Hai Tao and Harpreet S. Sawhney, Global matching criterion and color segmentation based stereo, in *Proc. Workshop on the Application of Computer Vision (WACV2000)*, pp. 246-253, December 2000.

Global Matching Criterion and Color Segmentation Based Stereo

Hai Tao and Harpreet S. Sawhney

Sarnoff Corporation, 201 Washington Rd., Princeton, NJ 08543
{htao, hsawhney}@sarnoff.com

Abstract

In this paper, we present a new analysis by synthesis computational framework for stereo vision. It is designed to achieve the following goals: (1) enforcing global visibility constraints, (2) obtaining reliable depth for depth boundaries and thin structures, (3) obtaining correct depth for textureless regions, and (4) hypothesizing correct depth for unmatched regions. The framework employs depth and visibility based rendering within a global matching criterion to compute depth in contrast with approaches that rely on local matching measures and relaxation. A color segmentation based depth representation guarantees smoothness in textureless regions. Hypothesizing depth from neighboring segments enables propagation of correct depth and produces reasonable depth values for unmatched region. A practical algorithm that integrates all these aspects is presented in this paper. Comparative experimental results are shown for real images. Results on new view rendering based on a single stereo pair are also demonstrated.

1 Introduction

This paper deals with the problem of estimation of dense scene structure using a generalized stereo configuration of a pair of cameras. As is the norm in stereo vision, it is assumed that the intrinsic camera parameters and the exterior pose information are provided. In general, this information can also be derived from images but the focus of this work is on dense 3D extraction. Extraction of dense 3D structure involves establishing correspondence between the pair of images. A variety of methods that rely on image matching under various constraints have been developed in stereo vision. An excellent review of early stereo vision work can be found in [Dhond89].

Stereo matching has to deal with the problems of matching ambiguity, image deformations due to variations in scene structure, delineation of sharp surface boundaries, and unmatched regions due to occlusions/deocclusions in the two images. Typically in order to handle ambiguities in matching, window operations are performed to integrate information over regions larger than a pixel. This leads to the classical matching disambiguation versus depth accuracy trade-off. In areas with sufficient detail, small windows may provide enough matching information but matching over a larger range of depth variations (disparities) may not be possible due to ambiguous matches. In textureless areas small windows are inherently ambiguous. A common strategy for combating this problem is to enforce depth smoothness using techniques such as cooperative/ competitive algorithms [Marr79, Zitnick99], multi-resolution schemes [Hanna93], graph methods [Roy98, Boykov99], and surface model fitting [Hoff89]. However, some of these techniques may introduce excessive smoothness that blurs the depth discontinuities and fails to capture details of the scene such as thin structures. Most of these algorithms face the classic depth smoothness and accuracy tradeoff in one way or the other. Using adaptive windows [Kanade94], more than two views [Okutomi93, Nakamura 96], and non-linear diffusion [Scharstein96] may alleviate this problem to some extent.

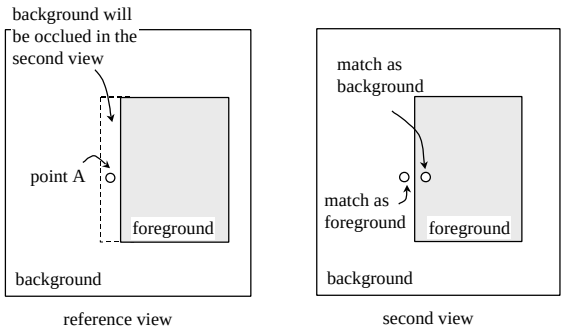


Figure 1. Invalid local matching in an occlusion region.

Invalid matches occur in unmatched regions due to occlusions/deocclusions. As shown in Figure 1, since point A is occluded in the other image, the correspondence induced by the correct depth will have a low matching score because the background is matched against the foreground. Better match will be achieved for spurious depth. This implies that the occlusion regions, ideally, should be recognized in the depth computation process and treated differently. Many smoothness enforcing algorithms either ignore this fact or assume that spurious depths give low matching scores and iteratively detect the occlusion region. Efforts have also been made to detect occlusion boundaries in the preprocessing stage to guide the subsequent matching process. Geiger at al. [Geiger95], Belheumer et al. [Belhumeur96] and Intille et al. [Intille94] define objective functions that account for the co-occurrence of depth boundaries and unmatched regions. However, this is done within a one-dimensional ordering constraint along scan lines under a simplified stereo model. Therefore, typically these algorithms produce jagged surface boundaries since they do not enforce surface smoothness across scan lines. In some cases [e.g. Belhumeur96] 2D smoothness is enforced

as a post-processing step with the scan line algorithm as the main constraint. In [Black99], a generative model of motion discontinuities was formulated and used for modeling and tracking local motion boundaries.

In this paper, we present a comprehensive framework for stereo matching that addresses all the important issues discussed earlier: (1) enforcing global visibility directly in the image space, not based on the local matching scores, (2) obtaining accurate depth boundaries and capturing thin structures, (3) obtaining correct depth for textureless region, (4) hypothesizing correct depth for unmatched regions.

Instead of enforcing global visibility based on local matching scores or along scan lines only, we propose to apply a more basic global matching criterion. It states that if the depth is correct, the image rendered according to the depth into the second viewpoint should be similar to the real view from that viewpoint. This criterion has been exploited recently as a quality evaluation metric for motion and stereo algorithms [Szeliski99]. Morris and Kanade [Morris00] also employed this criterion to find the best triangulation of a set of 3D scene points. In this paper, we apply this principle in stereo computation to enforce global visibility and obtain accurate depth for object boundaries and thin structures.

Inspired by [Black96, Moravec99], we exploit image information such as color segmentation to enforce depth smoothness and delineate sharp depth boundaries. A generally valid heuristic is that within a homogeneous color region the depth variation should be smooth. In this paper, the depth of each homogeneous region is modeled as a nominal plane with allowable additional smooth depth variations. Using this representation, depth smoothness is guaranteed in textureless regions. Furthermore, we present a logical way of deriving reasonable depth for unmatched regions through hypothesizing depth of a given region based on neighboring regions. A practical algorithm that integrates all these principles will be presented in the paper.

The paper is organized as follows. Section 2 explains the main ideas of the approach. Section 3 describes an efficient implementation of the approach. In Section 4, comparative experimental results are shown for real images. New view rendering results based on a single image pair are shown. Discussions and future work are presented in Section 5.

2 Approach

2.1 Global matching criterion

The proposed approach adopts an analysis-by-synthesis strategy that tests the goodness of a given depth map as a whole. The main idea is that if a depth map is good, warping the reference image to the other view according to this depth will render an image that matches the real view. In the depth based warping process, visibility is enforced by techniques such as Z-buffering or occlusion compatible ordering [McMillan95]. Instead of inferring depth indirectly from local matching measures, the depth map is found through searching in the solution space based on direct global image similarity measures.

Two immediate concerns regarding this approach are the huge solution space and the expensive synthesis process. For an image with N pixels, suppose each pixel may have d different quantized depth values, the total number of different possible depth maps is d^N. An exhaustive search will warp each of these configurations and find the best configuration as the solution. This also indicates that stereo vision is an NP-hard problem in general. Polynomial algorithms such as space carving and graph methods simplify the search space by either imposing an order or by localizing the solution space. For example, in space carving methods, the algorithm is polynomial due to the sequential peeling process. If a wrongfully carved voxel is to be put back, the carving process has to restart from that voxel again and this becomes an NP-hard problem. In dynamic programming methods [Geiger95, Belhumeur96], a left-right ordering is enforced to make the algorithm polynomial.

The second issue with the proposed approach is the cost of synthesis in every iteration. Even if the solution space is linear in the number of pixels, say $0.5N$, it is still computationally impractical to warp the reference image $0.5N$ times to find the best depth map.

We solve with the first problem by combining a color segmentation based representation and neighborhood depth hypothesizing method in a local search algorithm. More important than computational considerations, this approach enforces depth smoothness in homogeneous color regions and also makes it possible to infer reasonable depth for unmatched regions. In Section 3, a fast algorithm is presented which performs less than 10 image warps for each iteration of depth computation.

2.2 Color segmentation based depth representation

One of the difficult tasks in many existing stereo algorithms is to find correct depth in textureless regions. Because of the matching ambiguity involved in these regions, the depth map created by picking the best matching score is usually noisy. An observation employed in the proposed approach is that for a region with homogenous color, usually there is no large depth discontinuity. This observation implies a depth representation based on segmenting the reference image into homogeneous color regions. In this paper, we propose a plane plus residual disparity representation for each color segment. More specifically, in each color segment, the depth surface is modeled as a plane surface plus small depth variations for each pixel.

This model guarantees smoothness in textureless regions. For smooth but textured regions, where many small segments are present, smoothness is not enforced across segments. However, depth estimation tends to be reliable in these areas even without the smoothness constraint.

It is to be emphasized that the color segmentation is not an end goal of this work. Over-segmentation of smooth surfaces is tolerated. We rely on the generally valid heuristic that depth boundaries coincide with color segmentation

boundaries. Association of color segments with semantic/object regions is not attempted as in general color segmentation works. A way of initializing the representation for each segment is to compute an image-disparity based local matching volume with each voxel containing the local matching score. Then find the best match for each pixel in a segment and fit a plane. A simple recursive algorithm adjusts the plane recursively. The depth initialization problem will be described in more details in Section 3.2.

2.3 Hypothesizing Neighborhood Depths

Given a current depth map and the associated segmentation, errors in depth may be eliminated by a process of local search in the space of segments instead of in the space of pixels. Local search hypothesizes the depth of each segment to be that of the neighboring segments and tests the hypothesis by rendering. This process will help the correct depth to propagate because by hypothesizing the correct depth, the warped image induces better matching. For example, in Figure 2, the depth of a background segment is wrongfully computed as the foreground depth because of the propagation of depth from a nearby textured foreground region. However, the error can be corrected if the background segment is hypothesized to have the depth of the correct neighboring background segment and that hypothesis wins. Another benefit of the hypothesizing depths in neighborhoods is that it helps to derive reasonable depth for unmatched regions. For unmatched regions, the depth is more likely to be the extension of the neighboring background segment (Figure 3).

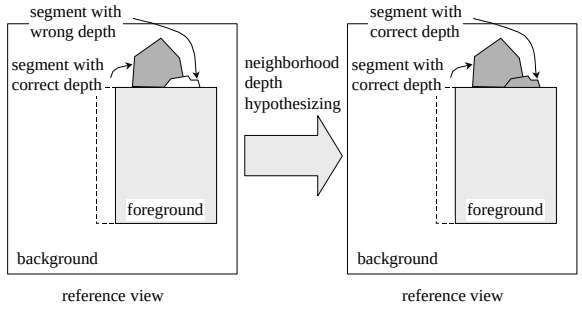


Figure 2. Correct depth is propagated to reduce the fattening effect.

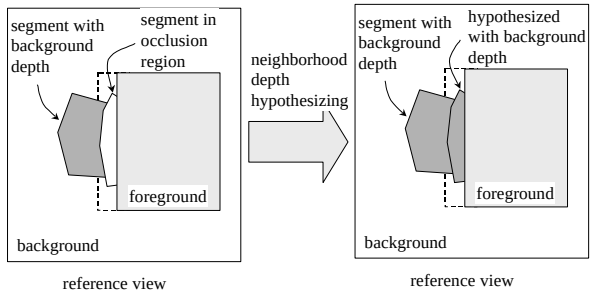


Figure 3. Background depth is hypothesized for an unmatched region.

2.4 Greedy searching

Suppose that there are s segments in the reference image and each segment has k neighboring segments. Then the number of depth hypotheses for each segment is $k+1$ (one hypothesis is from itself). The solution space for depth map is $(k+1)^s$. For example, if there are 1000 segments in an image and each segment has 5 neighboring segments, the solution space is $(5+1)^{1000}$.

Due to the NP-hard nature of the problem, we have no other choice but to find a locally optimal solution. A straightforward local greedy search algorithm is proposed. In this algorithm, we test all the neighboring depth hypotheses of each segment while all other segments are kept fixed. The neighborhood depth hypothesis that gives the best global matching score is recorded. After all segments have been tested, their depth is updated by choosing from the initial depth and the best neighborhood hypothesis according to the matching scores. This process is performed iteratively until either the total number of segments with depth changes is small or the number of iterations exceeds a certain value. We found experimentally that this process can tolerate large initial depth errors. The complexity is reduced to $sk+1$ per iteration: sk tests for k neighborhood hypotheses of s segments and one additional test for the initial depth. It should be noted here that in practice, the numbers of neighboring segments might vary for individual segments. The counting arguments presented here are just for illustrating the idea.

3 Algorithm and implementation

Summarizing the previous sections, a new stereo vision algorithm is proposed as follows.

Step 1: Color segmentation of the reference image

Step 2: Initial depth estimation for each segment.

Step 3: Hypothesizing neighborhood depth for each segment to generate a depth map solution space.

Step 4: Test all hypotheses of each segment using the global matching criterion with the depths of the other segments fixed.

Step 5: Batch updating of depth by choosing the best hypothesis for each segment.

Step 6: Repeat Step 3,4,5 until depth change is small or a certain number of iterations have been performed.

This section will elaborate on each step.

3.1 Color segmentation

Any algorithm that decomposes an image into homogeneous color regions will work for our purpose. The most important parameter in the algorithm is the threshold for splitting a region into multiple sub-regions. If this value is small, the image can be over-segmented. If this value is large, the

image is under-segmented. Since our algorithm enforces the depth continuity inside each segment strictly, under-segmentation should be avoided. For our implementation, we used the method proposed in [Comaniciu97]. All the tested images shown were segmented using a single set of parameters.

3.2 Initial estimation of depth representation

The three steps for the initial depth representation are (i) computing matching scores in an image-disparity volume, (ii) plane fitting in each segment, and (iii) residual disparity computation in each segment.

3.2.1 Local matching in image-disparity volume

For the standard (parallel) stereo setup, the correspondence of a point in the second view lies on the same scan line as the reference view. The horizontal displacement of the corresponding point is called disparity. Similarly, for any arbitrary two views, the matching point lies on the epipolar line in the second image. For a standard stereo setup, to compute the dense point correspondences, matching scores in an image-disparity volume are first computed. More specifically, the matching scores for all possible horizontal displacements (within a range and with a fix displacement interval) are computed first. This forms a three-dimensional matching score array, which we call image-disparity matching volume. Each cell (x, y, d) holds the matching score for the correlation between pixel (x, y) in the reference image and (x+d, y) in the second image. Then, for each pixel, the best score is picked and the corresponding displacement is transformed into depth. The same idea is applied to arbitrary views, except that the formulation is more complicated (See Appendix A). In both cases, the iso-disparity surface is a frontal plane in the reference view.

3.2.2 Plane fitting in each segment

Once the image-disparity matching volume is computed, a plane is fitted for each color segment. We first find the best depth value for each pixel in the segment and then compute the best fit plane to the depth values. More specifically, if the plane equation in each segment is given by:

$$\tilde{Z}_p = \frac{1}{Z} = ax + by + c$$

where (x, y) is an image point, and Z is its depth in the reference camera coordinate system. Then, a, b, c are the least squares solution of a linear system

$$A[a, b, c]^t = B$$

where each row of A is the $[x, y, 1]$ vector for a pixel and each row of B is its corresponding $\frac{1}{Z}$.

An iterative fitting process is adopted to reduce the effect of outliers. The idea is illustrated in Figure 4. First, the depth of every pixel is decided by picking the best matching score. Then, a plane is fitted in a segment. In the next iteration, the depth of each pixel is chosen within a given range of the fitted plane by finding the best matching score in that range. The plane parameters are updated accordingly based on this depth. This process iterates several times until the plane parameters do not change significantly. This process is particularly useful for fitting planes in large textureless regions where matching ambiguities occurs. More generally, any other robust method of plane fitting like M-estimation, least median squares or RANSAC may be employed.

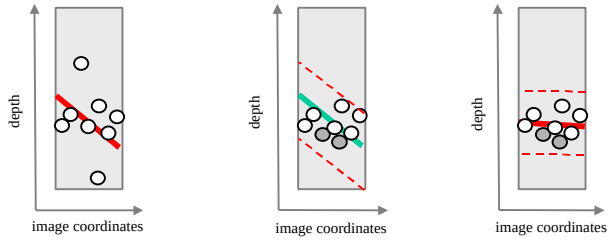


Figure 4. The iterative plane fitting process.

3.2.3 Residual disparity computation

Our representation allows small variations from the planar model in each segment. The actual depth of each pixel is

$$\frac{1}{Z} = \tilde{Z}_p + \tilde{Z}_r$$

Once the plane parameters are determined, for each pixel, $\tilde{Z}_p$ is known. $\tilde{Z}_r$ is computed by locating the best match in the image-disparity volume within a small range of $\tilde{Z}_p$. Residual disparity $\tilde{Z}_r$ is smoothed in each segment to obtain the initial color segmentation based depth representation.

3.3 Depth hypotheses generation and testing

To generate a hypothesis for a segment from a neighboring segment, the plane parameters are replaced using those of the neighboring segments. Then residual disparity for each pixel is found by searching around the plane and smoothing within the segment.

As mentioned in Section 2.4, we test the depth hypothesis of a single segment each time while all the other segments maintain the initial depth. There are total of $sk+1$ tests. The depth representations are updated after testing is done for all segments. Since only the depth of one segment is changed each time, only a small portion of the image needs to be tested. A fast algorithm takes advantage of this fact and performs equivalently only $k+1$ image warps. The algorithm is described as follows.

The reference image is first warped to the second view using the initial depth (i.e. initial depth hypothesis for each segment). We call this image the base warp (Figure 5b). In our implementation, all warping operations are done using the post rendering 3D warp method described in [Mark97].

Now if the depth of segment *s* is replaced by one of its neighborhood hypothesis, to compute its matching measure, we only need to consider those pixels affected by the depth change. For example, in Figure 5c, the depth of segment *s* is

changed. In the warped image, region B of segment *q* becomes visible while segment *s* becomes invisible. The matching score of the new depth map is computed by adding matching score of region B to the base warp score and subtracting matching score of segment *s*. This example suggests a fast algorithm for testing hypotheses. In this algorithm, for the base warp, for each pixel, the warped depths, the segmentation IDs, and the matching scores of the two top-most layers are stored. In hypotheses testing, changes in the matching scores over base warp are computed by adding the matching scores of pixels that become visible and subtracting scores of pixels that become invisible. Since for each test, we only change the depth of one segment, only the two top-most layers may become visible and information regarding these two layers should be recorded. The third layer will be blocked by at least one of the two layers originally in front of it and always invisible, therefore it does not affect the matching score.

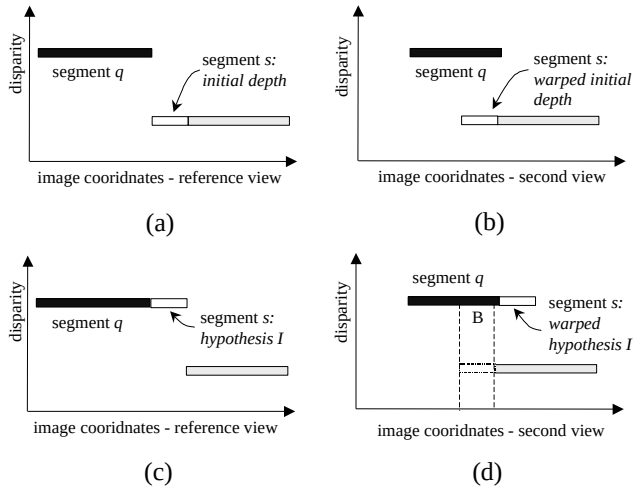


Figure 5. Fast hypotheses testing. (a) Initial depth in the reference view. (b) Warp the reference image to the second view using the initial depth produces the base warp. (c) Hypothesize the depth of segment *s*. (d) The warped image based on depth in (c). The new matching score is computed by adding matching score of region B on segment *q* to the based warp score, and subtracting matching score of segment *s*.

The algorithm is described in more details as follows.

Step 4.1: Base warp: Forward warp the reference image according to the initial depth to the second view. In the warped image, for each pixel i, record information for the two layers closest to the camera center. The information includes their depths d_{i1}, d_{i2}, the corresponding segments in the reference image s_{i1}, s_{i2}, and the matching scores c_{i1}, c_{i2}. Quantities with subscript $i1$ belong to the top-most layer (i.e. with smallest depth). If less than two layers presents at a point, the absent segment IDs and depths are undefined and the matching scores are set to zeroes. In our implementation, if I and I' denote the image intensities of a pixel in the second view and the warped image, the matching score is computed as $c = \exp\{-(I - I')^2 / 2s^2\}$, where s is a constant.

Step 4.2: For the *j*th neighborhood depth hypothesis in each segment s

4.2.1. Forward warp the segment using the depth hypothesis. For the *i*th pixel in the warped segment, let d_i and c_i denote its warped depth and its matching score

4.2.2. Compute the improvement value T_{sj} of the segment over the base warp as

$$T_{sj} = -C_s + \sum_{i \in s} sim(d_i, c_i, s, d_{i1}, c_{i1}, s_{i1}, d_{i2}, c_{i2}, s_{i2})$$

where the matching measure difference contributed by segment *s* in the base warp is

$$C_s = \sum_{s_{i1}=s} (c_{i1} - c_{i2})$$

The matching measure difference contributed by the *i*th pixel of segment *s* for various ordering cases according to the depth hypothesis is

$$sim(d_i, c_i, s, d_{i1}, c_{i1}, s_{i1}, d_{i2}, c_{i2}, s_{i2}) = \begin{cases} c_i - c_{i2} & s_{i1} = s \wedge d_i < d_{i2} \\ 0 & s_{i1} = s \wedge d_i \geq d_{i2} \\ c_i & s_{i1} = s \wedge s_{i2} = undefined \\ c_i - c_{i1} & s_{i1} \neq s \wedge d_i < d_{i1} \\ 0 & s_{i1} \neq s \wedge d_i \geq d_{i1} \\ c_i & s_{i1} = undefined \end{cases}$$

Step 5. Update the depth of each segment using the hypothesis with the best positive improvement. If none of the hypotheses gives positive improvement, keep the initial depth.

4 Experiments

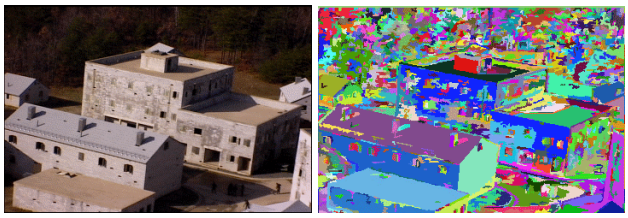
The proposed algorithm has been implemented on a PC platform. For a 360×240 reference image, the depth is computed within 30 seconds after the computation of the initial image-disparity matching volume. Some preliminary tests have shown promising results. Results on two real test image pairs are described in this section. These results are analyzed qualitatively by inspecting the depth map and by generating new views around the reference view. In the latter case, errors in the depth map will cause visual artifacts that are immediately noticeable.

4.1 Buildings

This data set contains two frames of a video sequence. The video was shot from a helicopter when it flew around a staged suburban area with buildings and woods. The video camera is calibrated. The relative camera pose of the two video frames is computed using a method described in [Sawhney 99].

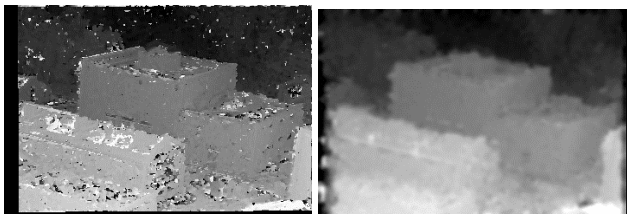
In Figure 6a, the reference image is shown. Figure 6b shows the segmentation result on the reference view. Though over-

segmented, it is good enough for our algorithm because all depth discontinuities are coincident with image boundaries. In Figure 7a, the initial depth map computed from the image-disparity matching volume is shown. This depth map is obtained by choosing the best matching score for each pixel. This is akin to the output of a standard window based stereo algorithm. The depth is very noisy on the textureless walls of the buildings and the roads. Depth bleeding (fattening) phenomenon is observed around the boundaries of the buildings. In Figure 7b, the depth map generated by a direct method [Hanna93] is shown. The noise in textureless regions has been suppressed significantly due to the coarse-to-fine incremental scheme. However, over-smoothing can be observed around depth boundaries. The building in the depth map is larger than the actual building in the reference image because of the fattening effect.

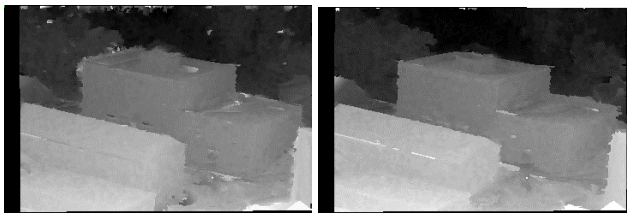


(a) (b)

Figure 6. The image pair *Buildings*. (a) The reference view. (b) Reference view color segmentation.



(a) (b)



(c) (d)

Figure 7. The computed depth maps. (a) Depth map from standard window based correlation method. (b) Depth map generated from a direct method. (c) The initial depth representation. (d) The final depth map.

The results using the proposed algorithm are shown in Figure 7c and 7d. In Figure 7c, the initial plane-plus-residual-disparity representation is shown. It should be noted that in this representation, a large amount of noise has been canceled by iterative plane fitting, and the depths of many textureless regions are correct. However, the fattening in many regions persists. This is noticeable especially among the small segments around the foreground objects. The reason is that for these regions, most initial depths are wrongfully computed as the foreground depth, therefore, the plane fitting can not correct the problem. Similar error can be observed for unmatched regions with small segments. Many of these problems are corrected in Figure 7d, which is the result after neighborhood depth hypothesizing. Furthermore, the depths of unmatched regions are correctly hypothesized. Compared to the results in Figure 7a and Figure 7b, the new depth map has more accurate depth boundaries without the loss of depth smoothness in textureless regions.

Figure 8. Two new views rendered by forward warp the reference image according to its corresponding depth map.

The correctness of the depth map is also verified by new view rendering. New views close to the reference view are rendered based on the reference image and the depth map. Animating these views enables indirect visual inspection of the depth map. Artifacts such as depth noise in textureless regions produce spikes hovering in the 3D space. Fattening artifacts cause the background texture to get attached to the foreground objects. In Figure 8, two rendered images are shown based on the reference image and the depth map shown in Figure 7d. It is observed that overall the depth map produces correct rendering. In particular, boundaries of buildings are accurately rendered.

It should be mentioned that the black regions next to the buildings are deocclusion regions that are not seen in the reference view. Since we generate these images from a single reference image, there is no texture information available for these regions.

4.2 Lady

The second image pair are taken using a standard stereo setup. In Figure 9, the reference image and the corresponding color segmentation are shown. The scene in this stereo pair contains objects of more irregular shapes and some thin structures such as the stems of the sunflowers. The widths of stems range from 1 to 3 pixels and their disparities are larger than 10 pixels.

As shown in Figure 11b, the direct method fails to capture the depth of some stems. The reason is that large disparity can only be captured at lower resolutions in a coarse-to-fine algorithm. However, for thin structures, the image features disappear at low resolutions.

Figure 11a shows the depth map computed using the simple window-based correlation method. The depths in a smooth region (upper-right corner) and the depths of the stems are noisy. In both depth maps, the fattening effect thickens the stems significantly and makes the foreground objects larger than their actual sizes.

(a) (b)

Figure 9. The image pair *Lady*. (a) The reference view. (b) Color segmentation.

(a) (b)

Figure 10. Two new views rendered by forward warping the reference image according to its depth map.

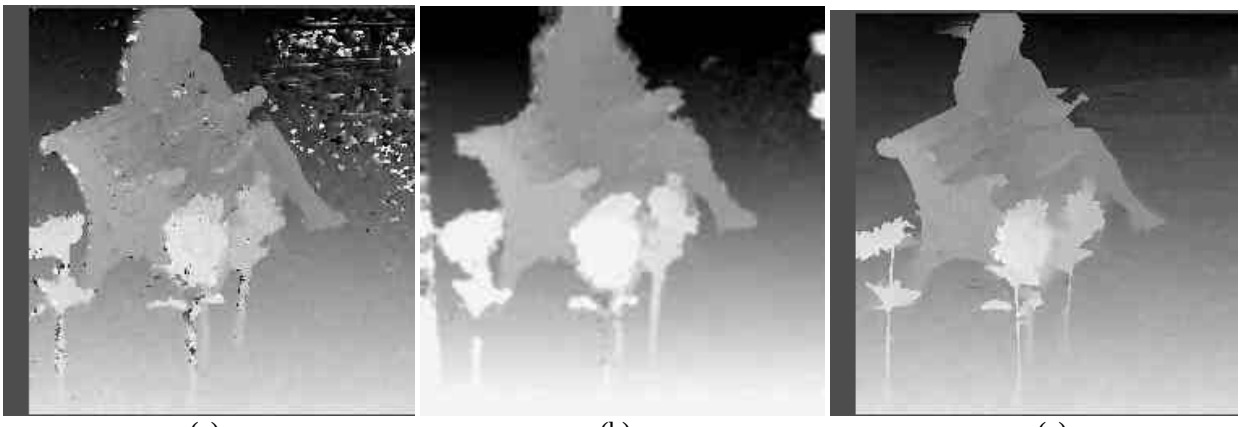
Figure 11c shows the depth map generated by the proposed algorithm. The object boundaries are more accurately estimated. The depths and the boundaries of the flower stems are correctly computed. Background depths are hypothesized for unmatched regions on the left side of the foreground object. The depths in the textureless regions are also correctly estimated. Figure 10 shows two rendered images from views around the reference view. Correct depth and accurate object boundaries are perceived from the animated new views.

5 Discussions and conclusions

A new analysis by synthesis stereo vision framework has been presented in this paper. In contrast to the traditional two-step local matching based approaches, a local hypothesize and test method is exploited. This approach includes global visibility check in the synthesis process and avoids invalid local matches. By adopting a color segmentation based plane-plus-residual-disparity depth representation, the smoothness constraint in textureless regions is enforced without sacrificing accuracy on depth discontinuities. This representation also dramatically reduces the depth solution space. Neighborhood depth hypothesizing method has further reduced the solution space and makes it possible for correct depth to propagate and depth of unmatched regions to be speculated in a principled way. A fast searching method has been proposed to find the local solution in linear time. We found that with reasonable initial depth representation, this algorithm converges to a local optimal solution. Promising results on real images have been obtained and demonstrated in this paper.

The proposed algorithm in this paper is a particular implementation of exploiting the proposed main principles. Many aspects need to be investigated more thoroughly. The plane plus small residual disparity representation may not be sufficient for objects with curved surfaces. A more flexible depth representation will solve this problem. In real images, it may occasionally occur that depth boundaries appear in homogeneous color segments. The current method is unable to handle this situation. A method of hypothesizing splits in problematic segments seems to be a reasonable cure. The accuracy of forward warping affects the final depth map significantly. We are investigating more accurate mesh based methods to improve the current algorithm.

A final note is that this approach is valid for flow computation too. We only need to replace the depth based forward warping by occlusion compatible traversal warping. Where of course, information such as positions of epipoles needs to be roughly known.



(a) (b) (c)

Figure 11. Depth maps computed using (a) standard window based correlation method, (b) a direct method (c) the proposed algorithm.

Acknowledgments

We thank IMAX Corporation for making their images available for our experiments (Figure 9,10,11).

References

[Belhumeur96] P. N. Belhumeur, "A Bayesian-approach to binocular stereopsis," *Intl. Journal of Comp. Vision, vol. 19, no. 3*, pp. 237-260, August 1996.

[Black96] M. J. Black and A. Jepson, "Estimating optical flow in segmented images using variable-order parametric models with local deformations," *IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 18, no. 10*, Oct. pp. 972-986, 1996.

[Black99] M. J. Black and D. J. Fleet, "Probabilistic detection and tracking of motion discontinuities," in *Proc. International Conf. on Computer Vision*, pp. 551-558, Sept. 1999.

[Boykov99] Y. Boykov, O. Veksler, and R. Zabih, "Fast approximate energy minimization via graph cuts," in *Proc. International Conference on Computer Vision*, Sept. 1999.

[Comaniciu97] D. Comaniciu and P. Meer, "Robust analysis of feature spaces: color image segmentation," in *Proc. of IEEE conference on Computer Vision and Pattern Recognition*, pp. 750-755, 1997.

[Dhond89] U. R. Dhond and J. K. Aggarwal, "Structure from stereo: a review," *IEEE Transactions on System, Man, and Cybernetics, vol. 19, no. 6*, pp. 1489-1510, 1989.

[Geiger95] D. Geiger, B. Ladendorf, and A. Yuille, "Occlusions and binocular stereo," *Intl. Journal of Comp. Vision, vol. 14*, pp. 211-226, 1995.

[Hanna93] K. J. Hanna and Neil E. Okamoto, "Combining stereo and motion analysis for direct estimation of scene structure," in *Proc. Intl. Conf. on Computer Vision*, pp 357-265, 1993.

[Hoff89] W. Hoff and N. Ahuja, "Surfaces from stereo: integrating feature matching, disparity estimation, and contour detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 11, no. 2*, pp. 121-136, February 1989.

[Intille94] S. S. Intille and A. F. Bobick. "Disparity-space images and large occlusion stereo," in *Proc. European Conference on Comptuter Vision*, J-O Eklundh (ed.), Stockholm, Sweden, Vol. 801, pp. 179-186. May 1994.

[Kanade94] T. Kanade and M. Okutomi, "A stereo matching algorithm with an adaptive window: theory and experiment," *IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 16, no. 9*, pp. 920-932, September 1994.

[Mark97] W. R. Mark, L. McMillan, and G. Bishop, "Post-Rendering 3D Warping," in *Proc. of 1997 Symposium on Interactive 3D Graphics*, pp. 7-16, Providence, Rhode Island, April 1997.

[Marr79] D. Marr and T. Poggio, "A computational theory of human stereo vision," *Proceedings of the Royal Society London B, 204*, pp. 301-328, 1979.

[McMillan95] L. McMillan, "Computing visibility without depth," UNC Technical Report TR95-047, University of North Carolina, 1995.

[Moravec99] K. Moravec, R. Harvey, J. A. Bangham, and M. Fisher, "Using an image tree to assist stereo matching," *ICIP99*, 1999.

[Morris00] D. D. Morris and T. Kanade, "Image-consistent surface triangulation," in *Proc. IEEE conference on Computer Vision and Pattern Recognition*, vol. 1, pp. 332-338, Hilton Head, SC, June 2000.

[Nakamura 96] Y. Nakamura, T. Matsura, K. Satoh, and Y. Ohta, "Occlusion detectable stereo - occlusion patterns in camera matrix," in *Proc. of IEEE conference on Computer Vision and Pattern Recognition*, pp. 371-378, 1996.

[Okutomi93] M. Okutomi and T. Kanade, "A multiple-baseline stereo," *IEEE Transactions on Pattern Analysis and Machine Intelligence vol. 15, no. 4*, pp. 353-363, April 1993.

[Roy98] S. Roy and I. J. Cox, "A maximum-flow formulation of the N-camera stereo correspondence problem", in *Proc. Int. Conf. on Computer Vision (ICCV'98)*, Bombay, India, January 1998.

[Sawhney99] H.S. Sawhney, Y. Guo, J. Asmuth, and R. Kumar, "Multi-view 3D estimation and applications to match move," in *Proc. of the IEEE Wkshp. on Multi-view Modeling & Analysis of Visual Scenes*, Fort Collins, CO, USA, June 1999.

[Scharstein96] D. Scharstein and R. Szeliski, "Stereo matching with non-linear diffusion," in *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'96)*, pp. 343-350, San Francisco, California, June 1996.

[Shashua96] A. Shashua and N. Navab, "Relative affine structure: canonical model for 3D from 2D geometry and applications," *IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 18, no. 9*, pp. 873-883, 1996.

[Szeliski99] R. Szeliski, "Prediction error as a quality metric for motion and stereo," in *Proc. Seventh International Conference on Computer Vision (ICCV'99)*, pp. 781-788, Kerkyra, Greece, September 1999.

[Zitnick99] C. L. Zitnick and T. Kanade, "A cooperative algorithm for stereo and occlusion detection," CMU-RI-TR-99-35, October, 1999.

Appendix A

For any two arbitrary views with known camera parameters, the image-disparity volume in the reference view can be formed as follows (the notations follow [Shashua96]).

Assume the calibration matrix for the reference view and the second view are M and M'. The 3D transform between the two view is $X' = RX + T$. The epipole in the second image is $v' = M'T$.

If $p = [x, y, 1]^T$ denotes the image of a 3D point in the reference view and Z is its depth in the reference coordinate system, then its image $p' = [x', y', 1]^T$ in the second view is projectively equal to

$$\begin{aligned} p' &\cong M'RM^{-1}pZ + M'T \\ &\cong M'RM^{-1}p + M'T\frac{1}{Z} \\ &\cong M'RM^{-1}p + v'\frac{1}{Z} \end{aligned}$$

Because $Z \in [0, +\infty]$, hence $1/Z \in [+\infty, 0]$. Suppose in a scene, $a \le 1/Z \le b$, quantize this range into l levels with interval $d_z = (b-a)/(l-1)$. The image-disparity volume is parameterized as (x, y, l) with $p' \cong M'RM^{-1}p + v'ld_Z$.