The 2010 IEEE/RSJ International Conference on
Intelligent Robots and Systems
October 18-22, 2010, Taipei, Taiwan

3D Room Modeling and Doorway Detection from Indoor Stereo Imagery using Feature guided Piecewise Depth Diffusion

Karthik Mahesh Varadarajan and Markus Vincze

Abstract—**Traditional indoor 3D structural environment modeling algorithms employ schemes such as clustering of dense point clouds for parameterization and identification of the 3D surfaces. RANSAC based plane fitting is one common approach in this regard. Alternatively, extensions to feature based stereo have also been used, mainly focusing on 3D line descriptions, along with techniques such as half-plane detection, real-plane or facade reconstruction, plane sweeping etc. Noise in the range data, especially in low texture regions, accidental line/plane grouping under lack of cues for visibility tests, presence of depth edges or discontinuities that are not visible in the 2D image and difficulties in adaptively estimating metrics for clustering can hamper efficiency of practical systems. In order to counter these issues, we propose a novel framework fusing 2D local and global features such as edges, texture and regions, with geometry information obtained from range data for reliable 3D indoor scene representation. The strength of the approach is derived from the novel depth diffusion and segmentation algorithms resulting in superior surface characterization as opposed to traditional feature based stereo or RANSAC based plane fitting approaches. These algorithms have also been heavily optimized to enable real-time deployments on personal, domestic and rehabilitation robots.**

I. Introduction

Traditional indoor 3D structural environment modeling algorithms employ schemes such as clustering of dense point clouds for parameterization and identification of the 3D surfaces. RANSAC based plane fitting [1] is one common approach in this regard. Alternatively, extensions to feature based stereo have also been used, mainly focusing on 3D line descriptions, along with techniques such as half-plane detection, real-plane or facade reconstruction, plane sweeping etc. Pioneering work in this regard is attributed to Baillard et al. [2,3] and Zisserman et al. [4]. Other important works include facade detection and multi-level regeneration by Lee - Nevatia [5]. Recent efforts at plane grouping based on PCA and visibility tests include [6] and [7]. The literature in building 3D line descriptor based structure analysis is also quite vast. Recent articles include Hausdorff measure based grouping [8] and model based recognition [9]. However, the performance of most of these techniques rapidly degrade in the presence of high amounts of noise (in range data such as stereo) under conditions of low illumination and in regions of low-texture or sparse features. Furthermore, accidental line/plane grouping (for eg. in the case of shelves/ cupboards), especially under lack of cues for visibility tests, presence of depth edges or discontinuities that are not visible in the 2D image and difficulty in adaptively estimating metrics for clustering can hamper efficiency of practical systems for door/doorway detection. On the other hand, traditional laser [10] or panoramic camera [11,12] (or multi-view) based room modeling and doorway detection systems (often using piecewise planar modeling [13], triangulation [14] or space carving [15]) are often impractical for cost-effective domestic robots. Moreover, machine learning based door recognition systems (usually from only 2D images) such as [16,17,18], perform poorly in cluttered scenes (especially with floor reflectance) and when the door is open or is viewed partially or the doorway is structurally similar to an arch and lacks the actual door and door frame, besides the inherent ambiguity in handling cupboards/shelves. Depth based doorway detection is more practical and useful in such cases and also provides cues for place learning.

In order to resolve these challenges, we propose a novel framework fusing 2D local and global features such as edges, texture and regions, with geometry information obtained from pixel-wise dense stereo for reliable 3D indoor scene representation. The strength of the approach is derived from the novel depth diffusion and segmentation algorithms resulting in superior surface characterization. Unlike earlier schemes, these algorithms also enable identification of depth edges that are critical to surface isolation. The pipeline also renders visibility tests and constraints superfluous.

The proposed framework follows a three step process – detection of walls, followed by the enclosing room and finally doorways. Walls and wall-like surfaces are detected using 2D edge, texture and region features. The 3D surfaces corresponding to the walls are then generated using piecewise depth diffusion techniques followed by depth segmentation to identify intra - wall depth discontinuities. The room model is built by classifying and selecting these wall-like surfaces to fit cuboidal (deformable) constraints. Finally, doorways in the room are hypothesized based on clustering of the dense stereo data pixels that do not conform to the concave room hypothesis.

II. Overview

This paper offers a number of novel contributions. The main contributions are listed below. Firstly, this paper presents an innovative framework for the processing of stereo images for 3D reconstruction of indoor structural

Manuscript received March 10, 2010. The research leading to these results has received funding from the European Community's Sixth Framework Programme (FP6/2003-2006) under grant agreement no FP6-2006-IST-6-045350 (robots@home).

Karthik Mahesh Varadarajan and Markus Vincze are with the Vienna University of Technology, Automation and Control Institute, Gusshausstrasse 30 / E376, A-1040 Vienna, Austria (email: {kv,mv} @acin.tuwien.ac.at).

environments and detection of doorways, corridors and other negative spaces in the given scene of interest. Secondly, this paper demonstrates efficient indoor wall segmentation by extending intrinsic gradient extraction and texture analysis algorithms. Thirdly, a new image-agnostic scheme of noise removal in range images is presented. Fourthly, a real-time depth diffusion algorithm, suitable for 3D surface generation, especially with extremely sparse range data is introduced. A fifth contribution is the demonstration of efficiency of intra-entity surface segmentation and depth discontinuity detection resulting from the novel diffusion framework. Other contributions include a framework for wall reconstruction using transformed plane fitting, a PCA based plane grouping scheme for building the room model and doorway detection using a concave room structure hypothesis.

The images used for evaluating the developed algorithms have been obtained in an indoor environment, from an experimental robot at a height of about 1m from the ground plane. A high dynamic range monochrome stereo camera is used to estimate the range images along with a centrally mounted inexpensive color camera. Note that the algorithms presented here are well suited for fusion of data from distinct color and range (stereo or otherwise) sensor systems. In order to simplify the algorithmic framework, it has been assumed that the fixed pose of camera and its height above the ground plane are known accurately, thereby establishing the approximate ground plane in any scene without further processing.

The approach for 3D room reconstruction and doorway detection presented in this paper follows a 3 stage modeling pipeline comprising of wall modeling, room modeling and doorway modeling. The various assumptions required for the modeling/ hypothesis at each stage are listed below.

A. *Wall Modeling*

Walls are typically characterized by

1. Homogeneous regions or areas with regular texture, usually with high numeric intensity values.
2. Largest single color regions in a given scene, especially when there are no large occluding obstacles in the vicinity.
3. Hold pixels with the farthest visible range information on planes parallel to the ground plane.
4. Frequent loss of homogeneity in color values owing to lighting and shading effects.

B. *Room Modeling*

Rooms are characterized by

1. Combination of walls typically as a cuboid.
2. Largest and most consistent of all possible cuboids in the scene (helps exclude walls internal to the room).
3. Often encompasses all extreme range pixels in the horizontal dimensions (along the image width).
4. Room fitting can be reduced in most cases (based on assumptions of known floor and ceiling) to fitting of a maximum of just three (largest) vertical walls.

C. *Doorway Modeling*

Doorways are characterized by

1. External outliers (or exclave points in range images) to the room model, that can be grouped to form regions with size bounds similar to that of typical doors or doorways.
2. These outliers should be at a jump discontinuity to the modeled room surfaces.
3. Floors are typically uniform across doorways.

III. ALGORITHM

The algorithmic pipeline presented in this paper follows from the above sequence of modeling. The framework can largely be divided into three sections. The first section deals with color image processing, wherein after pre-processing, reflectance image gradients are extracted from the 2D image and segmentation (along with region selection) is carried to identify walls and wall-like regions. The second section performs dense stereo depth data processing in a number of steps that include denoising, piecewise diffusion to reconstruct depth surfaces and depth segmentation to identify intra-object depth discontinuities. The last section deals with 3D indoor structure generation by fitting planes to the wall-like surfaces and grouping them to find room boundaries. The 3D reconstruction of the room from the depth map leads to detection of doorways and other negative spaces in the room. The various stages are detailed below.

Color Image Processing

A. *Color Pre-processing*

The color image, obtained from the centrally located camera is rectified and used as the reference image. Locations of dead or noisy sensor pixels are pre-determined and the intensity at these locations approximated by nearest-neighbor filling. All other images from the sensor system

Figure 1. Intrinsic Image Extraction and Segmentation (A) Input color image (B) Segmentation using the standard Felzenszwalb-Huttenlocher (FH) graph based algorithm – demonstrates high clutter in regions of the left wall with lighting changes (C) Shading intrinsic image (D) Reflectance intrinsic image – note that C and D (obtained by inversion of input image gradients classified as shading or reflectance respectively) are presented here only to demonstrate the separation of shading/lighting effects from material properties; segmentation is not carried out on these images (E) Segmentation on the input image using a low complexity multi-scale full gradient edge analysis scheme (F) Segmentation on the input image with the same scheme using reflectance-only gradients – shows superior performance in wall regions affected by lighting changes in comparison with full-gradient image segmentation schemes such as the graph based FH. Note that similar values of gradient and region size thresholds were used for the two approaches - FH and our scheme.

(range images - stereo pair and hence depth image, confidence image) are referenced to the coordinate system of the color image by registering with this image. As a pre-processing step, the noise in the color image is reduced using a bilateral filter that preserves salient gradient values and hence sharp edges that are crucial for algorithms in the following stages of processing, including 2D segmentation.

B. *Intrinsic Reflectance Gradients Extraction*

The gradients of the filtered color image are estimated and these gradients are decomposed into shading and reflectance components. The shading component captures the lighting and shadows in the scene while the reflectance component captures the distinction in the material surfaces. This step is helpful to eliminate the highlights and shadow patterns created by light fixtures typically mounted on walls. Since walls are the primary focus of this 3D room reconstruction and doorway detection solution, it is beneficial to use reflectance components since they are devoid of gradients pertaining to highlight and shadow artifacts, thus representing the wall faces as true homogenous surfaces. The algorithm we employ for intrinsic gradient extraction is based on the intrinsic image extraction algorithm developed by Weiss [19] and extended by Tappen et al. [20]. In the presented framework, gradients in the intensity channel of the color image are classified as 'shading' or 'reflectance' gradients by modeling an asymptotic linear color variation across neighboring pixels. The formulation for intrinsic image extraction [20] is,

$$I(x,y) = S(x,y) \times R(x,y) \quad (1)$$

where $S(x,y)$ is the shading image, $R(x,y)$is the reflectance image and $I(x,y)$is the input image defined in the dimensions $x$and y. Using a logarithmic transformation and applying multiple scale selective gradient/ derivative filters f_x, f_y we have the gradient images F_x and F_y, the (x,y) components of which can be classified as shading if the color pixels satisfy the constraints $c_{x+1} = \alpha c_x$and $c_{y+1} = \alpha c_y$respectively and as reflectance otherwise. Shading and reflectance component images can be reconstructed from the gradients as

$$C(x,y) = g * [(f_x(-x,-y) * F_{cx}) + (f_y(-x,-y) * F_{cy})] \quad (2)$$

where, * represents convolution, $F_{cx}$and $F_{cy}$are component gradients (image gradients classified as either shading or reflectance) and g is obtained from

$$g * [(f_x(-x,-y) * f_x(x,y)) + (f_y(-x,-y) * f_y(x,y))] = \delta \quad (3)$$

The shading and reflectance components as defined by equation (2) are shown in Fig. 1C and 1D. In our framework, the reflectance image gradients $F_{cx}$and F_{cy} are used directly in the segmentation process. One possible disadvantage of this scheme is that gradients at edges pertaining to depth discontinuities or surface orientation changes in walls and other structures may not be captured in the reflectance component. This disadvantage limits the application of algorithms for intrinsic image extraction. However, this is not a major issue in the presented framework as the additional step of depth segmentation detects these gradients, from the depth image.

C. *2D Reflectance Gradient based Segmentation and Region Isolation using Texture Analysis*

Using the gradients $F_{cx}$and F_{cy}obtained in the previous stage, segmentation is carried out using a low complexity multi-scale edge analysis scheme. The scheme links edges found at various scales (by analysis of reflectance only gradients) using proximity and similarity measures to form enclosed regions or segments. The choice of the segmentation algorithm is based on the goal of meeting real-time constraints for deployments on robots, which excludes the possibility of using algorithms like the Felzenszwalb-Huttenlocher (FH) graph based algorithm. It can be seen from Figure 1B, 1E and 1F that the performance of the proposed 'reflectance gradient only' segmentation approach is superior in terms of output to traditional full-gradient image segmentation approaches like FH (with gradients as grid graph edge weights) and the variant of the multi-scale edge analysis algorithm operating on full-gradients, in the given context of wall detection with lighting and shading changes. The performance is comparable in regions devoid of lighting changes. Any possible over-segmentation in regions of high texture does not affect the output since these regions are unlikely to be wall surfaces. The segmented regions are then subjected to a region selection algorithm that uses pixel spans and texture analysis to select walls and wall-like structural surfaces that are expected to support the room model. The characteristic features of walls such as homogeneity, large pixel spans and representations using high gray-scale intensity values are used in region selection. The current framework employs 2 levels of thresholds (hard and soft) on measures of entropy (E), homogeneity (H), uniformity energy (U), correlation (R), contrast (C) and other constraints based on the Grey Level Co-occurrence Matrix (GLCM) to select wall-like surfaces. Estimated soft threshold values, along with the assigned confidence values of the measures on condition conformance (in brackets) for the two-class separation (positive wall classification) are

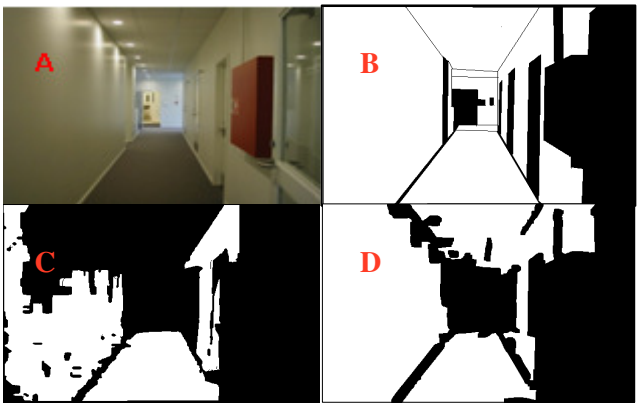

Figure 2. Segmentation and Regions of Interest Selection (A) Input color image (B) Ground Truth (Manual Segmentation) for walls and wall-like regions (floors, ceiling etc.) (C) Results using FH (mislabeled pixels: 70292) (D) Results using our framework (mislabeled pixels: 32069) – note the correspondence to Fig. 1B and 1F respectively, mislabeled pixels are estimated by XOR logical gating with the ground truth.

H>0.99 (1.0), C<0.0275 (1.0), R>0.9 (0.9) - can be undefined or slightly negative under near perfect surface homogeneity; reduces slightly below this threshold for walls with rough texture such as in the case of visible brick layouts, U>0.6 (0.3) - is unreliable and varies with the lighting changes; ideally 1.0 under no lighting change or natural lighting but drops to 0.3 under artificial lighting/ large lighting changes, E<5.5 (0.8). The hard threshold values are H>0.96 (0.8), C<0.0475 (0.7), R>0.85 (0.6), U>0.3 (0.1), E<7.0 (0.5). The surface is classified as a wall if the aggregate confidence value exceeds 3.0 out of a maximum of 4.0. On a representative data set of 80 image chips of various material textures found indoors, such as wood, tile, brick, rock, vegetation, carpet, cloth, curtain, steel, bronze, tree bark, granite etc., besides painted wall surfaces, the classifier achieved a classification rate of 95%, with a wall detection rate of 97.87%. False alarms were caused due to white curtains, steel and floor tiles that were homogenous or 'wall-like'. The features were robust to wall colors and surface roughness. Objects such as uniformly colored doors and cupboard doors had confidence measures close to that of walls. Since these surfaces are also helpful in the room reconstruction, they are used in further analysis. Thresholds on pixel spans of the surfaces ($> I_w*I_h/15$, where, I_w is image width and I_h is image height) and average gray-scale intensity (> 100/255), further help reduce the detected segments to the set of primary room surfaces. The framework can be extended to use machine learning techniques to adaptively estimate these thresholds based on in situ training in the environment of deployment of the robot. This is the scope of future work. Figure 2 demonstrates the results of the segmentation and region selection approach and compares the performance with the output of the region selection approach in combination with the FH segmentation algorithm. While the number of mislabeled pixels is 70292 with the FH approach, this number is less than half (at 32069) for our framework. The region masks thus obtained are used for piecewise isotropic depth diffusion.

Stereo Depth Image Processing

D. Depth Pre-processing

Depth pre-processing involves registration and transformation of the depth pixels to the coordinate system of the color camera. This is followed by noise removal. This is done using a novel sparse de-noising algorithm, employing iterative hysteresis filtering and morphological reconstruction. The various steps in the proposed noise removal algorithm are detailed below:

1. The input depth map is divided into core-blocks and the standard deviation (σ_c) of each core-block is estimated using values of depth pixels that have high confidence measures (obtained from the confidence map). Macro-blocks corresponding to each core-block are created, composing of a larger number of pixels and centered at the core-block and the standard deviation estimated as σ_m. Macro and core block linear dimensions are selected as $I_w/10$ and $\sqrt{I_w/10}$ based on expected spans of depth surfaces at mean ranges from the camera (for given FOV). Typical core and macro block sizes are 7x7 and 50x50 respectively.
2. For each core-block and macro-block, logical maps corresponding to all valid pixels, the values of which fall within a pre-determined threshold times σ_c and σ_m, respectively, are estimated. The threshold for the macro-block is set higher than that for the core-block, thereby permitting greater deviation. Testing on a number of images yielded a rough rule of the thumb calculation for the macro-block and the core-block thresholds. The thresholds can be calculated as 0.25/P and 0.175/P respectively, where P is the percentage pixel density with typical values of the thresholds being 1.0 and 0.7 for 25% pixel density.
3. Valid pixels in the core-block that are flagged true in both the logical maps retain their original values in the filtered depth map. These pixels are well-behaved, in the sense that they satisfy topological smoothness constraints and are likely to belong to the same surface.
4. Pixels that are flagged true in only one of the logical maps are categorized as hysteresis pixels. These pixels might belong to other surfaces at a depth discontinuity with respect to the most prominent depth surface in the current core-block. For each hysteresis pixel, neighborhood pixels are ascertained (based on limits - set at 300 for the current 16-bit depth pixel range, on depth values from the current hysteresis pixel value and connected component analysis) in the macro-block map (which is expected to contain much of the surface supporting the hysteresis pixel, while only a small portion of this surface is present in the core-block that had resulted in the hysteresis pixel being classified as an outlier in the core-block logical map). If the size of the neighborhood pixel region exceeds a certain threshold, indicating the presence of a high confidence depth surface as opposed to spurious depth pixels, the hysteresis pixel along with the neighborhood pixels are added to the filtered map.
5. The above steps are iterated for the entire depth map until

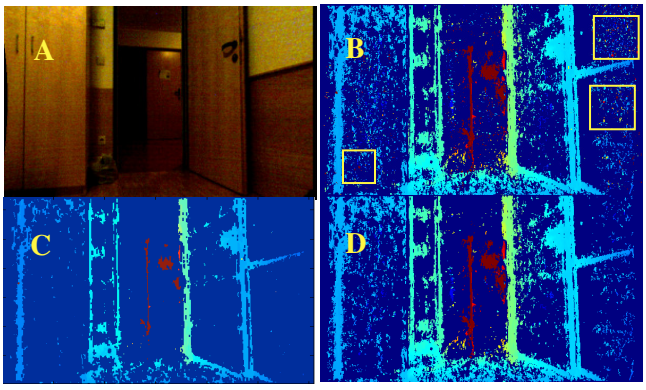

Figure 3. De-noising depth data. (A) 2D Color image (B) Noisy depth map from the stereo system – Note the presence of red pixels in areas that are predominantly blue and vice-versa (the depth varies from blue – nearest to red – farthest). Bounding boxes in orange are shown around large clusters of noisy pixels (C) Results from a median filter – Note that though much of the noise is removed, only pixels at surface segment boundaries have been preserved while almost all of the valid depth values in the interior of surface segments (which are crucial for proper surface curvature reconstruction) are lost (D) Results of the proposed scheme – Note that while most of the noise is removed, surface depth pixel values are preserved

the number of pixels classified as noise pixels between iterations falls below a threshold.

The final noise filtered depth map is obtained by morphological reconstruction of the marker under the mask map, where the iterative hysteresis filtered depth map obtained in the previous step is used as the marker and the original depth map is used as the mask. Results presented in Figure 3 show superior performance of our algorithm in relation to another image-agnostic filtering scheme – the median filter. On a representative set of images from the Middlebury dataset, the filtering scheme reduced the noise in the dense depth data from ICM stereo by an average of 26% (MSE drop from 2300 to 1700).

E. Depth Diffusion

Since the input depth map is quite sparse, it is required to convert it into a dense cloud for reliable and coherent surface estimation. This step is necessary since the span of the surfaces (in terms of pixels) is crucial for reliable weighting in the fitting and room reconstruction process and for outlier rejection. Diffusion of depth values is carried out using a Piecewise Isotropic Laplacian Partial Differential Heat Linear Equation (PDE) Solver that operates only in regions identified by masks obtained in step C. By combining Multi-grid and Iterative Back Substitution (IBS) schemes to solve the PDE equation, rapid convergence is obtained, demonstrating suitability for real-time deployments.

The PDE representing the flow of heat in a 2 dimensional isotropic medium [21] is given by

$$\frac{\partial u(\mathrm{r},t)}{\partial t} = c\left(\frac{\partial^2 u(\mathrm{r},t)}{\partial x^2} + \frac{\partial^2 u(\mathrm{r},t)}{\partial y^2}\right) \quad (4)$$

where, $u(\mathrm{r},t)$ represents the heat measured in the two dimensional space r(x,y) at time t. If c varies in the space of the depth map dimensions, the equation becomes anisotropic. Equation (4) can be also be used to represent depth diffusion, where $u(\mathrm{r},0)$ represents the original depth values and $u(\mathrm{r},t_{ss})$ final depth values obtained after diffusion (at steady state). This equation is equivalent to

$$\frac{\partial u(\mathrm{r},t)}{\partial t} = c\nabla^2 u(\mathrm{r},t) \quad (5)$$

where, ∇^2 is the Laplacian operator. While it is possible to define c using the confidence map or reflectance gradients in order to preserve edges, we hold c constant within regions identified by the segmentation mask in the current framework, giving rise to a piecewise isotropic formulation. This scheme preserves the wall edges (segment boundaries) identified in the 2D image. Using the tuple *(i,j)* for the row and column indices of the image, we have

$$\begin{aligned}\frac{\partial u(r_{(i,j)},t)}{\partial t} &= u(r_{(i,j)},t+1) - u(r_{(i,j)},t)\\ &= \varphi\big[\, c_{(i-1,j)}.\nabla u(r_{(i-1,j)},t)\\ &+ c_{(i,j-1)}.\nabla u(r_{(i,j-1)},t)\\ &+ c_{(i+1,j)}.\nabla u(r_{(i+1,j)},t)\\ &+ c_{(i,j+1)}.\nabla u(r_{(i,j+1)},t)\big] \end{aligned} \quad (6)$$

where, the constant $\varphi \le 0.25$ controls the overall rate of diffusion. In the steady state,

$$\begin{aligned}(1/\varphi).u(r_{(i,j)},t_{ss})&\\ - \big[\, c_{(i-1,j)}.u(r_{(i-1,j)},t_{ss})&\\ + c_{(i,j-1)}.u(r_{(i,j-1)},t_{ss})&\\ + c_{(i+1,j)}.u(r_{(i+1,j)},t_{ss})&\\ + c_{(i,j+1)}.u(r_{(i,j+1)},t_{ss})\big]&\\ = 0 \end{aligned} \quad (7)$$

Representing $(1/\varphi)$ as λ and linearizing the tuple indices, (7) can be reduced to a matrix system. A sample matrix for a 3x3 depth image, is shown in (8)

$$\begin{bmatrix} \lambda & c_{12} & 0 & c_{21} & 0 & 0 & 0 & 0 & 0\\ c_{11} & \lambda & c_{13} & 0 & c_{22} & 0 & 0 & 0 & 0\\ 0 & c_{12} & \lambda & 0 & 0 & c_{23} & 0 & 0 & 0\\ c_{11} & 0 & 0 & \lambda & c_{22} & 0 & c_{31} & 0 & 0\\ 0 & c_{12} & 0 & c_{21} & \lambda & c_{23} & 0 & c_{32} & 0\\ 0 & 0 & c_{13} & 0 & c_{22} & \lambda & 0 & 0 & c_{33}\\ 0 & 0 & 0 & c_{21} & 0 & 0 & \lambda & c_{32} & 0\\ 0 & 0 & 0 & 0 & c_{22} & 0 & c_{31} & \lambda & c_{33}\\ 0 & 0 & 0 & 0 & 0 & c_{23} & 0 & c_{32} & \lambda \end{bmatrix} \begin{bmatrix} u_{11}\\ u_{12}\\ u_{13}\\ u_{21}\\ u_{22}\\ u_{23}\\ u_{31}\\ u_{32}\\ u_{33} \end{bmatrix} = \begin{bmatrix} 0\\ 0\\ 0\\ 0\\ 0\\ 0\\ 0\\ 0\\ 0 \end{bmatrix} \quad (8)$$

$$Ax = B \quad (9)$$

This system of equations forms a block-tridiagonal matrix system with fringes. In matrix *A*, from equation (8), the blocks are denoted by red squares, the upper (called $c_1(i,j)$) and lower ($a_1(i,j)$) tri-diagonals by the violet and blue indices along with the main diagonal, the upper fringe ($c_2(i,j)$) in green and lower fringe ($a_2(i,j)$) in orange, along with the main diagonal ($b_1(i,j)$). For the case of pixels with known depth values, the corresponding row in the *A* matrix has only one non-zero element (at the main diagonal and equal to 1), the row in the *x* vector is non-zero and equal to the known depth value and that in *B* is set to 1. IBS algorithm has been used for isotropic diffusion of grayscale images, in the context of image compression [22]. This paper extends the scope of IBS to perform piecewise isotropic diffusion of depth data. Adapting the IBS scheme [23] for the case of depth maps, the pseudo-code for solving the system is:

FringeTriDiagSolver := {InitializeSolution,
InitializeMatrixComputation, i_{iter} *-> 0,*
While[{CurrEps > EpsTol && i_{iter} *< MaxItr &&*
AbsErr > AbsErrTol},{
i_{iter} *->* i_{iter} *+ 1,*
StorePreviousResult,
ForwardSubstitution,BackwardSubstitution,
ComputeMaximumResidual}] }

where, *InitializeMatrixComputation* estimates the values of intermediate matrices G, Q_i, $P_i$as,

$$\begin{aligned} G(i,j) &= 1/(-a_1(i,j) * Q_1(i-1,j) - a_2(i,j)\\ &\quad * Q_2(i,-1) - b_1(i,j));\\ Q_1(i,j) &= G(i,j) * (a_2(i,j) * Q_1(i,j-1) * Q_2(i+1,j\\ &\quad -1) + c_1(i,j));\\ Q_2(i,j) &= G(i,j) * c_2(i,j);\\ P_1(i,j) &= Q_1(i,j) * X(i+1,j);\\ P_2(i,j) &= Q_2(i,j) * X(i,j+1); \end{aligned} \quad (10)$$

ForwardSubstitution and *BackwardSubstitution* modules are iterated until convergence of *X* estimated as,

$$M(i,j) = G(i,j) * (a_1(i,j) * (M(i-1,j) + P_2(i-1,j) + P_3(i-1,j)) + a_2(i,j) * (Q_1(i,j-1) * (M(i+1,j-1) + P_1(i+1,j-1)) + M(i,j-1) - S(i,j));$$

$$P_1(i,j) = Q_1(i,j) * X(i+1,j);$$

$$P_2(i,j) = Q_2(i,j) * X(i,j+1);$$

$$X(i,j) = M(i,j) + P_1(i,j) + P_2(i,j) + P_3(i,j); \quad (11)$$

where G is an inverse matrix, Q_i, P_i, M are intermediate matrices and S is the solution matrix (the right side of the equation). Traditional isotropic diffusion solvers smooth out edge regions, while direct application of anisotropic diffusion to depth data smoothens depth edges in regions where image gradients are weak (such as in the case of the edge of intersection of two homogenous wall surfaces). In the piecewise isotropic diffusion solver, the calculation of the forward and backward substitution modules is suppressed for known depth pixels, thereby propagating and preserving depth edges across iterations in addition to those identified during the segmentation process (Fig. 4).

While the above solution is reasonably fast (of the order of 0.5 sec on a 3.2 GHz single core PC with 512 MB RAM, for a 320x240 depth image), the convergence rates are to be further enhanced for real-time operation on resource constrained systems. In our frameworks, we use a variant of the multi-grid approach (that employs the solved equation systems at smaller scales as pre-conditioners for higher scales) to speed-up calculations of the IBS. Results of piecewise depth diffusion for an input depth map (Figure 4B) are presented in Figure 4C. On the representative Middlebury dataset, diffusion reduced the MSE from 1700 (post-filtering) to 1100.

F. *Depth Segmentation*

An additional step of depth segmentation is necessary to detect depth discontinuities and hence surface boundaries that are not captured in 2D edge segmentation. A good example is the case of a discontinuity in a wall surface as a result of a pillar or column like structure or a depth edge created at the intersection of two wall faces of a room (Fig. 4C, 4D). Since the faces of the room are expected to be of the same color, it is possible that a reliable edge is not detected at the junction of these faces or at locations of surface orientation changes on a column during color processing. As explained in the previous section, our novel diffusion step renders these edges detectable and regions separable using a standard segmentation approach. A number of range segmentation algorithms such as relaxation labeling, planar and linear region growing, clustering are available for segmentation of dense depth data [24]. In our approach, the simple, low-complexity multi-scale edge detection and linking approach explained in section C is sufficient for intra-object (here intra-wall) depth discontinuity detection as demonstrated in Fig. 4. This approach is chosen with a view of minimizing computation requirements. This also removes noisy depth surfaces.

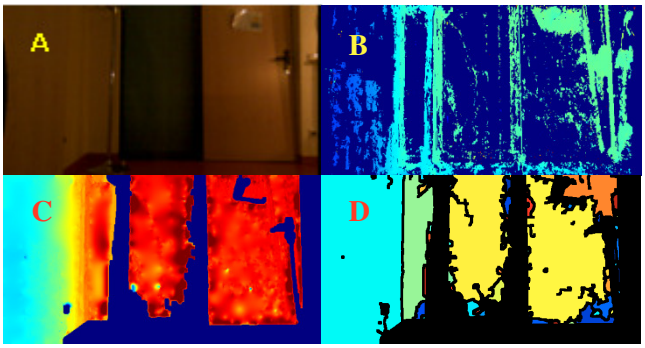

Figure 4. Depth Diffusion and Segmentation (A) Input image – please note that the depth edge formed by the junction of wall plane parallel to the observer (in front) and perpendicular to the observer (to the left) is hardly visible and hence color image segmentation produces one single surface (B) Input depth map (C) Diffused depth map (after filtering & within mask from step C) – here the depth edge is clearly visible (D) Segmentation in depth identifies the depth edge (between the blue and green segments)

3D Scene Generation

G. *Surface Fitting*

The detected wall-like depth segments are then fit to planar surfaces. This process helps parameterize the depth surfaces, rendering surface orientation analysis easier. All surfaces that do not conform to planar constraints are eliminated based on the measure of error obtained from the fitting process. Since walls are expected to satisfy Manhattan constraints and the floor plane is approximated to be perpendicular to the image plane, all depth surfaces that are not perpendicular to the floor plane (within tolerance limits) are also excluded from further analysis.

The plane fitting is carried out using Iteratively Re-weighted Least Squares Robust Linear Regression. In order to overcome the effect of propagation of errors to the 3D planar coordinates - X and Y (ideally independent variables) from the depth coordinate Z (ideally dependent variable) during point cloud estimation, the 3D fitting is carried out using a reprojected equation in the image plane given by (12). The equation is solved using a transformation of variable (1/Z) to a temporary variable z, with x and y being image coordinates.

$$\frac{1}{Z} = \left(\frac{A}{f_x D}\right)x + \left(-\frac{B}{f_y D}\right)y + \left(\frac{C}{D} - \frac{A(1-C_x)}{f_x D} + \frac{B(1-C_y)}{f_y D}\right) \quad (12)$$

where, A, B, C and D are the true plane equation coefficients in the 3D world, f_x and f_y are camera focal lengths, while (C_x, C_y) is the principal point (The additional negative sign is due to an inverted Y reference system).

H. *Room Boundaries Detection*

The depth surface planes are projected onto the ZX plane and PCA is used to find the principal orientation of each wall like surface. The planar surfaces are then classified by orientation and mapped to a cuboidal (deformable – to permit some deviation) structure. Planes that do not support the cuboid hypothesis are rejected. Planes are ranked based on consistency, span, texture content and the degree of meeting Manhattan constraints. Higher ranked planes are preferred (using a rule based framework) in the room boundary establishment. This scheme permits use of wall-like surfaces like doors (at reasonable orientations) and cupboards for approximating the room reconstruction

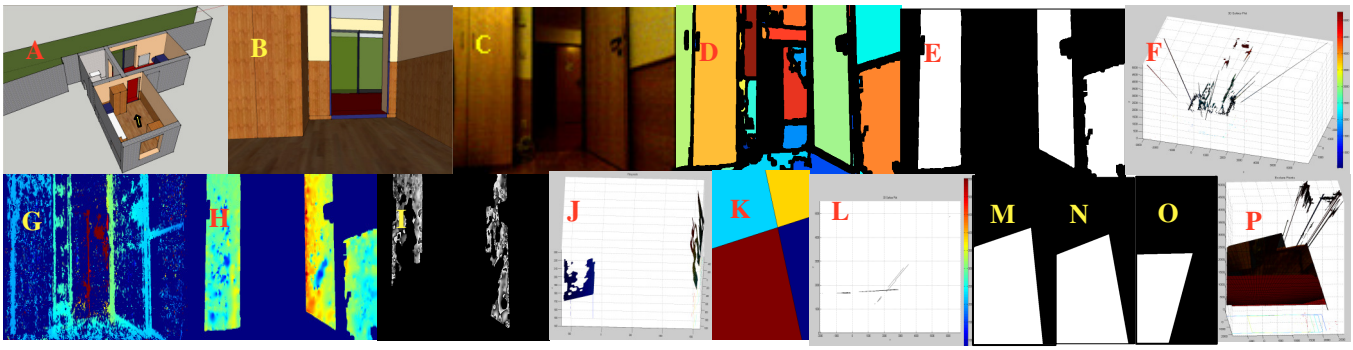

Figure 5. Complete Algorithmic Pipeline (A) 3D ground truth of the scene (yellow arrow indicates position of camera) (B) Synthetic view from camera location – shows corner of a room composed of two intersecting wall planes and consisting of a doorway leading to a second room (C) 2D input image (D) Image segmentation output (E) Region selection output (of wall-like structures) (F) Input 3D point cloud (G) Input depth map (H) Depth diffusion output (I) Depth segmentation output (J) Surface fitting results (near top view) (K) Surface categorized and PCA based room boundary detection (L) Comparative results with RANSAC based plane fitting on segmented point clouds- inliers only (M) Ground truth room sector (N) Room sector results of proposed framework (mislabeled pixels = 8713) (O) Comparative results for RANSAC scheme (mislabeled pixels = 40125) (P) 3D Reconstruction (top view – note similarity to 5N) with exclave points corresponding to the doorway

whenever the current camera viewpoint does not contain significant wall surfaces. In case of contention between wall-like and other large (door/cupboard door) planar surfaces, the algorithm adaptively chooses the wall-like surfaces for reliable reconstruction. The active room sector is also identified by the scheme.

This framework renders any visibility tests and constraints superfluous. This is because the depth diffusion uses the values of all know depth pixels (noise suppressed) to build the depth surfaces, irrespective of the curvature and the depth segmentation step breaks those surfaces that would not have satisfied visibility constraints. Also, by ensuring that those surfaces that do not fit the planar constraints are removed from the final room boundary modeling and by permitting the PCA based approach to build wall sectors, only the most consistent and visible surfaces are used in the room modeling process (Fig. 5).

I. 3D Room Reconstruction and Doorway Detection

Using the detected room sector map, height of the camera above the ground and standard room height measurements, the 3D structure of the room is reconstructed. Doorways are detected by clustering depth pixels that do not support the concave room structure hypothesis. Typical measurements of doors are used to improve localization of doorways. These doorways are modeled as open regions in the 3D representation with the exclave pixels (those belonging to the room seen through the doorway) as sparse 3D points (Fig. 5P). The 3D reconstruction has also been texture mapped to enable easy identification of the exact position and width of the doorway. The use of bounds on depth discontinuity ranges (between current room boundaries and exclave pixels for categorization as a surface from the room beyond the doorway) make the scheme robust to presence of small cupboards and other enclosures, leading to high recognition rates for true doorways (Refer Fig 6C & 6D).

IV. Analysis

The results presented in the previous section demonstrate the robustness of the framework. As described in Figure 5 (with results for all stages of the algorithmic pipeline and comparative analysis), the proposed scheme outperforms traditional RANSAC based plane fitting and room boundary detection algorithm that uses the output of our surface segmentation approach. The pixel mislabeling error is 5 times higher for RANSAC (our framework: 8713, RANSAC: 40125). Direct application of RANSAC to the 3D data set produces even worse results due to the high amount of noise in the input depth data. This is reflected by the fact that the orientation of the vertical plane (as it appears in the top view image in Fig. 5(O)) has a high amount of error. The algorithm is also shown to be robust for a variety of complex scenes. Fig. 6 describes robust and consistent performance for two positive scenarios (true doorways) and two scenes

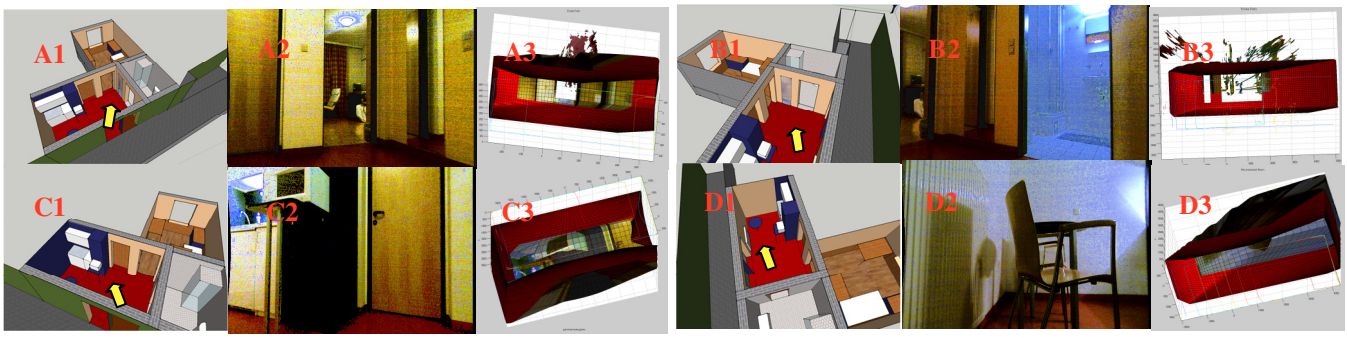

Figure 6. 3D Scene Reconstruction and Doorway Detection - Sets A and B are scenes with true doorways, while sets C an D are cluttered scenes with plenty of negative spaces but no doorways. Images indexed 1 present 3D ground truth of the scene, 2 are 2D input images and 3 are final reconstructions. Note that while true doorways have been estimated in sets A (at the intersection of two perpendicular wall faces) and B (2 doorways - a large one leading to the brighter room directly in front and a small one at the extreme left), the algorithm builds the 3D scene in sets C and D without detecting any doorways, demonstrating the robustness of the scheme to clutter.

with high clutter but no true doorways. While in the given environment (as modeled in Fig. 6 A1), all four doorways were reliably detected, the system was also demonstrated to work robustly in larger environments with multiple doorways and heavy clutter such as cupboards and closets. These are presented in Figure 7 in the form of reprojections of the built 3D models on to the image plane. Dark blue and red denote doorways and the floor, while walls are labeled in other colors. The reprojection error, measured here in terms of a rough metric of number of mislabeled pixels when compared with the manually labeled doorway, floor and wall regions in the image are also shown. It is seen that reprojection error is about 5% in typical scenes and exceptions are due to large and dynamic occlusions (such as humans). With a view to achieving real-time operation, critical modules such as depth diffusion have already been heavily optimized using novel techniques. Table 1 presents run-time comparisons of the proposed and other standard depth diffusion schemes (for 320x240 images with error tolerance of 0.01). Optimization of other modules for real-time deployment is ongoing work.

Method	Time (sec)	System configuration
Our Scheme	0.048	Core 3.2 GHz, 512 MB
CGHS – MG	1.100	Core 3.2 GHz, 512 MB
YC04[25]	3.600	PIII 1.1 GHz
ZBV08[26]	21.50	PIV 3.2 GHz, 256 MB

Table 1. Run-time comparison of Depth diffusion schemes

V. Conclusion and Future Work

In this paper, we have demonstrated the benefits of a novel framework of fusing 2D local and global features such as edges, textures and regions with geometry information obtained from pixel-wise dense stereo for reliable 3D indoor structural scene representation. The strength of the approach is derived from the novel depth diffusion and segmentation algorithms that result in better surface characterization as opposed to traditional feature based stereo or RANSAC based plane fitting approaches. While the presented framework is related to other color/range sensor fusion algorithms such as [27] and [28], it should be noted that in the context of indoor 3D room reconstruction, the presented framework is highly efficient with extremely sparse range data, preserves and detects depth edges in regions where there are no visible edges in the color data and also handles shadows and specular highlights effectively, unlike [27] and [28]. While the complete framework has been presented with a focus on achieving low computational cost and proof of concept established for a reliable 3D indoor environment modeling system, actual deployment of the algorithm on a robot (using optimized GPU routines) for real-time operation and testing is the scope of ongoing work. It should however be noted that, it might be sufficient to have the framework update the model/map of the environment every few seconds or whenever a large change in the viewpoint is expected or observed.

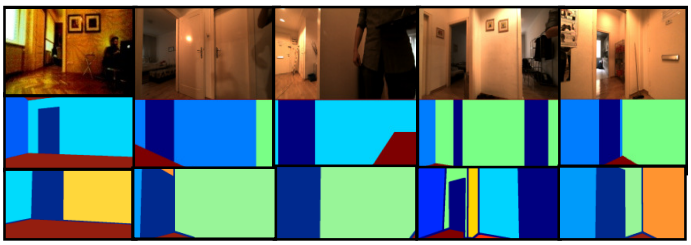

Figure 7. Results from test environment (Top to bottom) (a) Input scenes (b) Results of reprojection of the developed 3D model on to the image plane (c) Ground truth – manual labeling of doorway, wall and floor regions. The number of mislabeled pixels in the 5 test cases were 5%, 4%, 17%, 12% (due to large human occlusions) and 5% respectively.

References

[1] A. Bartoli, "A random sampling strategy for piecewise planar scene segmentation", Computer Vision and Image Understanding, 2007.

[2] C. Baillard, C. Schmid, et al., "Automatic line matching and 3D reconstruction of buildings from multiple views", ISPRS 1999.

[3] C. Baillard and A. Zisserman."A plane-sweep strategy for the 3D reconstruction of buildings from multiple images". ISPRS, 2000.

[4] C. Schmid and A. Zisserman."Automatic line matching across views". CVPR, 1997.

[5] S.C.Lee and R. Nevatia."Interactive 3D Building Modeling Using a Hierarchical Representation", ICCV2003.

[6] JK. Lee, S. Ahn et al., "A Prospective Algorithm for Real Plane Identification from 3D Point Clouds and 2D Edges", ICHIT '08.

[7] JK. Lee, S. Ahn et al., "Visibility-Based Test Scene Understanding by Real Plane Search", Advances in Visual Computing, Springer, 2008.

[8] C. Guerra et al."Line-based object recognition using Hausdorff distance from range images to molecular secondary structures" IVC05

[9] SH. Chang, S. Lee, D. Moon, et.al,"Model based 3D Object Recognition using Line Features", ICAR 2007.

[10] P. Biber, et al., "3D Modeling of Indoor Environments by a Mobile Robot with a Laser Scanner and Panoramic Camera", IROS 2004.

[11] P. Doubek, T. Svoboda, "Reliable 3D reconstruction from a few catadioptric images", OMNIVIS 2000.

[12] M. Nevado, et al. "Obtaining 3D models of indoor environments with a mobile robot by estimating local surface directions", RAS 2004.

[13] A. Dick, R. Cipolla et al. "Combining Single View Recognition & Multiple View Stereo for Architectural Scenes", ICCV 2001

[14] DD. Morris, T. Kanade, "Image Consistent Surface Triangulation", CVPR 2000.

[15] K Kutulakos, "Approximate N View Stereo", - Lecture Notes in Computer Science, 2000

[16] AC. Murillo, J. Kosecka, et al.,"Visual door detection integrating appearance and shape cues", RAS 2008.

[17] Z. Chen, S.T. Birchfield, "Visual Detection of Lintel-Occluded Doors from a Single Image", CVPRW 2008.

[18] R. Munoz-Salinas, E. Aguirre, M. Garcıa-Silvente, "Detection of doors using a genetic visual fuzzy system for mobile robots", Autonomous Robots, 2006.

[19] Y. Weiss, "Deriving intrinsic images from image sequences", ICCV, 2001.

[20] MF. Tappen, WT. Freeman, EH. Adelson, "Recovering intrinsic images from a single image", PAMI 2005.

[21] P. Perona and J. Malik, "Scale-space and Edge Detection Using Anisotropic Diffusion. PAMI,Vol.12, No.7,pp.629-639,1990.

[22] H. Wei, S. Zabuawala, KM. Varadarajan, et.al. "Adaptive pattern-based image compression for ultra-low bandwidth weapon seeker image communication", SPIE 2009.

[23] Aldo Dall'Osso, "An iterative back substitution algorithm for the solution of tridiagonal matrix systems with fringes", JCAM, 2003.

[24] A. Hoover, G. Jean-Baptiste, et al., "A Comparison of Range Image Segmentation Algorithms", PAMI, 1996.

[25] J Yin, JR Cooperstock, "Improving depth maps by nonlinear diffusion", WSCG, 2004.

[26] H Zimmer, A esBruhn, et al. "PDE-based anisotropic disparity-driven stereo vision", VMV 2008.

[27] J. Diebel and S. Thrun, "An application of Markov random fields to ranage sensing", NIPS 2005.

[28] J. Dolson, et.al. "Upsampling range data in dynamic environments", CVPR2010.