SENSOR DATA FUSION FOR THE ACQUISITION AND COMPRESSION OF RGBZ CONCENTRIC MOSAICS

Ingo Bauermann[1], Subhasis Chaudhuri[2], and Eckehard Steinbach[1]

[1]Media Technology Group, Institute of Communication Networks, Technische Universität München
[2]Department of Electrical Engineering, Indian Institute of Technology, Bombay

ABSTRACT

In this paper we present an algorithm for fusion of intensity and range data in the context of image based scene representation and compression. The fusion algorithm is embedded into an energy minimization framework incorporating active depth measurements using a 2D laser range scanner and passive geometry reconstruction from an image sequence. The joint disparity field is modeled as a Markov Random Field (MRF) and a globally optimal configuration is approximated using Bayesian Belief Propagation (BP). The recovered depth information is used for compression based on disparity compensated prediction. The improvement in terms of coding efficiency by utilizing an active range system in addition to a conventional imaging system is investigated. Results show that adding range data is valuable for calibration purposes and retrieving a global geometry model and also improves the coding efficiency.

1. INTRODUCTION

Based on the 7D plenoptic function introduced by Adelson and Bergen [7], image based rendering (IBR) is capable of representing a complex 3D real world scene in photo quality. Based on simple interpolation of reference images, pure IBR provides an efficient approach for modeling and rendering of complex 3D real world scenes without recovering the object's geometry. However, with the scene's structure known, a very efficient compression can be achieved by utilizing the geometric model for disparity compensated prediction [4]. Furthermore rendering is more accurate and view approximations can promptly be presented.

Geometric scene reconstruction techniques can be divided into an active and a passive group. The former one utilizes active systems like laser scanners based on time of flight measurements or active triangulation. In the latter group point correspondences in at least two 2D reference images are used for structural reconstruction via triangulation. Both approaches have their limitations and advantages partly complementing each other. The key aspects summarize as follows:

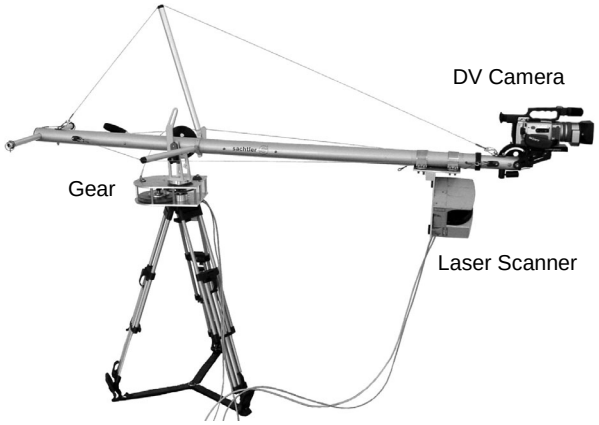

Figure 1: Device for joint image and depth data acquisition. A camera crane with gear moves the Sony DX2000 consumer camera and the SICK LMS290 laser range finder.

- **Noise.** Sensor noise, blurring, varying lighting conditions, and inaccurate calibration have to be taken into account.
- **Untextured regions.** While it is not possible to establish correspondences in untextured regions using visual sensors, active measurement systems using e.g. a laser beam can still provide reliable estimates.
- **Depth discontinuities and occlusions.** Flat areas and discontinuities should be preserved. Active systems can provide sharp edges while passive systems are sensitive to mismatches near object boundaries because of occlusions.
- **Physical limitations.** While active sensors mostly provide a low spatial resolution and samples often are subject to systematic errors, imaging systems allow for a dense sampling of intensity values.

For a more detailed analysis of the two approaches see e.g. [3],[5]. With respect to the given issues we investigate fusion of intensity and range data in order to get an improved scene model for fast view approximation and disparity compensated prediction in the context of image based rendering and compression. An algorithm for combining both is given.

The remainder of this paper is structured as follows. Chapter 2 gives some notes on related work. In Chapter 3

we briefly describe the system setup and the resulting capture geometry. Chapter 4 discusses the energy minimization technique used to reconstruct scene geometry and the fusion algorithm. Results are presented in Chapter 5. Some notes on the learned lessons are given in Chapter 6. Chapter 7 concludes the paper.

2. RELATED WORK

Much work has been done on geometry reconstruction from image pairs and sequences (see e.g. [9] for an extensive comparison). Recently graph cut and belief propagation algorithms emerged in stereo and allow for excellent reconstruction of scene geometry with a relatively low computational complexity by modeling depth maps as markov random fields [9]. Also much work has been done in geometry aided coding of video and image based scene representations (e.g. [4]). Models based on object shape, volumetric reconstruction, or implicit disparity information are used for compression of object light fields. In most cases geometry retrieval is done using intensity based methods. Where both active and passive methods are used simultaneously, geometry retrieval from images is performed using feature based methods. Only sparse depth maps are produced (e.g. [3]). In our work a dense joint disparity field is considered and approximated.

3. SYSTEM SETUP

An acquisition device assembled of a SICK LMS290 laser range finder and a consumer DV camera SONY VX2000 is mounted on a camera crane as shown in Figure 1 and captures a concentric mosaic [8]. The video camera provides high quality DV images (720x576) with a field of view of approximately 40 degrees. The scanner provides 50 range samples per scan along one scanline covering 50 degrees of vertical field of view. Figure 2 shows the capture geometry.

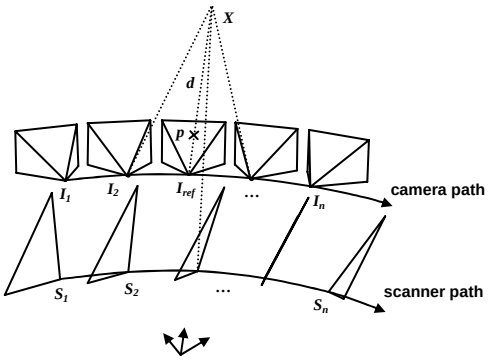

Figure 2: Capture geometry for joint acquisition of intensity and range data. The dotted lines denote corresponding projections of one scene point.

A scene point *X* has multiple correspondences in the images taken along the camera path and shows up once in the range data captured along the scanner path. Depth *d* can be determined for a pixel ***p*** with known intrinsic and extrinsic calibration parameters of the acquisition system determined using the method described in [1].

4. ENERGY MINIMIZATION AND DATA FUSION

The general framework for the minimization problem we look at is defined as follows. For a set *P* of pixels and a set *D* of depth labels, find the labeling *f* that assigns a label $d \in D$ to each pixel $\boldsymbol{p} \in P$ so that the global energy

$$E(f) = \sum_{\boldsymbol{p} \in P} C(\boldsymbol{p}, d) + \sum_{(\boldsymbol{p},\boldsymbol{q}) \in N} V(d_{\boldsymbol{p}}, d_{\boldsymbol{q}}) \qquad (1)$$

is minimized. Here, ***q*** is a pixel in a local neighborhood *N* of ***p***. *C* is the local matching cost for pixel ***p*** if a depth corresponding to depth label *d* is assigned to ***p***. *V* describes the disparity cost of two neighboring pixels having different depths.

4.1. Data cost

C consists of the intensity matching cost C_I and the scanner matching cost C_S.

$$C(\boldsymbol{p}, d) = \lambda_I \cdot C_I(\boldsymbol{p}, d) + \lambda_S \cdot C_S(\boldsymbol{p}, d) \qquad (2)$$

Here, λ_I weighs the intensity matching cost and λ_S is a regularization parameter to weigh passive and active geometry retrieval. λ_S is set to zero whenever no active measurement is available, otherwise it is set to one.

4.1.1 Intensity matching cost

The intensity matching cost function

$$C_I(\boldsymbol{p}, d) = \frac{1}{|S|-1} \sum_{r \in S} \left(I_r(\boldsymbol{p}, d) - \bar{I}(\boldsymbol{p}, d) \right)^2 + \lambda_{ref} \cdot \max_{r \in S} \left(I_r(\boldsymbol{p}, d) - I_{ref}(\boldsymbol{p}) \right)^2 \qquad (3)$$

uses an image set *S* which contains all frames in which ***p*** is visible. This set of images is determined by iteratively calculating a depth map with respect to an occlusion map, and vice versa, similar to the method described in [6]. The first term of C_I takes care of the color consistency and is the variance of the intensity I_r of pixel ***p*** projected into any frame *r* of a sequence given a depth *d*. $\bar{I}$ denotes the mean intensity of projections of ***p*** in *S*. The second term addresses the deviation of the intensity I_{ref} in a reference frame compared to the observed intensity in any other frame *r*. λ_{ref} controls the weight of the reference frame for the intensity controlled part of the data cost.

4.1.2. Scanner cost

The second term in Equation (2) takes care of the measurements from the laser scanner.

The scanner cost is defined as:

$$C_S(\boldsymbol{p},d) = min\left(\left\|d_{Scanner}(\boldsymbol{p}) - d\right\|, d_{S\,max}\right) \qquad (4)$$

This is a truncated linear model and defines a distance measure between actively measured depth $d_{Scanner}(\boldsymbol{p})$ and passively determined depth d. The parameter d_{Smax} limits the maximum cost for a mismatch.

4.2. Disparity cost

The second part of Equation (1) defines the cost for discontinuities and can also be referred to as the smoothness term. V denotes the cost for assigning depth labels to neighboring pixels:

$$V\left(d_{\boldsymbol{p}},d_{\boldsymbol{q}}\right) = min\left(\lambda_v(d_{\boldsymbol{p}},d_{\boldsymbol{q}})\cdot\left\|d_{\boldsymbol{p}} - d_{\boldsymbol{q}}\right\|, d_{max}\right) \qquad (5)$$

Again, a truncated linear model is used as a distance measure between the depth for neighboring pixels d_p and d_q with d_{max} as the maximum cost for a depth discontinuity. Additionally, a color segmentation is performed by incorporating $\lambda_v(d_p,d_q)$ as a regularization of the smoothness dependent on the difference in normalized intensity of neighboring pixels:

$$\lambda_v(d_{\boldsymbol{p}},d_{\boldsymbol{q}}) = a_v\cdot\left(1 - \left\|I(\boldsymbol{p}) - I(\boldsymbol{q})\right\|\right) \qquad (6)$$

Here, a_v defines the weight of the color segmentation. This term causes discontinuities to be more probable at segment boundaries.

4.3. Fusion

By carefully choosing the weighting factors for the intensity and scanner costs a joint disparity field is modeled as a MRF. Minimizing the energy E corresponds to the MAP estimation problem for this MRF. An optimal labeling f and therefore an optimal depth map can be approximated using graph cut algorithms or Bayesian belief propagation algorithms such as described in [9]. In this work we use a modified implementation of the belief propagation algorithm described in [2].

5. RESULTS

To evaluate our algorithm we perform geometry reconstruction and sensor data fusion for an image sequence consisting of 61 frames spanning about 20 degrees of a concentric mosaic with a rotation radius of 1.3m. This results in an effective baseline of about 0.45m between the first and last frame of the sequence. In the two experiments we performed, the number of depth labels used was 71 in a distance of 0.1m each, ranging from 1m to 8m measured from the center of projection of frame 31.
In the first experiment, scene reconstruction is based on scanner measurements and all available intensity information of the 61 frames. The scene was reconstructed by image interpolation of the first and last frame of the sequence based on a retrieved depth map. The two reference images and the depth map as well as the residual errors are encoded in a rate scalable manner using JPEG2000 [10]. A rate-distortion plot using four different depth maps is given in Figure 3. A constant depth assumption, chosen according to the mean scene depth, is the standard way for rendering concentric mosaics [8]. The depth map only acquired from the scanner data was enhanced for a fair comparison as holes due to device parallax were compensated for by interpolation. The result only using image data performs about 1dB better than only using range data, while the fused version of the scene geometry gives another gain of about 0.2 dB. The parameter set according to the notation in Chapter 4 for the different scenarios is given in table 1.

	λ_I	λ_S	λ_{ref}	a_v	d_{Smax}	d_{max}
Fused	0.5	0.5	0.5	4	1	2
Scanner	0	1	0	λ_v=1	1	2
Image	1	0	0.5	4	0	2

Table1. Parameters for the performed experiments.

The second experiment considers image extrapolation and is performed by only reconstructing the geometric model from frames 25 to 35, but using this model to predict the whole image sequence. The rate distortion performance using these new depth maps is shown in Figure 4. The fused model performs up to 4dB better than using a constant depth and up to 1db better than just using range or intensity data.
A comparison of both experiments shows that the fused model only gives a small gain over the geometric model retrieved only from images when used for image interpolation. For image extrapolation this gain is significant.
Figure 5 shows parts of the acquired depth maps along with the original data and reconstructed data. The higher resolution of the fused depth map compared to the depth recovered from intensity is obvious in rather flat areas. Furthermore, specularities are covered as geometry deformation.

6. LESSONS LEARNED

The proposed algorithm shows robustness to sensor noise and varying lighting conditions during image acquisition. Untextured regions are recovered correctly due to the MRF model and its global approximation, and the support of range scanner data. Even small details are recovered with a careful choice of system parameters. The good performance of intensity based scene reconstruction is only possible due to the accurate calibration which again relies on the range data as shown in [1]. Though the range scanner also provides reflectance data which could be used for a reliability measure of λ_S, our experiments show that this does not give a gain in reconstruction quality. Though the compression performance is not as good as frame to

frame motion compensation algorithms, random access can be provided to a view using our approximation.

7. CONCLUSIONS

In this paper we present an algorithm for fusion of intensity and range data. Our fusion algorithm is embedded into an energy minimization framework and a globally optimal depth map is approximated using Bayesian Belief Propagation (BP). Our experiments show that the improvement in terms of coding efficiency by utilizing an active range system for concentric mosaics is about 1dB for image extrapolation and 0.2dB for image interpolation. Future work will include the optimal choice of reference images and the distribution of depth labels.

8. REFERENCES

[1] I. Bauermann and E. Steinbach, "Joint Calibration of Video and Range Data for the Acquisition of RGBZ Concentric Mosaics," VMV 2005, accepted for publication.

[2] P.F. Felzenszwalb and D.P. Huttenlocher, "Efficient belief propagation for early vision," CVPR04, pp. 261-268, 2004.

[3] P. Dias, V. Sequeira, J. Gonçalves, F. Vaz, "Registration and Fusion of Intensity and Range Data for 3D Modeling of Real World Scenes," In proceedings of Fourth International Conference on 3-D Digital Imaging and Modeling, pp. 418-425, 6-10 October 2003.

[4] B. Girod, P. Eisert, M. Magnor, E. Steinbach, and T. Wiegand, "3-D image models and compression—Synthetic hybrid or natural fit?," in Proc. Int. Conf. Image Processing (ICIP'99) Kobe, Japan, pp. 525-529, Oct. 1999.

[5] S.F. El-Hakim, J.A. Beraldin, F. Blais, "A Comparative Evaluation of the Performance of Passive and Active 3-D Vision Systems," SPIE Proc. 2646, St.Petersburg Conf. on Digital Photogrammery, pp. 14-25, June 1995.

[6] J. Sun, Y. Li, S. B. Kang, and H.-Y. Shum, "Symmetric stereo matching for occlusion handling," CVPR 2005.

[7] E. H. Adelson and J. Bergen, "The plenoptic function and the elements of early vision," Computational Models of Visual Processing, pp. 3–20, MIT Press, Cambridge, MA, 1991.

[8] H. Shum and L. He, "Rendering with concentric mosaics," Computer Graphics (SIGGRAPH '99), pp. 299-306, Aug. 1999.

[9] http://cat.middlebury.edu/stereo/

[10] D. Taubman and M. Marcellin, "JPEG 2000: Image Compression Fundamentals, Standards and Practice," Kluwer International Series in Engineering and Computer Science, Secs. 642, Nov. 2001.

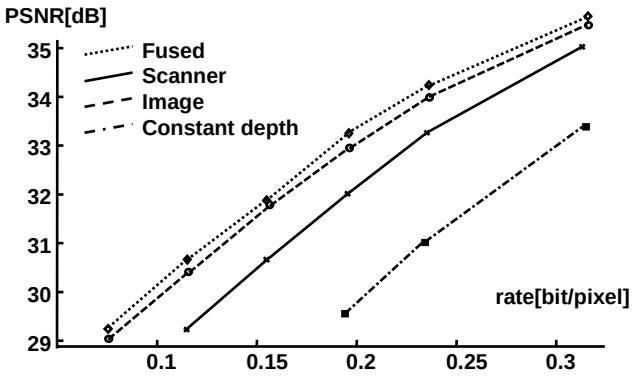

Figure 3: Rate-Distortion curve for encoding a 20° sector of the concentric mosaic consisting of 61 images using different depth maps. All images are used for geometry reconstruction.

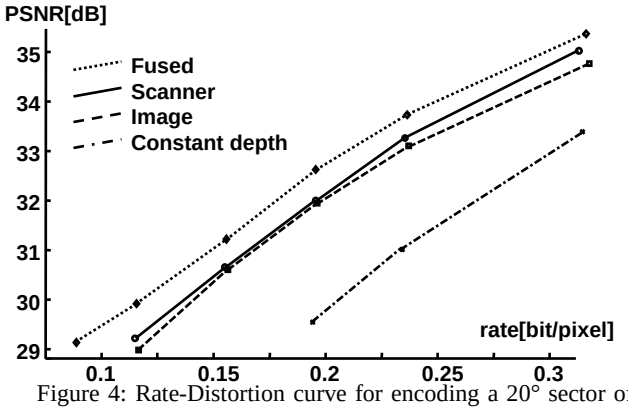

Figure 4: Rate-Distortion curve for encoding a 20° sector of the concentric mosaic consisting of 61 images using different depth maps. Only 11 frames in the middle of the sequence are used for geometry reconstruction.

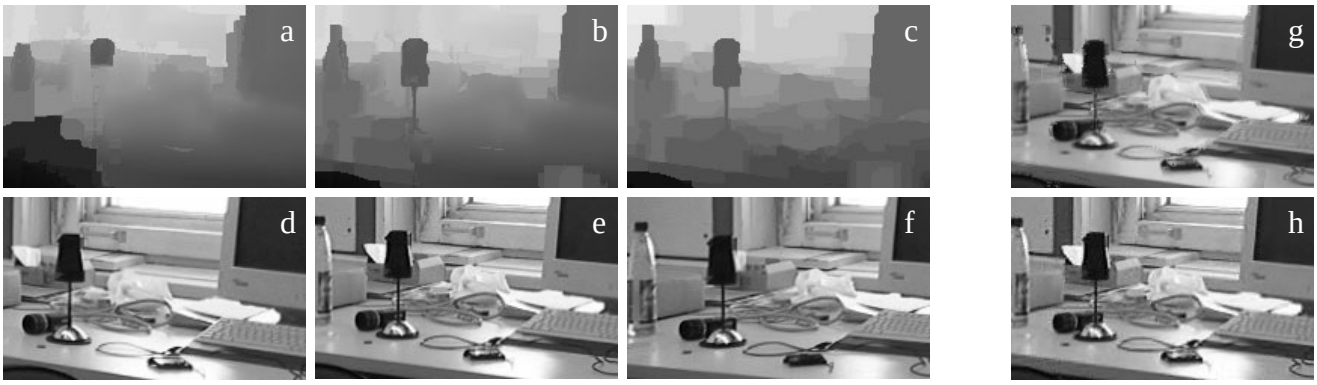

Figure 5: a) shows one part of the scanner depth map, warped to the reference image; b) is a cutout of the fused depth map using our algorithm; c) same cutout of the depth map calculated onlz from the images; d),e),f) show original first, reference, and last frames of the sample sequence (cutouts); g) gives the prediction of e) at 0.04bit/pixel and 25.6dB; h) same image at 0.15bit/pixel and 31.7dB;