In ICIP 2017, Beijing, China

MONDRIAN STEREO

Dylan Quenneville *Daniel Scharstein*

Middlebury College
Middlebury, VT, USA

ABSTRACT

Untextured scenes with complex occlusions still present challenges to modern stereo algorithms. We consider the pathological case of *Mondrian Stereo*—scenes consisting solely of solid-colored planar regions, inspired by paintings by Piet Mondrian. We analyze assumptions that allow disambiguating such scenes and present a novel stereo algorithm employing symbolic reasoning about matched edge segments. We demonstrate compelling stereo matching results on synthetic scenes and discuss how our insights could be utilized in robust real-world stereo algorithms for untextured environments.

***Index Terms*—** Stereo, untextured scenes, occlusion

1. INTRODUCTION

While there has been tremendous recent progress in stereo matching [1, 2], current techniques are still challenged by scenes with complex occlusions and lack of texture (Fig. 1). In this paper we consider the pathological case of scenes consisting solely of uniformly-colored planar surface patches arranged in 3D, resembling the abstract paintings by Piet Mondrian. Borrowing a phrase first suggested by Richard Szeliski, we name this problem *Mondrian stereo*, and present an algorithm for its solution (Fig. 2). The priors and search strategy employed in our algorithm yield important insights that promise to be useful for robust real-world stereo algorithms.

2. RELATED WORK

Stereo matching has a long tradition [1]. The first algorithms to successfully handle textureless objects and thin scene structures utilized color segmentation [4, 5, 6]. More recently, over-segmentation into superpixels has been used in complex energy minimization approaches [7, 8]. However, all existing approaches require at least some scene texture and cannot handle pure Mondrian scenes where edges between regions provide the only disparity cues. Dealing with slanted surfaces makes this even more difficult [9, 10]. While edge-based algorithms were among the earliest stereo methods [11], they only yield sparse disparity estimates and do not explicitly reason about the disparities of inter-edge regions.

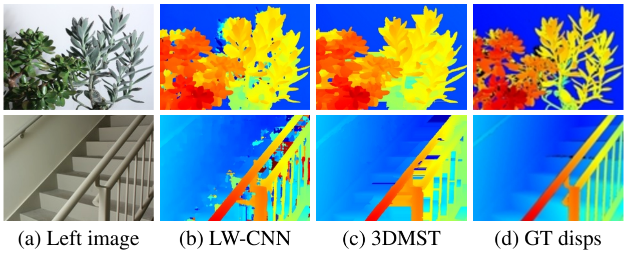


Fig. 1. Common failure cases for state-of-the-art stereo methods (b, c) from the Middlebury stereo evaluation [2]. Comparing the results with the ground truth (d) reveals that thin objects and untextured background regions cause problems.

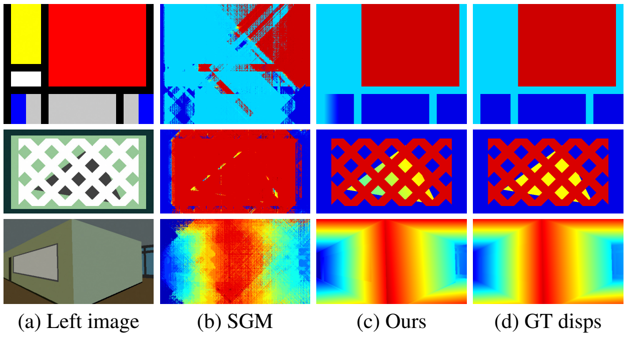


Fig. 2. (a) Synthetic Mondrian scenes devoid of texture. (b) Standard stereo methods like SGM [3] fail in the absence of texture and presence of complex occlusions. (c) Our Mondrian stereo results, employing symbolic reasoning and proceeding solely from matched edges. (d) Ground truth.

The failure cases in Fig. 1 reveal that proper depth assignment of disconnected background regions requires reasoning about object color, which is commonly done in image segmentation [12]. This idea has been applied to stereo matching [13], but proper occlusion reasoning yields higher-order energy functions that are difficult to minimize.

In this work we propose a symbolic algorithm that uses edge disparities to constrain and infer disparities of adjacent surfaces, while also taking color similarity into account. Similar to Ishikawa and Geiger [14] we hope to inspire new prior models for stereo algorithms, partially motivated by observations about human perception [15, 16, 17].

Support by NSF grant IIS-1320715 is gratefully acknowledged.

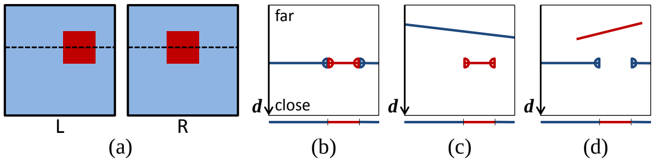


Fig. 3. (a) Simple Mondrian scene. (b–d) Multiple valid depth interpretations of the marked scanline, with different edge ownerships indicated by semicircles. Note that the only disparity cues are provided by the square's vertical edges.

3. ALGORITHM

We now describe our algorithm for Mondrian stereo matching. We first discuss assumptions necessary to disambiguate Mondrian scenes, then detail the different processing steps.

3.1. Assumptions

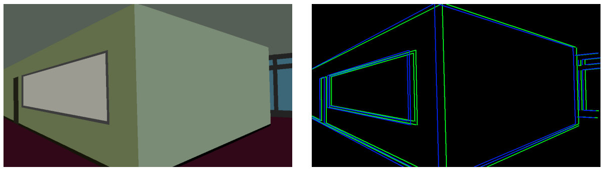
We assume as input a rectified stereo pair with known disparity range $d = [0, d_{\max}]$. Each stereo pair depicts a Mondrian scene consisting of single-colored planar surface patches (segments), arranged in 3D. Segments can be concave and also can have holes. We assume that adjacent segments have different colors so that depth discontinuities always give rise to a color change. Segments with different colors may or may not be coplanar. The same holds for disconnected segments with the same color, though in this case we assume coplanarity is likely.

Since the segments are textureless, non-horizontal edges between segments are the only source of 3D information. In general, Mondrian scenes have an infinite number of valid 3D interpretations (Fig. 3). An important concept is *edge ownership*: each edge must be owned by at least one of its adjacent segments [7]. If it is only owned on one side, it forms an object boundary (the other side has farther depth); if it is owned on both sides, it is a surface crease or simply a surface color change (Fig. 3b–d).

In order to select the most likely interpretation, we need to make additional assumptions. These assumptions are chosen to favor 3D interpretations by human observers. While we did not perform an extensive user study, the "correct" interpretation chosen by human observers is usually clear and makes intuitive sense. Our algorithm makes two key assumptions:

1. In the absence of other evidence, each surface is assumed to be as close as possible. That is, we prefer Fig. 3b (a red square painted on a blue surface) over Fig. 3d (a red surface seen through a square hole in the blue surface).

2. We prefer to assign disconnected segments with the same color to a single surface as long as there is a valid geometric interpretation (i.e., closer occluding segments) [13].



(a) Left image (b) L+R edges overlaid

Fig. 4. Edge extraction. In each image locations of color changes are approximated with polylines.

3.2. Edge matching

The first phase of our algorithm is edge extraction and edge matching. Given our fairly clean synthetic images, which only contain a small amount of Gaussian noise, we can extract segments using a simple union-find connected-component scanline algorithm. If the sum of the absolute color difference across bands is less than a threshold $\lambda = 35$ for two neighboring pixels, they are considered part of the same component.

After components are found, we find all non-horizontal edges and approximate them with polylines. To do so, we find *edgels* (component label changes) on each scanline, and track these edgels from one scanline to the next, as long as the component labels stay the same on both sides. Each sequences of connected edgels is stored as an "edgel curve." We approximate the edgel curve with a sequence of line segments using the standard split-and-merge algorithm, using a maximum distance of $\epsilon = 1$ pixel. Fig. 4 shows an example.

We then match each extracted edge segment in the left image with candidate edge segments in the right image. First we check that the match hypothesis segment in the right image has the same colors on both sides within a threshold. To allow for partial occlusion of an edge we allow "hallucinating" edges beyond their endpoints to the maximal vertical extent of the two edges. If the calculated disparities are outside the disparity range, the hypothesis fails. We also check for contradictory matches by disqualifying a hypothesis that matches an edge in the right image to two edges in the left image, backtracking over earlier assignments if necessary. If no match is found for an edge in the left image, the edge is disregarded.

3.3. Plane generation

We generate a disparity plane hypothesis for each component using two or more non-collinear edge segments attached to the component. We fit a plane to the (x, y, d) coordinates of the four endpoints from two edges. We disqualify plane assignments that yield disparity residuals above threshold τ_1 = 1.0 at any of these four base points, or residuals above threshold τ_2 = 1.5 along any of the segment's other matched edges. Once a legal assignment is found, we use the remaining edges of the plane as points to refine the plane fit as long as the residuals of the refined fit stay within the thresholds. In case no two edges yield a legal plane fit, the component's plane is left uninitialized and is handled in later processing steps described below.

An important feature of our algorithm is that plane fitting yields constraints on edge ownership. In this first phase of the algorithm we seek the closest consistent plane hypothesis for each component (assumption 1). Thus, each edge that contributed to the plane fit is tentatively owned by that plane. Edges can be owned by both adjacent planes (which can be coplanar or form a crease at that edge), and at most one of these ownerships might be undone at a later stage. An edge owned by only one adjacent plane indicates an occluding edge (an object boundary / depth discontinuity); in this case the ownership is firm and we consider the endpoints of the edge *fixed points*. If a plane has at least three fixed points, it is a *fixed plane* whose plane equation cannot change later.

3.4. Component merging

After the first pass we have a tentative assignment of disparity planes to segments satisfying assumption 1. To account for occlusions that divide planes into multiple components, we now turn to assumption 2 and attempt to merge each pair of like-colored components by assigning them to a common plane. There are three possible scenarios. First, if one of the two planes is fixed (has at least 3 fixed points), then the other plane must coincide (fit with acceptable residuals). If this is not the case, the two components cannot be merged. Second, if neither plane is fixed but there are at least 3 fixed points between the two components, there is only one possible plane that can fit both original components. If this new plane has acceptable residuals and is not inconsistent with edge disparities (does not violate visibility constraints), the two components are merged. Finally, if there are less than 3 fixed points between the two components, we attempt to fit a new plane using all points owned by the two planes. Only if a new edge-consistent plane is found, the components are merged.

While plane hypotheses are solely generated from edge matches, we can optionally utilize any weak texture present in the scene to choose between competing plane assignments. This is the only point in our algorithm where we propose to utilize a "data term." We demonstrate below that this can help disambiguate certain difficult scenes. We use as matching cost the average absolute color difference between the segments from the two images, aligned by the proposed plane equation; more robust costs such as NCC or census are clearly possible. We compare the costs resulting from the original plane and the proposed new plane merging the components. If for either component the new plane generates a higher matching cost than the original plane, the hypothesis is disqualified.

For every successful merge hypothesis, we attempt to assign the remaining like-colored components to this new plane, using the same three checks: acceptable residuals, edge consistency, and (optionally) non-increasing matching costs. For each complete hypothesis, we calculate the total number of components that were merged. We compare this count with the best hypotheses for all other like-colored components and keep the hypothesis with the highest merge count, i.e., the lowest number of planes remaining.

In addition to merging disconnected components with the same color, we also attempt to merge adjacent components with different colors if their planes are roughly coplanar. If we can fit a new plane to both segments with acceptable residuals, we use this joint plane for both components, which simplifies the scene description and increases its accuracy.

Finally, we also attempt to extend planes into adjacent segments with uninitialized planes, which are typically caused by occlusion. If this does not result in edge inconsistencies, we keep the extended plane, otherwise we leave it uninitialized. Usually at least one edge of an uninitialized component will generate an edge-consistent extended plane.

4. EXPERIMENTS

Figure 5 shows results for ten challenging Mondrian scenes. Each row contains the original image pair as well as five disparity maps: comparison results by SGM [3] and ELAS [18], our initial results after phase 1, our final results after component merging, and the ground truth. It can be seen that the comparison methods (c, d) fail completely in the absence of texture. Our initial results (e) are already mostly correct, except that some segments are assigned to the foreground depth (scenes 2, 3, 4, 6) due to assumption 1, while other segments have uninitialized planes (scenes 1, 2, 4, 5, 7, 8) due to unmatched edges and/or an insufficient number of supporting edges. Our final results (f) are almost identical to the ground truth for scenes 1–9 except for minor differences in the plane equations caused by discretization errors during edge extraction. Comparing (e) and (f) it can be seen that component merging is able to both assign correct planes to disconnected background regions, and propagate planes into partially occluded regions. Scene 6 demonstrates that utilizing texture can resolve ambiguous cases such as the depth of the left two squares, which have the same edge disparities. Scene 7 demonstrates that our edge matching step is able to deal with ambiguous matches due to repetitive patterns by searching for a globally consistent solution. Finally, scene 10 demonstrates a failure case: if segment colors are too similar, multiple scene surfaces can get merged into a single component (the three gray "walls" in this example). Our algorithm still finds a valid geometric interpretation by treating the boundaries as occlusion edges and placing a plane at a greater distance, but the algorithm is unable to detect and correct the segmentation error. Human observers, in contrast, readily add the necessary surface creases to arrive at the correct interpretation without depth discontinuities. This matches the results reported Ishikawa and Geiger [14].

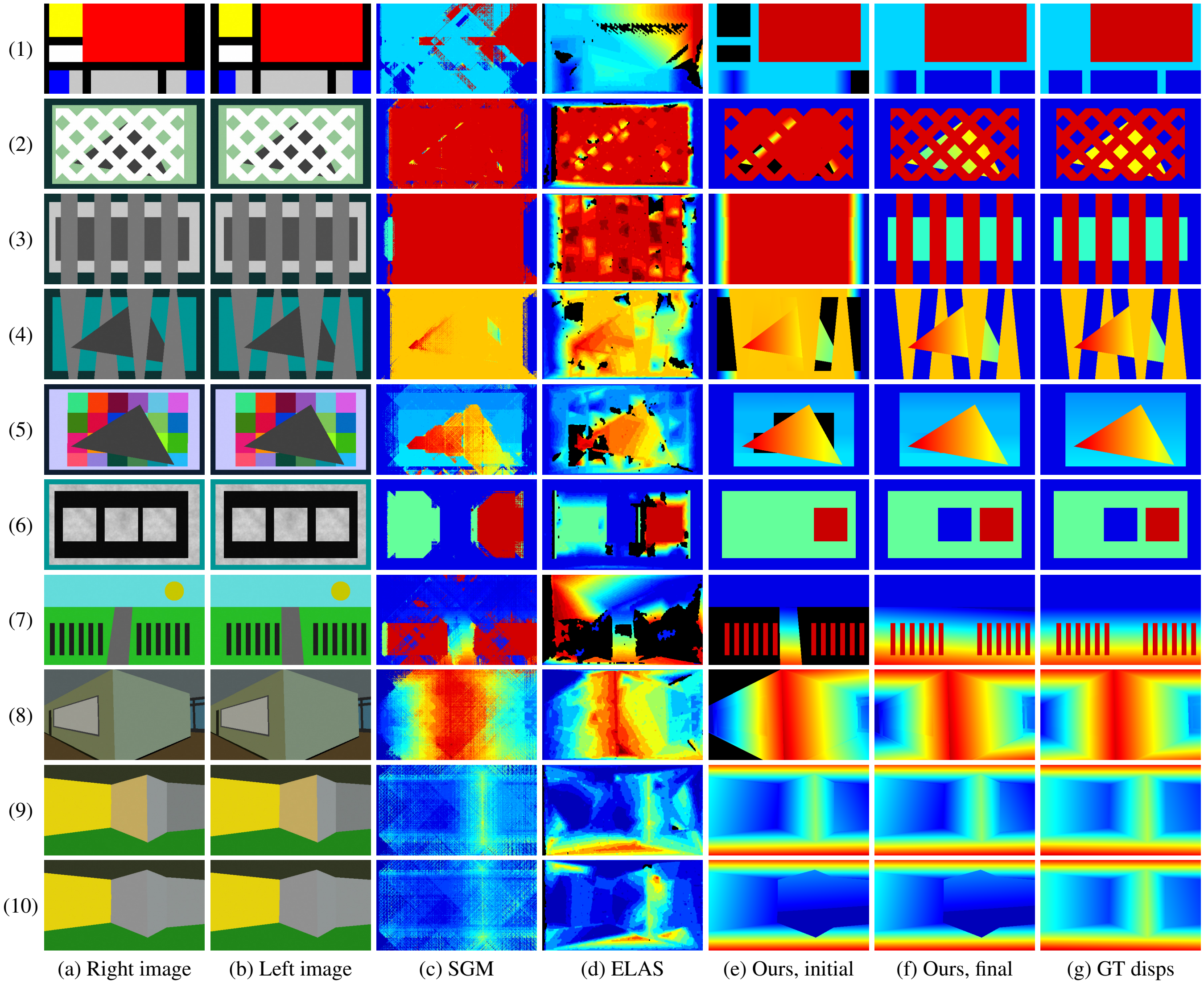


Fig. 5. Experimental results on 10 synthetic Mondrian scenes. (a, b) Input images, displayed in reverse order to allow crossed fusion. (c, d) Results by traditional stereo methods SGM [3] and ELAS [18], which fail due to the lack of texture. (e) Our results after the first pass, satisfying assumption 1. (f) Our final results after component merging, satisfying assumption 2. (g) Ground-truth disparities. See Section 4 for a detailed discussion.

5. CONCLUSION

We have presented a novel stereo algorithm that can solve challenging synthetic image pairs depicting "Mondrian" scenes consisting of untextured planar 3D objects with complex occlusion relationships. Such scenarios still present significant challenges for current stereo methods.

Our algorithm proceeds in two stages. It first fits planes to matched edge segments in order to construct a valid initial 3D interpretation of the visible surface patches in the scene. It then merges like-colored components, while performing proper occlusion reasoning, in order to recover larger surfaces separated by occluding objects. Weak surface texture can optionally be used for disambiguation.

Our long-term goal is to employ similar assumptions and strategies in robust algorithms for difficult real-world images. While not all components of our current method will readily translate to real images (for instance, achieving a consistent segmentation in both images is difficult), the assumptions we identified and our multi-step plane reassignment strategy bear significant promise. Many current stereo methods are formulated as monolithic energy minimization problems [7, 8, 13, 19] that are difficult to optimize. At the same time, their energy functions are not complex enough to allow complete visibility reasoning. Solving a sequence of simpler optimization problems while explicitly reasoning about untextured surface segments could prove a promising alternative.

6. REFERENCES

[1] D. Scharstein and R. Szeliski, "A taxonomy and evaluation of dense two-frame stereo correspondence algorithms," *IJCV*, vol. 47, no. 1-3, pp. 7–42, 2002.

[2] D. Scharstein, R. Szeliski, and H. Hirschmüller, "Middlebury stereo vision page," http://vision.middlebury.edu/stereo/, 2017.

[3] H. Hirschmüller, "Stereo processing by semiglobal matching and mutual information," *IEEE TPAMI*, vol. 30, no. 2, pp. 328–341, 2008.

[4] H. Tao, H. Sawhney, and R. Kumar, "A global matching framework for stereo computation," in *ICCV*, 2001, vol. I, pp. 532–539.

[5] Y. Zhang and C. Kambhamettu, "Stereo matching with segmentation-based cooperation," in *ECCV*, 2002, pp. 556–571.

[6] M. Bleyer and M. Gelautz, "A layered stereo matching algorithm using image segmentation and global visibility constraints," *ISPRS Journal of photogrammetry and remote sensing*, vol. 59, no. 3, pp. 128–150, 2005.

[7] K. Yamaguchi, T. Hazan, D. McAllester, and R. Urtasun, "Continuous Markov random fields for robust stereo estimation," in *ECCV*, 2012, pp. 45–58.

[8] K. Yamaguchi, D. McAllester, and R. Urtasun, "Efficient joint segmentation, occlusion labeling, stereo and flow estimation," in *ECCV*, 2014.

[9] A. Ogale and Y. Aloimonos, "Stereo correspondence with slanted surfaces: critical implications of horizontal slant," in *CVPR*, 2004, vol. I, pp. 568–573.

[10] O. Woodford, P. Torr, I. Reid, and A. Fitzgibbon, "Global stereo reconstruction under second-order smoothness priors," *IEEE TPAMI*, vol. 31, no. 12, pp. 2115–2128, 2009.

[11] Y. Ohta and T. Kanade, "Stereo by intra- and inter-scanline search using dynamic programming," *IEEE TPAMI*, vol. 7, no. 2, pp. 139–154, 1985.

[12] C. Rother, V. Kolmogorov, and A. Blake, "GrabCut—interactive foreground extraction using iterated graph cuts," *ACM TOG (Proc. SIGGRAPH)*, vol. 23, no. 3, pp. 309–314, 2004.

[13] M. Bleyer, C. Rother, P. Kohli, D. Scharstein, and S. Sinha, "Object stereo—joint stereo matching and object segmentation," in *CVPR*, 2011, pp. 3081–3088.

[14] H. Ishikawa and D. Geiger, "Rethinking the prior model for stereo," in *ECCV*, 2006, pp. 526–537.

[15] D. Marr and T. Poggio, "A computational theory of human stereo vision," *Proc. Royal Soc. London*, vol. B 204, pp. 301–328, 1979.

[16] K. Nakayama, S. Shimojo, and G. Silverman, "Stereoscopic depth: its relation to image segmentation, grouping, and the recognition of occluded objects," *Perception*, vol. 18, no. 1, pp. 55–68, 1989.

[17] R. Blake and H. Wilson, "Binocular vision," *Vision research*, vol. 51, no. 7, pp. 754–770, 2011.

[18] A. Geiger, M. Roser, and R. Urtasun, "Efficient large-scale stereo matching," in *ACCV*, 2010.

[19] C. Zhang, Z. Li, Y. Cheng, R. Cai, H. Chao, and Y. Rui, "Meshstereo: A global stereo model with mesh alignment regularization for view interpolation," in *ICCV*, 2015, pp. 2057–2065.