34th Annual International Conference of the IEEE EMBS
San Diego, California USA, 28 August - 1 September, 2012

Phosphene Vision of Depth and Boundary from Segmentation-based Associative MRFs

Yiran Xie, Nianjun Liu and Nick Barnes

Abstract— **This paper presents a novel low-resolution phosphene visualization of depth and boundary computed by a two-layer Associative Markov Random Fields. Unlike conventional methods modeling the depth and boundary as an individual MRF respectively, our algorithm proposed a two-layer associative MRFs framework by combining the depth with geometry-based surface boundary estimation, in which both variables are inferred globally and simultaneously. With surface boundary integration, the experiments demonstrates three significant improvements as: 1) eliminating depth ambiguities and increasing the accuracy, 2) providing comprehensive information of depth and boundary for human navigation under low-resolution phosphene vision, 3) when integrating the boundary clues into downsampling process, the foreground obstacle has been clearly enhanced and discriminated from the surrounding background. In order to gain higher efficiency and lower computational cost, the work is initialized on segmentation based depth plane fitting and labeling, and then applying the latest projected graph cut for global optimization. The proposed approach has been tested on both Middlebury and indoor real-scene data set, and achieves a much better performance with significant accuracy than other popular methods in both regular and low resolutions.**

I. Introduction

Stereo matching has been one of the most intensively investigated research topics as it is useful in a variety of applications such as scene reconstruction and navigation. Based on different representations of depth estimation, existing methods can be sorted into two categories: pixel-wise and segment-wise. Pixel-based algorithms often suffer from local noises and have insufficient cues of the scene. As people generally identify the object and reconstruct the scene by partitioning the scene into a set of groups each with the same or similar visual features such as color or texture, researchers have developed segment-based algorithms upon the similarity.

Segment-based algorithms [1][2] have dominated the Middlebury Benchmark [3] due to their good performance on reducing ambiguity of disparities in textureless regions. They usually share the assumption that the scene structure can be approximated by a set of non-overlapping visually homogeneous regions where each region corresponds to its own depth surface. In other words, all pixels in the same segment should lie on the same depth surface and discontinuities only occur on boundaries. This assumption certainly enhances the tolerance of local noise as the depth surface is now decided by a group of pixels, the risk of assigning incorrect disparities to occluded or textureless individual pixels is decreased. Typical procedures for these approaches are as follows: first, segmenting the reference image using color-based segmentation and getting an initial disparity by doing pixel-based local match; then fitting disparity planes to every segment using plane fitting techniques; finally the optimal assignment of planes is approximated by using global-based optimization tools to minimize a certain energy function.

However, with segments being purely grouped on visual features, they are still likely to be influenced by local noises. Imagining a piece of colorful newspaper lying on a planar table. Clearly the newspaper should locate on the same planar depth surface. However in segment-based algorithms, every individual character and color region may be segmented into different sized segments. Segment-based approaches usually are not concern with the dimension of the segment, and simplify each segment as an individual node in the model for further optimization. Therefore robustness will not be guaranteed due to the existence of these small segments. Certainly, it may be regularized by adding smoothness interaction between neighboring segments, but the parameter of smooth scale is always hard to tune. If the parameter is too small, these small segments will not be as consistent as desired, but if too large it will lead to undesired blurring along surface boundaries because the neighboring segments that actually cross the surface boundaries are smoothed as well. An alternative solution is to introduce depth surface boundaries to distinguish the smoothness of neighboring segments along the surface boundaries. An experiment motivates us is that given perfect or near perfect surface boundaries, state-of-the-art results can be achieved by over-smoothing segments within the same depth surface.

In the paper, we investigate a novel approach of depth computation and then down-sampled in low resolution, which is crucial for some specific applications for artificial visual simulation[4][5][6]. Under the present hardware limitation of low-vision devices, the depth must be down-sampled to a qualified low-resolution. Apparently, some popular image resizing methods(nearest-neighbor, bilinear, cubic and so on) will be the straight-forward solution, but they may bring some serious distortions into the results in which the surface boundaries are blurred, and depth of foreground merges into background. This is partially due to the equally treatment of boundary regions and no boundary regions. Down-sampling within the same depth surface is straightforward and easy to implement, but boundary regions should be handled carefully. Therefore, surface boundary can

*This work was supported by

Yiran Xie, Nianjun Liu and Nick Barnes are with College of Engineering and Computer Science of Australian National University and Canberra Research Laboratory of National ICT Australia.(email:{yiran.xie, nianjun.liu, nick.barnes}@nicta.com.au)

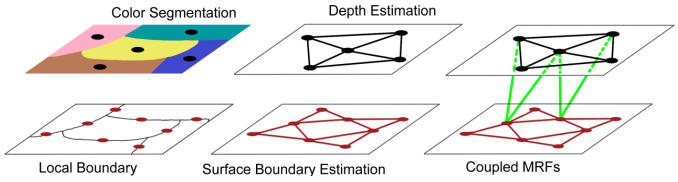

Fig. 1. The proposed two-Layer MRFs Framework. We use color segmentation as our inputs. For depth estimation in the upper-layer, every segment is modeled as one node(black). For surface boundary estimation in the downer-layer, boundaries are further broken into piecewise ones(red). The green lines are the connection between two layers. For simplicity, here only draws the two-layer connections(green) of two boundary nodes.

be used as clues into down-sampling process.

Experiments demonstrates our novel approaches could provide 1)significant improvements by eliminating depth ambiguities and increasing its accuracy, 2) explicit clues of depth and boundary for human navigation under low-resolution phosphene vision, 3) foreground obstacles are clearly discriminated from surrounding background by integrating boundary clues into downsampling process.

II. Proposed Method

The above challenges motivate us to integrating depth surface boundary estimation into the existing stereo matching framework so that these two-layer variables could be inferred together and interact each other. Inspired by Ren's work[7], we use one layer of Markov Random Field (MRF) to model the connectivity of locally found edges, but instead of using *constrained Delaunay Triangulation* to approximate the edges, we novelly break boundaries into pieces so that two neighboring segments will only have one unique boundary piece between them. And we treat such boundary pieces as individual variables in the boundary layer of associative MRFs. The connection between boundary nodes are simplified from higher-order to pairwise relationship due to computational purpose. After these two layers are modeled separately, we align and associate two layers based on the topological structure. An example is given in Figure 1. Under such modeling, a two-layer MRFs is built where one layer represents depth and the other represents surface boundary.

Along with surface boundaries determined dynamically, smoothness scaling between segments can be decided as need, and will only apply within surface boundaries. In some sense, it can be seen that segments are grouped dynamically according to boundaries and in some sense segments are re-partitioned dynamically. And both surface boundary and depth obtained simultaneously facilitates further recognition and scene understanding.

Generally, optimizing such framework is quite complex and challenging. The third-order interaction between two layers makes standard graph cut approach difficult to apply. Also widely existing loops will lengthen the time taken to converge in message-passing algorithms. Thanks to the latest projected graph cut[8], it minimize the energy by making projected moves iteratively, in which it fixes one layer of MRFs at a time, and uses ST-min cut[9] to optimize

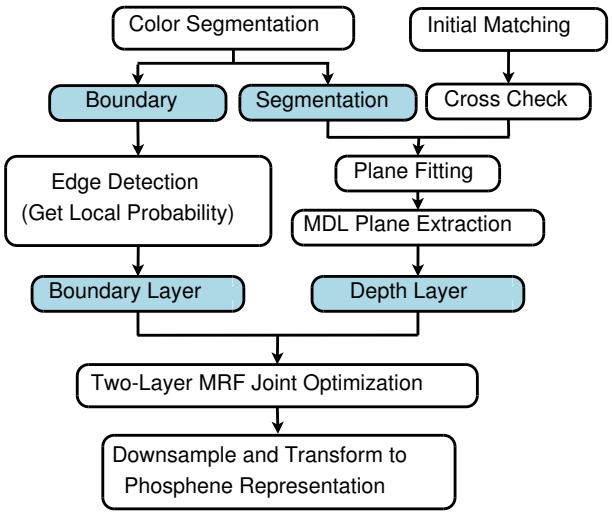

Fig. 2. Overview of the proposed approach. The frames in white are the processes, frames in blue are the intermediate results.

the other layer. It converges when no lower energy can be reached. Experiments has demonstrated such inference algorithm performs well in our application.

The flow chart of our proposed approach is shown in Figure 2.

A. Typological Structure of two-layer MRFs

In our joint framework, we have two-layer MRFs that one layer represents depth and the other represents surface boundary. More formally, two sets of variables are used, X for depth and Y for boundary. Let $L_x = \{1, 2, ..., n\}$ be a set of n different discrete depth plane labels, and $L_y = \{0, 1\}$ be a two-variable set for the labels of surface boundary in which 0 is off and 1 is on. The task is to find a labeling configuration f that allocates the labels from L_x to each variables $X_i \in X$ and L_y to each $Y_i \in Y$ respectively. Then each possible labeling f has its own *posterior* probability, the goal is to find the f^* that has the maximum probability. According to the Hammersley-Clifford theorem, maximum a *posterior* labeling f^*(MAP) is equivalent to the minimum of the Gibbs energy. We define the proposed energy function as:

$$E = \underbrace{E^S(x)}_{Stereo} + \underbrace{E^B(y)}_{Boundary} + \underbrace{E^I(x,y)}_{Interaction} \quad (1)$$

Note that the energy function not only contains energy terms for stereo matching and boundary estimation alone but also has energy term describing their interactions.

The first stage of the proposed approach is color segmentation[10] on the reference image. For stereo matching, every segment is taken as an individual depth node disregard of their sizes. And for each pair of neighboring segments, define their unique piece of boundary connection as one boundary node. An illustration of this process is given in Figure 1. Note that the boundaries are actually between pixels, and do not occupy any pixels themselves.

B. Surface Boundary Potentials

The energy potentials for surface boundary estimation is defined as

$$E^B(y) = \psi_i^B(y_i) + \psi_{ij}^B(y_i, y_j) \tag{2}$$

$$\psi_i^B(y_i) = \sum_{y_i \in Y} (1 - pb_i) \cdot y_i \tag{3}$$

The unary term $\psi_i^B(y_i)$ only penalizes when the boundary y_i chooses to appear. The lower its local probability pb_i is, the higher penalty it takes. To capture pb_i, we apply probability of boundary detector[11] on the reference image and normalize the result into $[0, 1]$, so every pixel will have a probability value. For every boundary y_i, its pb_i is given as the average of the pixels' probabilities it passing through.

The pairwise term $\psi_{ij}^B(y_i, y_j)$ encourages two connecting boundaries to be both turned on or turned off. It takes the form of the Potts model:

$$\psi_{ij}^B(y_i, y_j) = \sum_{y_i, y_j \in N} \begin{cases} 0, & if\ y_i = y_j, \\ 1, & otherwise. \end{cases} \tag{4}$$

C. Stereo Matching Potentials

Firstly, we employ the fast local pixel dissimilarity measure[12] to construct the correlation volume for both left and right images as the reference image. Secondly, we apply mutual consistency check on the result. Pixels passing it will be labeled as $stable\ pixel$ and if the percentage of $stable$ members of a segment exceeds a certain threshold, then the segment will be labeled as $stable\ segment$. Thirdly, a segment-based RANSAC plane fitting is carried out on $stable\ pixels$ inside each $stable\ segment$, the plane with the least error will be put into label set L_x. If there are duplicated planes, we keep records of their occurrence frequency f_{l_x}.

Plane Extraction with MDL Regularization To cut down the volume of depth planes in L_x, a plane extraction procedure is executed. Its energy function is given as:

$$E_{MDL} = \psi_i(x_i) + \psi_{ij}(x_i, x_j) + \underbrace{\sum_{l \in L_x} e^{-f_{l_x}} \cdot \delta_l}_{label\ cost} \tag{5}$$

$$\delta_l = \begin{cases} 1, & \exists x, l_x \in L, \\ 0, & otherwise. \end{cases} \tag{6}$$

where $\psi_i(x_i)$ is the sum of pixel-based absolute difference between its original plane and new mapping plane. $\psi_{ij}(x_i, x_j)$ is a Potts model penalizing on difference. Our modified label cost term[13] penalizes on occurrence frequency of planes. Roughly, the size of L_x is cut down to less than 20 after this step.

The energy function for stereo is then defined as:

$$E^S(x) = \psi_i^S(x_i) \tag{7}$$

The unary term $\psi_i^S(x_i)$ is the sum of absolute difference between current labeling and initial disparity map. We do not have a conventional pairwise term for stereo here is that we modified it into an interaction term with boundary, it will be described in details in next section.

D. Interaction Potentials

For each pair of neighboring x_i and x_j there will be an unique piece of boundary namely y_k. The interaction potential is defined as:

$$E^I(x, y) = \sum_{x_i, x_j, y_k} \psi_{ij}(x_i, x_j) \cdot \overline{y_k} \tag{8}$$

where $\psi_{ij}(x_i, x_j)$ is a Potts model. The principle of the projected graph cut is to fix one layer in MRFs at a time while optimizing the other. When layer X is fixed, and neighboring x_i and x_j do not belong to the same depth surface($\psi_{ij}(x_i, x_j)$ = 1), the energy potential will intend to decrease itself by encouraging the boundary between to be appeared($y_k = 1$). And when layer Y is fixed and y_k is turned on, the energy potential will be 0 thus the smoothness requirement of x_i and x_j will no longer be executed.

E. Joint Inference

This two-layer MRFs have the set of variables up to $\{X, Y\}$ and label space up to $L_x * L_y$. Graph with such complexity is generally difficult to optimize. We bring the idea of Projected graph cut (PGC) [8] to α-expansion optimization, it gives an approximation of the true labeling at an acceptable efficiency.

The basic steps for the inference is as follows. We start randomly either from the initial labeling f_X or f_Y, and do the optimization recursively. For instance, when we optimize Y in one iteration, suppose the optimal labeling achieved so far are f_X^* and f_Y^*. We fix X in $E^I(x, y)$ by taking the values from f_X^*, and put the transformed term together with the stand alone term $E^B(y)$, and use ST-min cut to optimize variable Y alone. If a lower energy with solution f_Y' is found, we keep the f_X^* unchanged and set $f_Y^* = f_Y'$. Optimizing X is applied in a similar subsequent way. When no lower energy can be achieved in $L_x * L_y$ iterations, the optimization stops and returns f_X^*.

F. Downsampling and Phosphene Representation

There exists a variety of image down-sampling methods. Interpolation of bilinear and cubic will compose new values for anti-aliasing purpose which may cause confusion in depth-based human navigation. Although simple nearest neighbor will not add new value, it is not robust for low-vision navigation either as it may omit some critical information in the foreground. In this paper, we propose a novel down-sampling method by integrating the boundary clues to the down-sampling process, which clearly help to discriminate the obstacle object from the surroundings in phosphene-based low-resolution navigation trial.

A brief example is given in Figure 3. The principle of nearest neighbor down-sampling is to project every down-sampled node(pixel) to original image and obtain its sub-pixel location and coordinates, and then simply select the

value of its nearest neighbor as its own. However in low-vision navigation, the priority is to avoid the nearest obstacles. Therefore during the down-sampling process, nearest neighbor algorithm may omit some critical information of foreground obstacles which merged into background, and this will cause serious problems in navigation. Such errors always happens in surface boundaries where the depth significantly changed. In the paper, we have modified and improved nearest neighbor algorithm by integrating the boundary clues to efficiently solve the problem. During the down-sampling process, it takes advantages of the boundary map, for the sub-pixel projected in the original image, if any of its neighbors in a limited scope locating on the boundary, the sub-pixel will take the largest depth value among its neighbors, otherwise it takes the value of its nearest neighbor. The experiments demonstrates such modification could emphasize the foreground objects significantly in low-resolution vision.

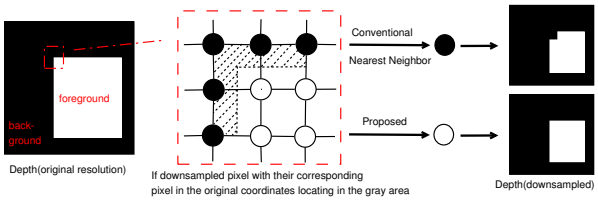

Fig. 3. An example demonstrates the advantage of our downsampling algorithm comparing to the conventional nearest neighbor.

For stimulated phosphene rendering after down-sampling, each phosphene is represented by a circular Gaussian whose center value and standard deviation are modulated by the depth at that point. In addition, phosphene sums their values when they overlap. For complete description, please refer to [14].

III. Experiment

The proposed method has been tested on Middlebury's benchmark images[3] and our indoor navigation real-scene dataset. The performance of other two popular global pixel-wise matching approaches, Graph Cut(GC) and Belief Propagation(BP) have been compared. The analysis on the real-scene dataset is presented in Figure 4 and Figure 5, while the comparisons on the Middlebury's images are in Figure 6 and Figure 7. The testbed is on a desktop computer with Intel $I3$ 2.93Ghz CPU and the proposed algorithm takes less than 100 seconds to process a high-resolution image pairs.

From the results of the indoor image pairs in Figure 4, It clearly presents that our approach has more natural and continuous depth than traditional graph cut under both obstacle and non-obstacle image pairs, as well as the obstacles stand discriminatively from the background. When comparing the performance of downsampled results, the obstacle objects are clearly discriminated from the surroundings after integrating the boundary clues into down-sampling process and it is valuable for further object detection use. While the obstacles in the traditional down-sampled look vague. In Figure 5 of

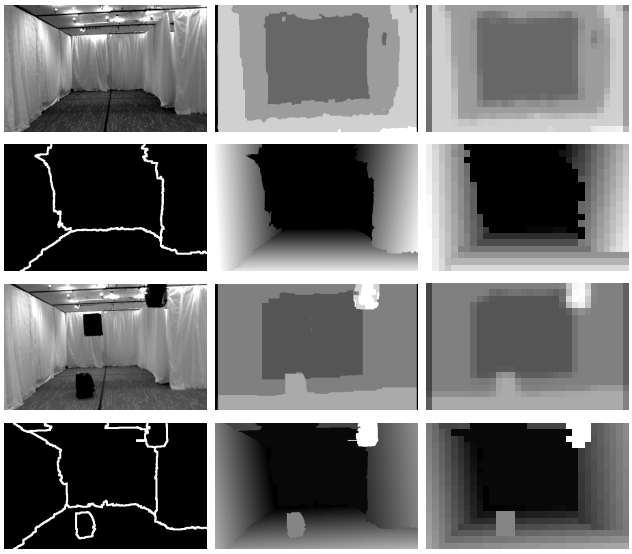

Fig. 4. The first row includes the original image without obstacles, its original-size and downsampled depth computed by Graph Cut, followed by second row with the results obtained by our algorithm, respectively surface boundary, depth and its downsamples. The third and forth rows are the results of the images with obstacles. (Original image size: 500 * 312, downsampled image size: 32 * 20)

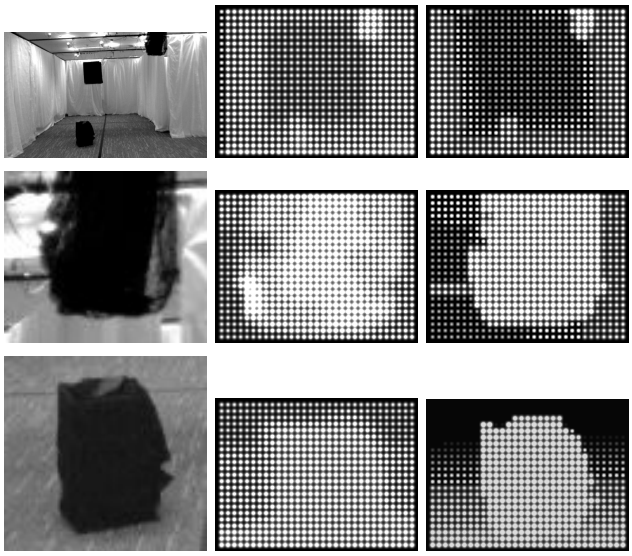

Fig. 5. Phosphene simulation of indoor scene with obstacles. The first row uses the full camera size image as the input, while the last two rows are the obstacles zoom-in effect which could be crucial in real navigation. The second and third columns are the result by Graph Cut and the proposed algorithm respectively. It can be seen that the latter one has obvious advantage in obstacle distinction.

zooming out interest regions, those obstacles could be more clearly observed in phosphene visualization.

For quantitative analysis, the proposed method has been tested together with Graph Cut and Belief Propagation on two classical Middlebury image pairs Venus and Teddy, under three different scales of original size, 1000 and 100 samples respectively. The accuracy is calculated in the following way. For every unoccluded pixels, the absolute difference of their depth with ground truth is calculated.

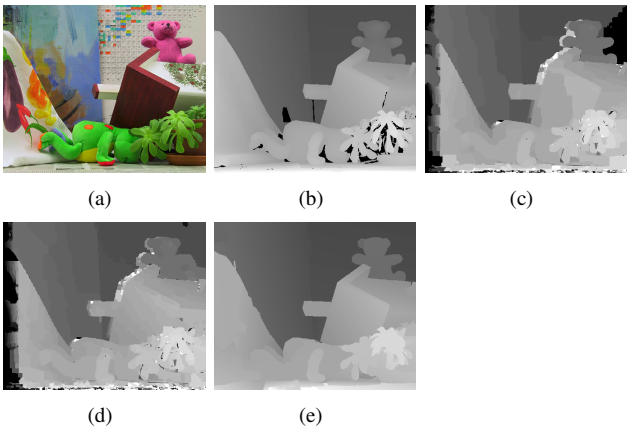

Fig. 6. Results on Middlebury's Teddy image pair: (a) original image, (b) ground truth, (c) result by Graph Cut, (d) result by Belief Propagation, (e) result by proposed method.

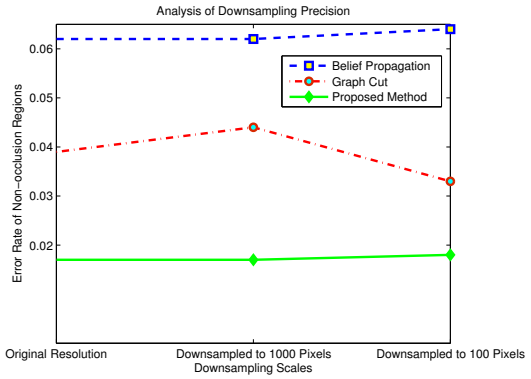

Fig. 7. Quantity analysis of precision of the proposed algorithms comparing to Graph Cut and Belief Propagation in three scales. The accuracy is computed as the average of Teddy and Venus image pairs.

Pixel with difference large than 1.0 will be labeled as $bad\ pixel$. The error rate is the average percentage of these $bad\ pixels$ over all unoccluded pixels in two Middlebury images. The original ground truth and occlusion map are all down-sampled to align the comparison under difference scaling. The results of Figure 6 and Figure 7 clearly demonstrate our method outperforms other two approaches at all three scales consistently and achieved the best accuracy with the error rate less than 2%.

IV. CONCLUSION

The paper proposed a novel two-layer associative MRFs for both depth and boundary estimation. The topologies of energy interactions and minimization are well obtained. The experiments demonstrate both depth and surface boundary have been significantly improved through the positive mutual energy interactions among two MRFs layers both quantitatively and qualitatively, when comparing with other traditional methods.After integrating the boundary clues into down-sampling process, the objects are clearly discriminated from the surroundings in both traditional and phosphene visualization under low resolution. Future research will extend such approach to object recognition and assist human navigation better.

V. ACKNOWLEDGEMENTS

Thanks are due to Paulette Lieby and Adele Scott for writing the simulated prosthetic vision software[14] used in this paper.

NICTA is funded by the Australian Government as represented by the Department of Broadband, Communications, and the Digital Economy, and the Australian Research Council (ARC) through the ICT Centre of Excellence Program. This research was also supported in part by ARC through its Special Research Initiative (SRI) in Bionic Vision Science and Technology grant to Bionic Vision Australia (BVA).

REFERENCES

[1] Z. Wang and Z. Zheng, "A region based stereo matching algorithm using cooperative optimization," in *CVPR*, 2008.

[2] Q. Yang, L. Wang, R. Yang, H. Stewenius, and D. Nister, "Stereo matching with color-weighted correlation, hierachical belief propagation and occlusion handling," in *CVPR*, 2006.

[3] D. Scharstein and R. Szeliski, "A taxonomy and evaluation of dense two-frame stereo correspondence algorithms," *IJCV*, 2002.

[4] R.; Maingreaud F. Pissaloux, E.; Velazquez, "Intelligent glasses: A multimodal interface for data communication to the visually impaired," in *MFI*, 2008.

[5] E.E. Pissaloux, "A vision system design for blinds mobility assistance," in *EMBC*, 2002.

[6] M.; Sivaprakasam M. Wentai Liu; Fink, W.; Tarbell, "Image processing and interface for retinal visual prostheses," in *ISCAS*, 2005.

[7] X Ren, C. Fowlkes, and J Malik, "Scale-invariant contour completion using conditional random fields," in *ICCV*, 2005.

[8] Russell.C Sturgess.P Bastanlar.Yalin Clocksin.William Torr. P.H.S. Ladicky.L, Sengupta.S, "Joint optimisation for object class segmentation and dense stereo reconstruction," in *BMVC*, 2010.

[9] Y Boykov and V Kolmogorov, "An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision," *TPAMI*, 2001.

[10] D. Comaniciu and P. Meer, "Mean shift: A robust approach toward feature space analysis," *TPAMI*, 2002.

[11] David R. Martin, Charless C. Fowlkes, and Jitendra Malik, "Learning to detect natural image boundaries using local brightness, color, and texture cues," *TPAMI*, 2004.

[12] Stan Birchfield and Carlo Tomasi, "A pixel dissimilarity measure that is insensitive to image sampling," *TPAMI*, 1998.

[13] A. Isack H.N. Boykov Y. Delong, A. Osokin, "Fast approximate energy minimization with label costs," *IJCV*, 2010.

[14] Paulette Lieby, Nick Barnes, Chris McCarthy, Nianjun Liu, Liu Dennett, Janine Walker, Viorica Botea, and Adele Scott, "Substituting depth for intensity and real-time phosphene rendering: Visual navigation under low vision conditions," in *EMBC*, Boston USA, September 2011.