MVA2007 IAPR Conference on Machine Vision Applications, May 16-18, 2007, Tokyo, JAPAN

Evaluation of Stereo Matching Systems for Real World Applications Using Structured Light for Ground Truth Estimation*

Martin Humenberger
Austrian Research Centers GmbH - ARC, Vienna
martin.humenberger@arcs.ac.at

Daniel Hartermann
TU Vienna
Austria
hartermann@acin.tuwien.ac.at

Wilfried Kubinger
Austrian Research Centers GmbH - ARC, Vienna
wilfried.kubinger@arcs.ac.at

Abstract

In this paper we present an evaluation method for stereo matching systems and sensors especially for real world indoor applications. We estimate ground truth reference images by illuminating scenes with structured light. The paper starts with the selection of appropriate scenes, goes over ground truth estimation to the finally resulting evaluation of the stereo sensors. Three different stereo vision sensors are tested with three different input scenes. Beside the subjective review of the disparity map images a pixel wise evaluation is carried out. We mainly describe an evaluation method for existing stereo sensors which should help developing new ones.

1 Motivation

The need of dense 3D information to support robots and autonomous systems is increasing permanently. Service robots require dense depth information to fulfill their tasks [4] and intelligent vehicles rely on accurate sensor information to support navigation along tracks and avoid obstacles [1]. Stereo vision is a dedicated technology to deliver dense depth information and therefore plays a key role in sensing the environment of a robot or vehicle to get reliable 3D data.

Consequently, our research on stereo vision for robotic and embedded applications led to the need of an evaluation method for the reliability of 3D information. Checking the correctness of a disparity map is not simple because normally there is no reference dataset ("ground truth") available. The acquisition of such a high-precision depth map can be very complex and time intensive. There are already several platforms available, like the good and well known one from Middlebury College [6] which tries to solve this problem providing input images for stereo algorithms with the corresponding dense ground truth disparity maps. On the one hand side this gives a good possibility to quantify disparity maps in respect to others but on the other hand side there is no flexibility in the choice of input images or scenes. In robotic applications, researchers want to know how accurate their 3D sensors work in real world environments. Here, the classical difficulties of stereo matching, like textureless surfaces, occlusions and reflections get in focus again. This issue led us to develop an evaluation platform which attaches importance to flexibility in scene selection. The goal of our approach is to find a possibility to quantify 3D-data produced by different stereo sensors under the aspect of a setup placed in natural indoor environments.

*This research has been supported by the European Union projects MOVEMENT under grant #IST-2003-511670 and ROBOTS@HOME under grant #FP6-2006-IST-6-045350 and the Austrian Science Foundation grant S9101-N04.

2 Previous Work

We did not want to produce any results with images optimized for stereo algorithms or synthetic generated scenes, but we needed an easy-to-obtain quality metric to evaluate results of stereo vision systems. We looked for a low-budget and portable solution without the need of special equipment. There are just a few low-cost techniques dealing with this issue, i.e. the prediction error as a metric [7], Self-Consistency [3] or a pixel-wise comparison with a "ground truth" data set. We decided to use a method based on "ground truth", so the problem is how to obtain such a high accuracy depth map without wasting a lot of time in hand labeling. Finally we found an adequate technique in [5] which fits for our requirements. The idea behind this method is to use structured light to uniquely label each pixel in a sequence of images, so that the correspondence becomes trivial.

In further sections, we use the term "ground truth" for denoting ground truth estimation. This means that the ground truth images, created by our platform, are reliable and very dense, but not complete. In this paper, the term "stereo (vision) sensor" always refers to the systems (sensors) described in sec. 3.3.

3 Workflow

The workflow of our evaluation approach consists of four different steps and is illustrated in fig. 1. At first, a suitable scene is selected. Then a pair of stereo images of this scene, as input for the stereo sensor systems, is recorded. After that, the scene is illuminated by the structured light patterns and captured accordingly. In the next step these images are used for our ground truth approximation. This results in a rather good disparity map which is the basis for evaluation of the stereo vision systems in the last step.

3.1 Scene selection

Because of the fact that we wanted to use scenes as close to reality as possible, labor environments should be excluded. The acts of the scenes should be typical camera views of a robot for indoor applications. This

means the main objects a robot has to deal with are chairs, tables, many kinds of boxes, and furniture in general. We tried to catch as many common indoor obstacles as possible to build realistic scenes for our evaluation. All scenes were recorded in our office under real world conditions (e.g. no controlled illumination).

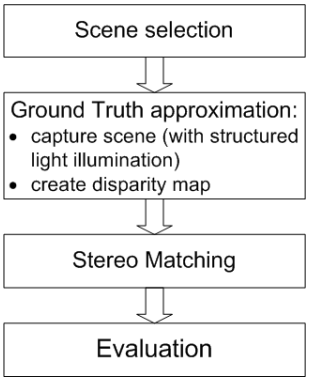

Figure 1: Workflow

3.2 Ground truth estimation

As mentioned above the goal of our approach is to gain depth information from real world scenes using two cameras for ground truth estimation. In our setup we project binary gray-code patterns with a NEC LT245 video beamer to a scene and take the images with a stereo rig which is described in the next section. The image capture setup is shown in fig. 2. The

Figure 2: Image capture setup.

processing pipeline consists of the following stages:

- Camera calibration, acquisition and rectification of each stereo image pair using the "Caltech Toolbox" for Matlab
- Decode the light patterns for each illumination source to get unique codes for each pixel and computation of disparity map
- Creation of a LUT (lookup table) consisting of correspondences between each view and each illumination source, determination of the projection matrices to reproject the code labels using the LUT
- Combine all disparities

Currently we use ten patterns to encode the horizontal and ten to encode the vertical component. Due to the limited quality of the camera images we decided to use also inverse pattern to compute the stripes[1]. This doubles the required number of images. For reliable detection of shadow areas we use three more images (black, white and gray illumination). The results of the decoding process are strongly depend on the scene. Fogging inside the projector, inter reflections and highly varying albedos in the scene are causing wrong code values. Therefore we use different methods to enhance the quality of the binarized images. For determination of the position of the stripes, we use linear interpolation [8]. The detection of stripe-edges is also possible within sub pixel-accuracy. These edges are used to smooth the stripes in the first estimation of the binary images by a bicubic interpolated (in the appropriate code direction) "greater or lesser" comparison over all color channels of the normal and inverse illuminated image. Afterwards it is downscaled with a majority filter and the pixels which contain shadow or have not sufficient contrast to its inverse are labeled as unknown. The disparities $d(u, v)$ (left to right and vice versa) are computed with a robust block matching algorithm. As input for the next step only pixels which passed a left-right consistency check are taken. Using the DLT-method (Direct Linear Transformation) which is based on the pinhole camera model and treats the first disparity map as a 3D-reconstruction of the scene, we calculate the projection matrix P between each view and its illumination source. Note that the 3D-reconstruction is obviously registered with the proper view. We can solve the parameters $a11, ..., a34$ of the DLT (shown in equ. 1), in a Linear Least Square-sense, by eliminating wi and using a constraint to avoid the trivial solution.

$$\begin{pmatrix} u_i w_i \\ v_i w_i \\ w_i \end{pmatrix} = \begin{pmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ a_{21} & a_{22} & a_{23} & a_{24} \\ a_{31} & a_{32} & a_{33} & a_{34} \end{pmatrix} \begin{pmatrix} x_i \\ y_i \\ d(u_i, v_i) \\ 1 \end{pmatrix} \quad (1)$$

The constraint $a34 = 1$ as proposed in [5] produces the best fit. In practice, a small number of pixels with large disparity errors can strongly affect the fit. We therefore use an iterative algorithm to detect outliers. The next step is to compute a new set of disparities with equ. 1 and the LUT for all illuminated pixels. This includes pixels, which are not visible from one view (left or right occluded). Now we repeat the whole process with a different video projection position (the stereo rig remains untouched) to reduce shadow areas and to calculate 3D data in occluded areas. At last we combine all disparity maps to create a robust result.

3.3 Stereo sensors and matching

Following stereo vision systems are evaluated with our approach under real world indoor conditions.

1. Point Grey Research: Digiclops[2], Triclops[3] ©
2. Videre Design: Small Vision System (SVS) [2]
3. Our own stereo vision sensor (still work in progress)

[1] binary images created from the captured images of the illuminated scenes

[2] http://www.ptgrey.com/products/digiclops/Digiclops.pdf

[3] http://www.ptgrey.com/products/triclopsSDK/triclops.pdf

Table 1: Image parameters.

	Sensor 1	Sensor 2	Sensor 3
image size	640×480	640×480	640×480
baseline	10 cm	12 cm	12 cm
focal length	6 mm	6 mm	6mm
block size	11	21	31[4]
disparity range	0 - 40	40 - 168	10 - 140
matching alg.	SAD	SSD	census

Ad 1): Point Grey Digiclops is a three-camera stereo vision system which uses a SAD [6] correlation algorithm for stereo matching. For our tests, we used the three-camera system with the appropriate software (Triclops). Ad 2): Videre Design Small Vision System is a stereo vision platform consisting of a two-camera stereo head and a stereo software package. For our tests, we used the software only, which is the implementation of an area SSD [6] correlation algorithm. Ad 3): The third stereo vision system is our own and is still work in progress. We use a 12cm baseline stereo head consisting of two The Imaging Source GmbH FireWire, color, 640×480 cameras and our software uses the census transformation [9] for neighborhood dependent matching. These cameras were also the input source for the SVS software and the ground truth estimation.

As mentioned above sensor 1 is a complete standalone system. It has its own stereo head and software (including matching and calibration). Sensors 2 and 3 both use images captured from sensor 3. It can be assumed that these cameras are calibrated (Caltech Camera Calibration Toolbox for Matlab [**?**]) and that the images are rectified.

The matching procedure takes a rectified stereo image pair as input, processes it with the proper matching algorithm and finally saves a disparity map (8 bit grayscale bitmap) as output. The output disparity map will be evaluated according to our ground truth.

Tab. 1 lists the most important parameters of the three sensors. The blocksize describes the size of the window in block matching and is chosen accordingly to the best sensor results. Remarkable in tab. 1 is the big difference in the disparity range of sensor 1 compared to the others. The reasons for this are the different optics and baseline.

4 Evaluation

After obtaining the ground truth images and all disparity maps of the stereo sensors, the next step is an evaluation of these results.

Fig. 3 and fig. 4 illustrate our work on two different real world scenes. The image on the top left of the figures shows the left stereo image taken with sensor 3. The ground truth image can be found beside it. The middle image row shows the disparity maps of sensor 2 (left) and sensor 3 (right). In the last row, the results from sensor 1 are shown (left stereo image and disparity map).

Black pixels in ground truth indicate areas without illumination, areas without sufficient contrast between normal and inverse pattern illumination or areas where the decoding process failed. Notably is that occlusions between the left and right view are eliminated. This

[4]census transform: 31, block matching: 5

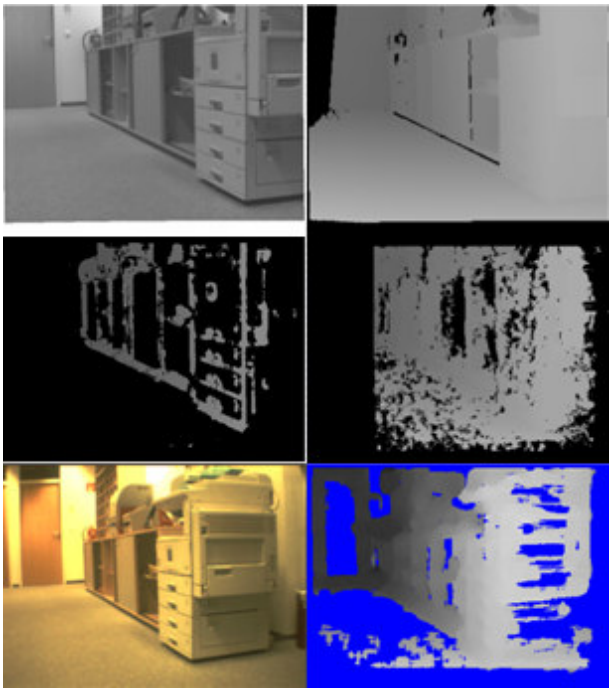

Figure 3: Scene "copier". top: left stereo image and ground truth image; middle: disparity maps of sensor 2 (left) and sensor 3 (right); bottom: results from sensor 1.

can be observed in the "chair" scene at the edges of the boxes. In the disparity images pixels are black where no correspondence according to the used algorithm could be found. Tab. 2 shows our results. It consists of three different scenes. We excluded sensor 1 from the pixel-comparison, because the available version of the software does not support an offline modus without additional implementations. The internal scaling of the disparities of sensor 2 also causes some problems. We had to treat it like a black box and used a linear model to map these disparities according to our ground truth disparities. We denote sensor disparities (disps) as *valid* if they are the same, within a range (+/- threshold in pixels), as the ground truth disps. Ground truth disps are *available* if they are not zero (not black) in the disparity map. Equally, sensor disps are *found* if they are not zero (not black) in the result disparity map.

Column *% present disps* shows the ratio between all *found* sensor disparities and all *available* ground truth disparities. It gives a general information about the density of the sensor disparities.

Column *% valid of all gt disps* shows the ratio between sensor disps and all *available* ground truth disps without respect if they were *found* by the sensor or not. It answers the question: How many of the available GT disps were found by the sensor?

Column *% valid of found disps* shows the ratio between all *found* sensor disps and all *valid* sensor disps. It answers the question: How many of the found sensor disps are valid?

In the values of the first two columns is a negative offset of pixels that are available in ground truth but not visible in the results, included. This offset can be seen on the borders of the result images and can be ascribed to the disparity search range, the blocksize and the camera field of view. To create offset independent

Table 2: Evaluation results.

Scene	% present disps		Threshold	% valid of all gt disps		% valid of found disps	
	Sensor 2	Sensor 3		Sensor 2	Sensor 3	Sensor 2	Sensor 3
couch	13.17	38.4	6	12.43	33.8	94.37	88.01
			2	11.7	3.3	89.1	8.65
copier	15.8	42.23	6	14.56	38.6	92.08	91.43
			2	11.35	25.76	71.78	61
chair	6.86	30.39	6	1.33	22.98	19.4	75.6
			2	0.44	14.16	6.46	46.58

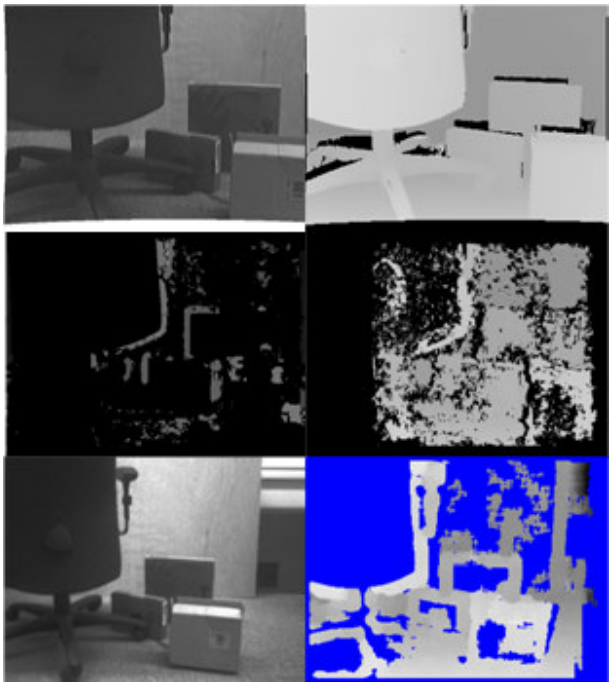

Figure 4: Scene "chair". top: left stereo image and ground truth image; middle: disparity maps of sensor 2 (left) and sensor 3 (right); bottom: results from sensor 1.

results, the disparity maps have to be cut accordingly. The results presented in tab. 2 can be interpreted as follows.

In the couch scene sensor 3 found more disps than sensor 2 but the results of sensor 2 are much more reliable. Sensor 2 handled the difficulties of this scene better than sensor 3. In the chair scene, sensor 3 wins the race. It delivers more dense disparity maps and they are also more reliable. The copier scene also goes to sensor 3. The reliability is a little worse but the results are much more dense. As a consequence, the indicator for a reliable stereo vision system is a high value in the last column and for a dense stereo sensor a high value in the first column. It is difficult to determine the best sensor for a certain application but our evaluation approach makes the decision a little bit easier.

5 Conclusion and Outlook

In this paper we presented our approach of evaluating stereo vision systems using estimated ground truth images of real world scenes. The ground truth images are created by illuminating the scenes with structured light, coding each pixel, stereo matching these code words and optimizing the result. After ground truth creation the evaluation of stereo sensors follows. Besides a subjective reviewing of the resulting images a pixel wise evaluation is used. During our work we came across many problems which highlighted some future work to do. First there is the calculation speed of the ground truth disparity maps. The solution is implemented in Matlab and therefore not speed optimized. Secondly the cameras and optics are candidates for further improvements. Also the evaluation itself could be more comprehensive, for example the scale problem could be solved. At last, an inclusion of matching speed to evaluate real-time requirements would be useful.

References

[1] R. Daily, W. Travis, D. M. Bevly, K. Knoedler, R. Behringer, H. Hementsberger, J. Kogler, W. Kubinger, and B. Alefs. Sciautonics-auburn engineering's low-cost, high-speed atv for the 2005 darpa grand challenge. *Journal of Field Robotics*, 23(8):579–597, 2006.

[2] K. Konolige. Small vision system: Hardware and implementation. In *Proceedings of Eighth International Symposium on Robotics Research*, Japan, 1997.

[3] Y. Leclerc, Q. Luong, and P. Fua. Self-consistency: A novel approach to characterizing the accuracy and reliability of point correspondence algorithms. In *Proceedings of Image Understanding Workshop*, pages 793–807, 1998.

[4] D. Murray and J. J. Little. Using real-time stereo vision for mobile robot navigation. *Autonomous Robots*, 8(2):161–171, 2000.

[5] D. Scharstein and R. Szeliski. High-accuracy stereo depth maps using structured light. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE Computer Society Press, 2003.

[6] D. Scharstein, R. Szeliski, and R. Zabih. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. In *Proceedings of the IEEE Workshop on Stereo and Multi-Baseline Vision*, Kauai, HI, 2001. IEEE Computer Society Press.

[7] R. Szeliski. Prediction error as a quality metric for motion and stereo. In *Proceedings of the Seventh International Conference on Computer Vision*, 1999.

[8] M. Trobina. Error model of a coded-light range sensor. In *Technical Report BIWI-TR-164, ETH-Zentrum*, 1995.

[9] R. Zabih and J. Woodfill. Non-parametric local transforms for computing visual correspondence. In *Proceedings of 3rd European Conf. Computer Vision*, pages 151–158, Stockholm, 1994.