as the result of a construction process driven by a set of elementary grouping laws. These laws are supposed to act for every new percept before any high level cognitive process. In the founding Wertheimer paper [32], one can distinguish two kinds of grouping laws. The first kind are elementary grouping laws that start from the atomic local level to recursively construct larger and larger groups (gestalts). The second kind are principles governing the interaction, collaborative or conflictive, between partial gestalts obtained by elementary grouping laws. In a broad overview of Gestalt theory, Metzger [20] showed that depth can be perceived in the absence of binocular correspondence. Although these results were well known at the time computer vision emerged as a new discipline, a great deal of effort has been invested by the computer vision community in coming up with algorithms to recover depth from stereo[18] and from other cues that requires multiples images, such as structure from motion [12] or depth from defocus [23]. More recently, several works on monocular depth perception are focusing on learning approaches that capture contextual information [25, 27, 13] and still involve more neurophysiology than phenomenology. To the best of our knowledge, the laws governing the primary process of depth perception, as opposed to a more cognitive secondary process have still not received an adequate mathematical and computational translation. This lack is mainly due to the qualitative nature of phenomenology. The mathematical definition of digital image was ignored by Gestaltists and the related issues of blur and noise in image formation were even not qualitatively considered. In this paper we attempt a mathematical and computational translation of gestalt laws and principles governing the monocular perception of depth, from the detection of sparse monocular depth cues such as T- and X-junctions and the local convexity to their synthesis into a global depth reconstruction.

In the next section we survey the literature related to the subject. In Section 3 we give a detailed description of the proposed approach to monocular depth perception. In Section 4 we discuss the experimental results and finally section 5 reports the main conclusions of the present work.

978-0-7695-3476-3/08 $25.00 © 2008 IEEE
DOI 10.1109/ICVGIP.2008.97

2. Related work

The first relevant works on monocular depth perception appeared at the beginning of the nineties and presented solutions based on two different perspectives: the contour processing and the region processing perspective. Due to the crucial role of depth perception in the interpretation of illusory contours, most of these seminal works were developed by psychologists and conceived as computational models of illusory contours.

From the contour-processing perspective, the formation of a global percept from local cues has been modeled as an optimization process with a contour interpretation mechanism. Williams [33] described the occlusion mechanism by a set of integer linear constraints. These constraints insure the physical consistency of a contour grouping process with the image evidence. The main limitation of this work is that it foregoes purely local use of local evidence. Saund [26] proposed a solution to this problem based on the use of a token-based algorithmic framework allowing locally derived constraints to propagate globally around a junction graph. The junction label assignment is conducted through annealing-style optimization, which is well known to be susceptible of local optima. Taking a neurophysiological perspective, Heitger et al. [11] proposed a grouping method which consists in convolving a representation of occlusion cues with a set of orientation selective kernels and nonlinear paring operations. This method cannot resolve ambiguities and tends to complete also the background.

From the region-processing perspective, the formation of a global percept from local cues has been modeled as an optimization process with a surface diffusion mechanism. Mumford and Nitzberg [22] proposed a variational formulation presented as a variant of the Mumford and Shahs segmentation model [5], allowing regions to overlap. They first compute edges and T-junctions and then minimize the functional combinatorially with respect to all possible ways of connecting the T-junctions by new edges that is consistent with a given ordering hypothesis. This work has inspired more recent theoretical investigation, addressing the main issues of the numerical minimization of the functional [6] and the computational complexity [30]. Maradarasmi et al.[17] proposed a Bayesian formulation: assuming that all surfaces in the scene are piece-wise constant or frontoparallel, the problem of finding a piece-wise smooth segmentation of the image into surfaces is equivalent to the problem of assigning a discrete depth value to each image pixel. Stella et al.[29] extended Maradarasmi's work by embedding into a Hierarchical MRF explicit decision rules that asserts continuity of depth assignments values along contours and within surfaces, and discontinuity of depth assignment value across contours. A linear diffusion formulation has indeed been proposed by Geiger et al.[9]. First, a set of local surface interpretations to local occlusion cues, such as junctions and corners, is assigned in the form of salient surface-states. Then, a linear diffusion algorithm that block diffusion coefficients at intensity edges is applied. The best image organization is selected based on a coherence measure between pairs of junctions.

A more neurophysiological approach is taken by Kogo et al. [16] and Mordohai et al.[21]. [16] proposed a feedback model based on a surface completion scheme. The relative depths are determined by convolution of Gaussian derivative based filters, while an anisotropic diffusion equation [24] reconstructs the surfaces. [21] integrated under a tensor voting framework first and second order information for automatic junction labeling and selection between modal and modal completion. Recently, Gao et al. [8] proposed a Bayesian inference framework which unify the contour-based and the region-based perspective. T-junctions are computed on atomic regions and broken into terminators. A graph representation is obtained consisting of two types of nodes: atomic regions and its corresponding terminators that make the problem a mixed MRF. The most recent works are learning-based approaches [25, 27, 13]. They are based on the use of a large database of images annotated with human-marked ground-truth to learn local figure/ground labels [25], or the set of parameters capturing the 3D location and orientation of small patches, or models of occlusion[13] based on both 2D and 3D depth cues. The inference is performed on a MRF [27] or on a Conditional MRF[25, 13] to enforce global consistency.

Most of described approaches have been tested only on a limited set of synthetic images [33, 26, 9, 17, 16, 21, 30], or on images previously segmented by interactive methods [29, 8, 25]. Impressive results on real images have indeed been showed when using learning-based approaches [25, 13, 27]. However, they are obtained using a ratio between the number of test images and the number of training images, with manual assignement of the ground truth, almost equal to 1.

In the next section, we propose a full automatic method completely based on gestalt phenomenology that can account for a variety of phenomena on real images.

3. Proposed Approach

The method proposed here involves three main steps. The first step detects a set of monocular depth cues arising from elementary grouping laws. The second step encodes all local and non-local depth relationships, by acting in an additive fashion under non-conflictive conditions (collaboration) and in a exclusive fashion under conflictive conditions (masking). The last step operates a synthesis of all available depth information to infer shape and spatial layout of depicted objects.

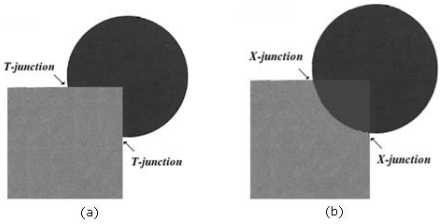


Figure 1. (a) Occlusion. (b) Transparency.

3.1. Computing Monocular Depth Cues

In this work we focus on a subset of monocular depth cues that do not require any *a priori* information about the scene and should be regarded as a direct, immediate response to retina stimulation. For each cue, we detail the psychophysical description as well as its mathematical and computational translation.

Probably the most important monocular depth cue is occlusion. Occlusion occurs when an opaque object partly obscures the view of another object further away from the viewpoint (Fig.1(a)). In this case, the projection of the object contours partially hiding each other creates T-shaped junctions in the image plane. The geometrical configuration of T-junctions encodes relative depth information of the objects in partial occlusion: the stem of the T belongs to the partially occluded object and the roof to the occluding object. A particular case of occlusion is transparency, which occurs when the occluding object is transparent and therefore the more distant objects are visible through the less distant transparent one (Fig.1 (b)). In this case, the projection of object contours creates X-shaped junctions in the image plane. Whereas the geometric characterization of T-junctions alone provides a local signature of occlusion, in the case of transparency a photometrical characterization is also needed. At points where transparency occurs two distinct depths lie in the same line of sight. The process of separating a single luminance value into two contributions is known as scission. Metelli [19] derived two constraints on the photometric conditions required for perceptual scission. The first constraint is known as *magnitude constraint*: a transparent medium cannot increase the contrast of the visible structures. As a consequence, a region can scissor only if its contrast is less than or equal to the contrast of its flanking regions. The second constraint is known as *polarity constraint*: a transparent medium cannot alter the contrast polarity of structures visible through it. Polarity constraints provide a photometrical signature of transparency. Once scission has been identified, the problem of assigning surface properties correctly to the two depths is solved by using the magnitude constraint: the contrast between the regions belonging to the transparent medium is always lower than the contrast between the regions of the underlying object.

From the above description results that figural signatures

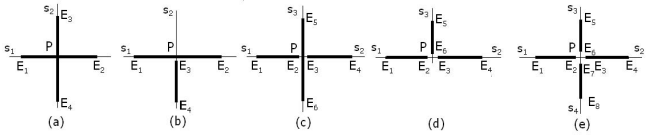


Figure 2. Possible configurations of segments conveying the perception of X-junctions ((a),(c),(e)) and T-junctions ((b),(d))

of occlusion and transparency are respectively T-junctions and X-junctions. Our method for detecting T-junctions and X-junctions is built on the response of the LSD proposed by Grompone et al. [10]. The perception of segments is related to the grouping law of constancy of direction (alignment), which is a special case of continuity of direction. LSD puts together two well known state of the art algorithms for segment-detection: the Burn's segment detector [3] and the meaningful segment detector developed by Desolneux et al. [4]. First the image is segmented into line-support regions using the Burn's strategy and the medium orientation is accurately computed for each support region. Then, following the approach of Desolneux et al. [4], segments are computed as outliers of an unstructured background model. The main advantage of this strategy is that the thresholds of the detection algorithm can be defined in order to control its expected number of false detection under the background model. In addition, the use of a previous line support-region detection step speeds up the computation leading to a line segment detector able to process images in linear time relative to the number of pixels. Furthermore, LSD leads to an easy visualization of T- and X-junctions, even if the junction center is often missing for the detection. In these cases, the visualization of junctions is the result of an interpolation process driven by the good continuation principle. Straight lines are extended and junctions are detected as intersection of straight lines. According to the number and the orientation of intersecting segments, junction points are classified. T-junctions can be detected as intersection of two or three segments (Fig.2(b) and (d)) whereas X-junctions can be detected as intersection of two, three, or even four segments (Fig.2(a),(c),(e)). The intersection of two segments may lead to a T-junction or an X-junction depending on the position of the intersection point P with respect to the tips E_i of the segments. When all tips have sufficient distance from the P, they convey the perception of an X-junction (Fig.2(a)), otherwise of a T-junction (Fig.2(b)). The intersection at point P of three segments s_1, s_2, and s_3 such that two of them, say s_1 and s_2, are aligned may lead to a T-junction or an X-junction depending on the position of P with respect to the tips E_5 and E_6 of the third segment s_3. When both tips of s_3 have sufficient distance from P, then they convey the perception of a X-junction (Fig.2(c)), otherwise of a T-junction (Fig.2(d)). The intersection of four segments at a point P leads to an X-junction when the seg-

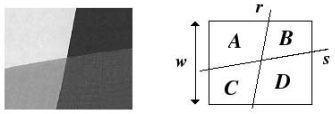


Figure 3. The polarity constraint tells us that s is the contour of the transparent object, since the polarity of the contrast between pairs of adjacent regions delimited by r ((A, B) and (C, D)), does not change when s is crossed

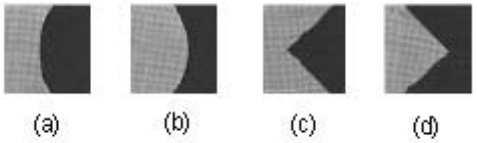


Figure 4. Convexity: contrast polarity and texture property being equal in ((a) and (b)) and in ((c) and (d)), the region with convex contour tends to be perceived as foreground.

ments are two by two aligned and all segment tips have sufficient distance from P (Fig.2(d)).

While occlusion is simply detected by using figural characterization of T-junctions, the detection of transparency involves a photometrical characterization as well since it requires to check the polarity constraint. Let A, B, C, and D be the four regions delimited by the contours r and s forming the X-junction and a squared window of size w centered at the junction center (see Fig.3). The gray level representative of each region, a, b, c and d is obtained as a median value on each region. If the regions A and B are separated by r and A and C are separated by s, then the polarity constraint is satisfied if the difference $a - c$ has the same sign as the difference $b - d$ or if the difference $a - b$ has the same sign as the difference $c - d$. In the latter case, s is the contour of the transparent object and r is the contour of the underlying object. In the former case the contrary is true.

In the absence of occlusion and transparency, the factors that determine which regions are perceived as foreground and which as background, given the complete description of the boundary contours, must be related to the shape of the regions and not to their contrast polarity, or any other texture property. With respect to other global shape properties, convexity has proved to have a stronger influence on figural organization. Its role has been illustrated by Kanizsa: any convex curve (even if not closed) suggests itself as the boundary of a convex body on the foreground (Fig.4). From a mathematical point of view, the convexity of a curve is related to the sign of its curvature. Let $u : R^2 \rightarrow R$ be an image, Du the gradient of u and x a point of u such that $Du(x) \neq 0$ and in a neighborhood of x the iso-level set of u through x is a C^2 Jordan arc Γ. Then the curvature vector $\kappa(u)$ at x is defined by

$$\kappa(u)(x) = -curv(u)(x)\frac{Du}{|Du|} \qquad (1)$$

where $curv(u)(x)$ is the curvature of u at point x. If x is a point of Γ, then the curvature vector $\kappa(u)(x)$ is normal to Γ at x as the gradient vector $\frac{Du}{|Du|}$ and points towards the center of the osculating circle.

A first example of conflict between elementary grouping laws arises in correspondence of T- and X- junctions. In fact, at these points the local interpretation of relative depth conveyed by convexity is never in agreement with the interpretation conveyed by occlusion or by transparency. This situation is called *conflict* by gestaltists and resolved by the *masking* or, in more neurophysiological terms, *inhibition*. As for any other case of conflict, the grouping law that gives the better global explanation of the figure inhibits the competing one. At T- and X-junction points, the masking phenomenon implies the inhibition of convexity.

Occlusion is one of two ways by which observation conditions lead to object obscuration. The second one is camouflage. In camouflage the occluding object is rendered invisible by matching the color or the texture of the background. In both cases the visual system interpolates missing data, a process known as visual completion. This process is important because it is one of the means by which the visual system organizes its depth measurements into meaningful bodies. In the case of occlusion, the perceptual completion of partially occluded objects is referred to as amodal completion. In the case of camouflage, the perceptual completion of occluding objects is referred to as modal completion (see Fig.7). In general, the regions of the image that are visible and lead to visual completion are referred to as "inducers" [7]. Inducers of visual completion are pairs of T-junctions that, when connected by extrapolating one stem and connecting it with the stem of the other element of the pair, obey the "good continuation" law. This means that the interpolated curve should be as similar as possible to the piece of curves it interpolates. According to the Kellman and Shipley's theory of relatability [15] human vision does not always complete contours in presence of T-junctions but uses geometric relationships among them to reduce the number of interpretations that are consistent with a given image. These geometric relationships are synthesized under the concept of relatability. The definition of relatability is as follows (see Fig.5(a)). *Two edges are said relatable if the process of interpolation begins and ends at the points of tangent discontinuity of the contour, called T-junctions, and, their linear extensions meet in their extended regions, forming an outer angle* Φ *less of* $\pi/2$. Psychophysical data suggest that within the category of relatable edges, there are quantitative variations in strength. As can be observed in Fig.5 (b), the strength of the perceived connection decreases when the angle between two edges increases (1, 2, 3) and/or the offset between two parallel edges increases (4, 5, 6). We use these quantitative variations to choose the best relatable T-junction for a given T-junction, when multiple candidate

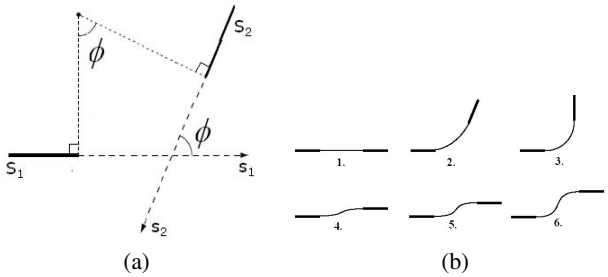


Figure 5. (a) Relatability geometry. (b) Strength variations of relatability.

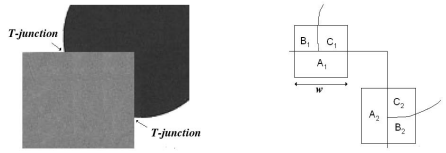


Figure 6. Amodal completion: pairs of relatable regions (A_1,A_2),(B_1,B_2), and (C_1,C_2) have similar gray level.

pairs satisfying the relatability conditions are possible. The angle is used as first criterion and, in case of angle parity, the offset between the two candidate edges is considered. As additional constraint for relatability we also impose a photometric condition. Let a_i b_i and c_i be respectively the medium gray level of regions A_i, B_i and C_i delimited by the contours forming the T-junction and a squared window of size w centered at the junction center (see Fig.6). The relatability condition is checked only at pairs of T-junctions such that the medium gray level of the region forming the top, say a_1, and the regions forming the stem b_1 and c_1 have a medium gray level similar respectively to a_2, b_2 and c_2 or a_2, c_2 and b_2. In the case of camouflage, T-junctions show up as line ends that correspond to the stem of the T. When the occluding object matches the color of only one of the two background objects, pairs of T-junctions that lead to modal completion show up as pairs of corners (see Fig.7). We shall call angles that lead to modal completion *degenerated T-junctions*. Pairs of degenerated T-junctions are detected using the quantitative variations of relatability. This criterion allows also to take a decision of which of the two segments forming the corner is the stem of the T-junction. For instance in Fig.7 the application of this criterion lead to see the triangle behind the square.

3.2. Computing Initial Depth Values

Let z be the depth image. We call source points the points for which the initial depth gradient Dz_0 is not zero and normal points the points for which $Dz_0 = 0$. Source points arise in correspondence of depth cues. In the following we shall call foreground source points (FSPs) all source

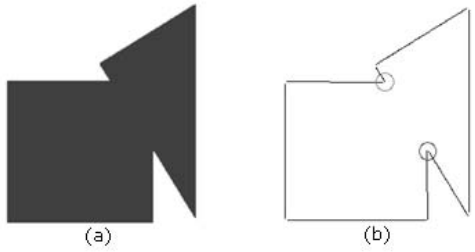


Figure 7. (a) Modal completion: modal contours through an homogeneous zone. (b) Boundaries objects visualized using LSD: T-junctions show up as corners

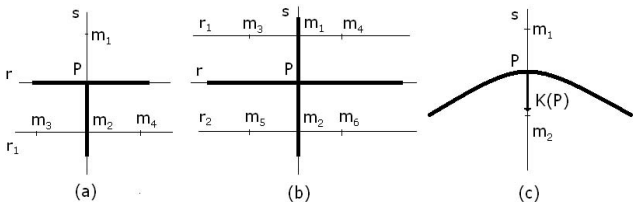


Figure 8. FSPs and BSPs arising from: (a) T-junctions, (b) transparency, (c) convexity

points marking the regions that are closer to the viewpoint and background source points (BSPs) the points more distant to the viewpoint. We assign a positive value to FSPs, and zero to BSPs. The rest of the image is initialized with value zero. The way source points are computed depends on the type of depth cue. Source points arising from a T-junction at point P are computed as follows (see Fig.8(a)). Let s be the line containing the segment that forms the stem of the T-junction. Let m_1 and m_2 be the points belonging to s and having distance d from P. If $m2$ is the point lying on the stem and $r1$ the line perpendicular to s and passing through m_2, then m_1 is the FSP and the points m_3 and m_4 belonging to $r1$ and having distance d from m_2 are the BSP. Source points arising from convexity at point P of a curve are computed in the following way (see Fig.8(c)). Let r be the line passing through P and having the direction of the gradient at P. Let m_1 and m_2 be the points belonging to r and having distance d from P. If m_1 is the point lying on the half-line having origin in P and oriented as the curvature vector at x, then m_1 is the BSP and m_1 the FSP. Source points arising from transparency at point P are computed as follows (see Fig.8(b)). Let s be the line containing the contour of the transparent object, and m_1 and m_2 be the points belonging to s and having distance d from P. Let $r1$ be the line perpendicular to s and passing through m_1, and $r2$ the line perpendicular to s and passing through m_2. Let m_3 and m_4 be the points belonging to $r1$ and having distance d from m_1, and m_5 and m_6 the points belonging to $r2$ and having distance d from m_2. If the gray level difference between m_4 and m_6 is larger than the gray level difference between m_3 and m_5, then m_3 and m_5 are FSPs whereas

m_4 and m_6 are the BSPs. The distance d is at least 4 pixels to take into account image blur. It allows one to jump over edges.

3.3. Depth Diffusion

Once source points have been computed, our goal is to extrapolate relative depth values to the entire image domain. To this goal we use a neighborhood filter. A neighborhood filter is any filter which performs an average of the values of pixels which are close in gray level value. The underlying assumption is that pixels belonging to the same object have a similar gray level. The average is commonly computed on pixels belonging to the neighborhood in spatial distance as in the Yaroslavsky neighborhood filter (YNF) [34], the SUSAN filter [28] and the bilateral filter [31], or in a fully non-local way as in the non-local means [2]. Let u be an image defined on a bounded domain $\Omega \in R^2$. The YNF computes a weighted average that can be written in a continuous form as

$$YNF_{h,\rho}u(x) = \frac{1}{C(x)} \int_{B_\rho(x)} u(y) e^{\frac{-|u(x)-u(y)|^2}{h^2}} dy \quad (2)$$

where $B_\rho(x)$ is a ball with radius ρ and center x, $x \in \Omega$ and $C(x) = \int_{B_\rho(x)} e^{\frac{-|u(x)-u(y)|^2}{h^2}} dy$ is the normalization factor. Neighborhood filters have been proved to be asymptotically equivalent to a Perona-Malik equation [1], one of the first nonlinear PDE used for image restoration.

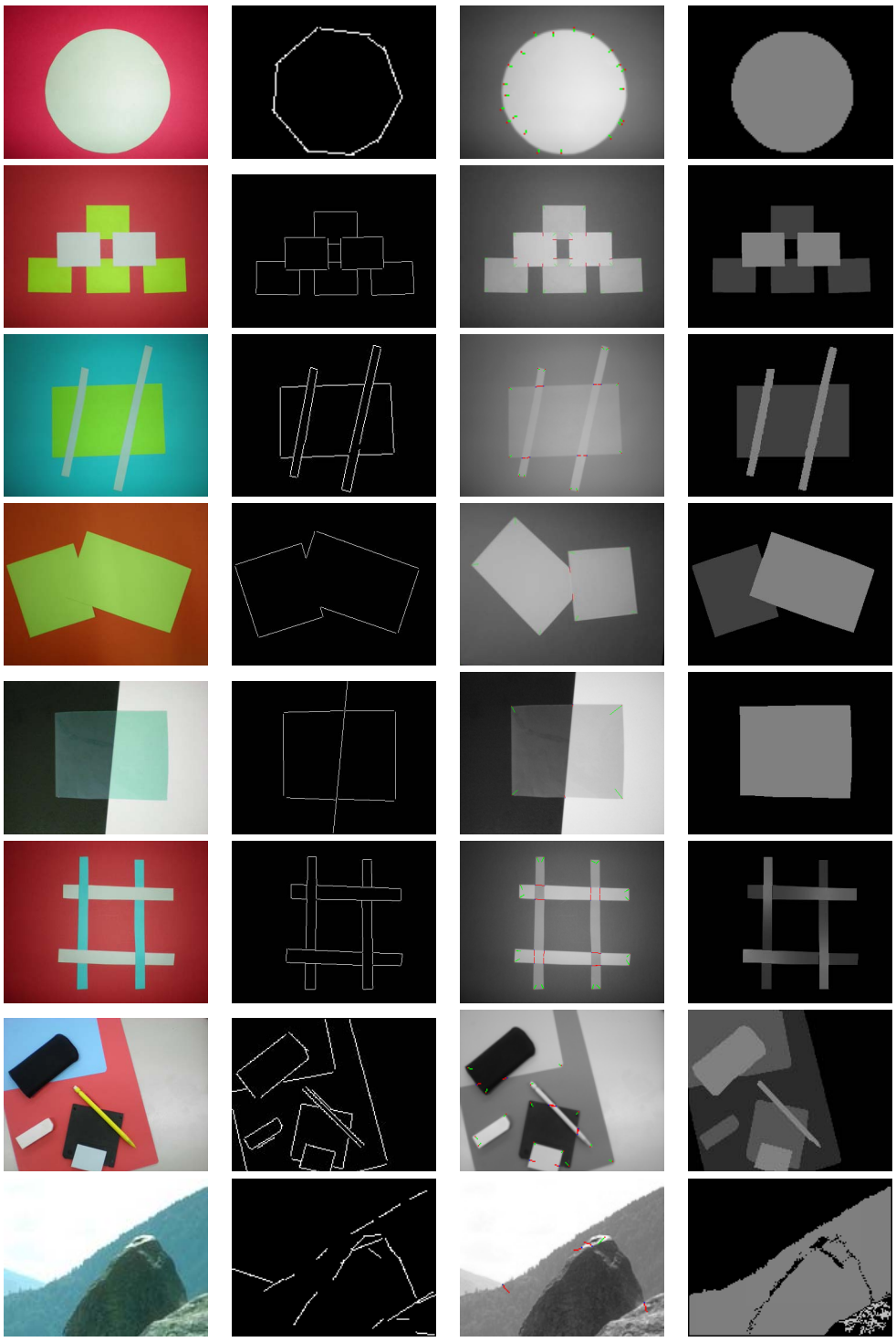
The diffusion process on the depth image z is performed using the gray level image u to define the neighborhood. In order to make the diffusion process faster, the sup of the neighborhood is taken instead of the average while the average is taken only in the last iterations. The depth diffusion filter (DDF) can be written in a continuous form as

$$DDF_{h,\rho}z(x) = \sup_{y \in B_\rho(x)} z(y) e^{\frac{-|u(x)-u(y)|^2}{h^2}} \quad (3)$$

This filter is applied iteratively until the stability is attained. After each iteration, the values of FSPs and BSPs are modified in order to hold at least the initial depth gradient. This constraint corresponds to Neumann internal boundary conditions which are understood as a prespecified jump on the $c\frac{Dz}{Dn}$ as the boundary crossed, where c is a positive constant and n is the normal to the boundary. This allows one to handle simple sorting when objects are located in multiple layers. In the case of occlusion and transparency there is also a depth order between the two regions separated respectively by the stem of the T and by the contour of the underlying object. Occlusion and transparency do not carry any information about the partial order between the underlying object and the background. This depth order can be inferred by other cues, such as convexity or visual completion. When information about this partial order is present, the depth gradient between one of the BSPs and the FSPs increases. This is the reason for which we force source points to hold "at least" the initial depth gradient. To handle visual completion, after each iteration pairs of relatable regions (see Fig.6) are forced to maintain the same depth. In the case of modal completion, one of the two BSP has a gray level similar to the one of the FSP. For this reason we modify the way the neighborhood is defined. Let r and s be the lines the modal contours lie on. The neighborhood N_ρ is defined as follows: $N_\rho = \{y \mid y \in B_\rho(x), y \in \alpha_r(x), y \in \beta_s(x)\}$, where $\alpha_r(x)$ is the half image plane including x with origin the line r and $\beta_s(x)$ is the half image plane including x with origin the line s.

4. Experimental Results

We tested our model on a set of real images (taken by a digital camera) involving occlusion, transparency, convexity, visual completion (both amodal and modal) and self-occlusion. For each experiment we show four images: the original image; the image showing the segments found by applying LSD on the original image; the image where the initial depth gradient at depth cue points is represented through vectors pointing to the region closer to the viewpoint (red vectors arise from T-junctions; green vectors arise from local convexity and each of them represents the point having the biggest curvature value of the connected components obtained by thresholding the curvature); the depth image obtained performing the proposed method. The depth map is rendered through gray level values (high values indicate regions that are close to the camera). In the example on the first row (Fig.9), local convexity induces to see the disk over the table. The second row is an example involving convexity and occlusion: it shows that the proposed method is able to handle simple sorting in presence of multiple depth layers. The third and the fourth rows are examples of amodal and modal completion respectively: in the former case, the detection of pairs of relatable T-junctions leads to see the green piece of paper partially occluded by the white strips as a meaningful unit; in the latter case, the detection of a pairs of degenerated T-junctions leads to see the rectangle in front of the square. In the example on the fifth row, the transparency phenomenon is correctly interpreted. In the example on the sixth row, occluding contours have different depth relationships at different points along its continuum. However, the proposed method performs well also in this ambiguous situation. The examples on the last two rows involve more realistic scenarios. While in the first example the solution is pretty contrived, in the second we show a case of failure. What has caused the break in this case is that a region with homogeneous texture (the mountain behind the rock) has been marked as FSP thanks to the T-junction



(a) (b) (c) (d)

Figure 9. (a) Original (b) Segments detected by LSD (c) Local Depth Information (d) Depth image

on the rock pick and as BSP thanks to the curvature vector. This example also demonstrates that the proposed DDF can handle homogeneous texture (see mountains and the biggest rock) but fails when shading conditions cause strong intensity oscillations (see the rock in the bottom-right corner).

5. Conclusions

In this work we have proposed a mechanism for monocular depth perception completely based on phenomenology. Experimental results involving occlusion, transparency, convexity, visual completion (both amodal and modal) and self-occlusion have shown a correct interpretation of several real images. In contrast with anterior state of the art, the cue detection was automatic and the depth synthesis led by a very elementary mechanism, namely an iterated neighborhood filter. However, the experiments shown here on real images give a high confidence to the DDF as a way to diffuse depth information from local depth cues. In particular, contradictory information given by conflicting depth cues were dealt with correctly by the proposed mechanism which permits two regions to invert harmoniously their depths, in full agreement with phenomenology, and very diverse gestalt laws were fused harmoniously within this simple and plausible mechanism. Although the experiments have been performed on real images, a new generation of detectors will be needed to deal with real world images, where T-junctions, convexity, etc. cannot always be computed from local information. Further research must therefore focus on more and more global cue detectors.

References

[1] A. Buades et al. Neighborhood filters and pde's. *Numerische Mathematik*, 105(1):1–34, 2006.

[2] A. Buades et al. The staircasing effect in neighborhood filters and its solution. *IEEE Tr. on IP*, 15(6):1499–1505, 2006.

[3] J. Burns et al. Extracting straight lines. *IEEE Tr.on PAMI*, 8(4):425 – 455, 1986.

[4] A. Desolneux et al. Meaningful alignments. *IJCV*, 40(1):7–23, 2000.

[5] D.Mumford and J.Shah. Optimal Approximations of Piecewise Smooth Functions and Associated Variational Problems. *J. on Communications in Pure and Applied Mathematics*, 42:577–685, 1989.

[6] S. Esedoglu and R. March. Segmentation with depth but without detecting junctions. *J.Math.Imaging and Vision*, 18:7–15, 2003.

[7] R. Fleming and B. Anderson. *The Visual Neurosciences. Cambridge,MA: MIT Press*, pages 1284–1299. Chalupa, L. and Werner, J.S. Eds., 2004.

[8] R.-X. Gao et al. Bayesian inference for layer representation with mixed markov random field. *LNCS*, 4679:213–224, 2007.

[9] D. Geiger and P. L. Visual organization for figure-ground separation. In *CVPR*, pages 155–160, 1996.

[10] R. Grompone von Gioi et al. Lsd: A line segment detector. Submitted to IEEE Tr. on PAMI, 2007.

[11] F. Heitger and R. von der Heydt. A computational model of neural contour processing: Figure-ground segregation and illusory contours. In *ICCV*, pages 32–40, 1993.

[12] H. Helmholtz. *Treatise on Physiological Optics*. James P. C. Southall, 1925.

[13] D. Hoiem et al. Recovering Occlusion Boundaries from a Single Image. In *ICCV*, pages 1–8, 2007.

[14] G. Kanizsa. *La Grammatica del Vedere*. Diderot, 1996.

[15] P. Kellman and T. Shipley. Visual interpolation in object perception. *Current Directions in Psychological Science*, 1(6):193–199, 1991.

[16] N. Kogo et al. Reconstruction of subjective surfaces from occlusion cues. In *Biologically Motivated Computer Vision: second workshop of BMVC*, pages 311–312, 2002.

[17] S. Madarasmi et al. Illusory contour detection using MRF models. In *World Congress on Computational Intelligence*, pages 4343 – 4348, 1994.

[18] D. Marr. *Vision*. W.H.Freeman and Co., New York, 1982.

[19] F. Metelli. The perception of transparency. *Scientific American*, 230:354–366, 1974.

[20] W. Metzger. *Gesetze des Sehens*. Waldemar Kramer, 1975.

[21] P. Mordohai and G. Medioni. Junction Inference and Classification for Figure Completion using Tensor Voting. In *CVPRW*, volume 4, pages 56–64, 2004.

[22] M. Nitzberg and D. Mumford. The 2.1-D Sketch. In *ICCV*, pages 138–144, 1990.

[23] A. Pentland. A new sense for depth of field. In *ICCV*, pages 839–846, 1985.

[24] M. Proesmans and L. Gool. Grouping based on coupled diffusion maps. *LNCS*, pages 196–216, 1999.

[25] X. Ren et al. Figure/ground assignment in natural images. In *ECCV*, pages 614–627, 2006.

[26] E. Saund. Perceptual organization of occluding contours of opaque surfaces. *Computer Vision and Image Understanding*, 76(1):70–82, 1999.

[27] A. Saxena et al. Learning 3-d scene structure from a single still image. In *ICCV*, pages 1–8, 2007.

[28] S. Smith and M. Brady. Susan - a new approach to low level image processing. *IJCV*, 23(1):45–78, 1997.

[29] X. Stella et al. A hierarchical markov random field model for figure-ground segregation. In *CVPR*, pages 110–133, 2001.

[30] S. Thiruvenkadam et al. Segmentation under occlusion using selective shape prior. *Scale Space and Variational Methods in Computer Vision*, 4485:191–202, 2007.

[31] C. Tomasi and R. Manduchi. Bilateral filter for gray and color images. In *ICCV*, pages 988–994, 1998.

[32] M. Wertheimer. Untersuchungen zur Lehre der Gestalt, II. *Psycologische Forschung*, 4:301–350, 1923.

[33] L. Williams. Perceptual organization of occluding contours. In *ICCV*, pages 133–137, 1990.

[34] L. Yaroslavsky. Digital picture processing - an introduction, 1985. New York: Springer-Verlag.