Proceedings of 3rd Iberian Conference on
Pattern Recognition and Image Analysis (IbPRIA)
June 6 - 8, 2007, Girona, Spain

Mosaicking Cluttered Ground Planes Based on Stereo Vision

José Gaspar[1], Miguel Realpe[2], Boris Vintimilla[2], José Santos-Victor[1]

[1]Computer Vision Laboratory
Inst. for Systems and Robotics
Instituto Superior Técnico
Lisboa, Portugal
{jag,jasv}@isr.ist.utl.pt

[2]Vision and Robotics Center
Dept. of Electrical and Computer Science Eng.
Escuela Superior Politécnica del Litoral
Guayaquil, Ecuador
{mrealpe, boris.vintimilla}@fiec.espol.edu.ec

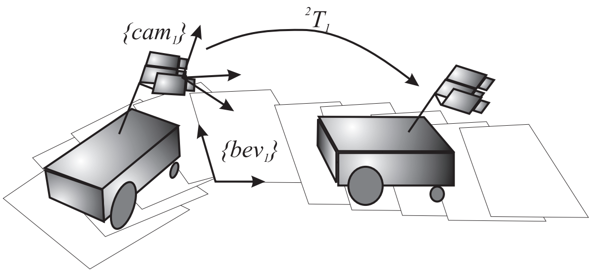
Abstract. Recent stereo cameras provide reliable 3D reconstructions. These are useful for selecting ground-plane points, register them and building mosaics of cluttered ground planes. In this paper we propose a 2D Iterated Closest Point (ICP) registration method, based on the distance transform, combined with a fine-tuning-registration step using directly the image data. Experiments with real data show that ICP is robust to 3D reconstruction differences due to motion and the fine tuning step minimizes the effect of the uncertainty in the 3D reconstructions.

1 Introduction

In this paper we approach the problem of building mosaics, i.e. image montages, of cluttered ground planes, using stereo vision on-board of a wheeled mobile robot. Mosaics are useful for the navigation of robots and for building human-robot interfaces. One clear advantage of mosaics is the simple representation of robot localization and motion: they are simply 2D rigid transformations.

Many advances have been made recently in vision based navigation. Flexible (and precise) tracking and reconstruction of visual features, using particle filters, allowed real time Simultaneous Localization and Map Building (SLAM) [1]. The introduction of scale-invariant visual features brought more robustness and allowed very inexpensive navigation solutions [2, 3]. Despite being effective, these navigation modalities lack building dense scene representations convenient for intuitive human-robot interfaces. Recent commercial stereo cameras came to help by giving locally dense 3D scene reconstructions. Iterative methods for matching points and estimation their rigid motion, allow registering the local reconstructions and obtaining global scene representations. The Iterated Closest Point (ICP) [4] is one such method that we explore in this work.

The ICP basic algorithm has been extended in a number of ways. Examples of improvements are robustifying the algorithm to the influence of features lacking correspondences or using weighted metrics to trade-off distance and feature similarity [5]. More recent improvements target real time implementations, matching shapes with defects or mixing probabilistic matching metrics with saturations to minimize the effect of outliers [6–8]. In our case, the wheeled mobile



$$\begin{tikzcd}
\{cam_1\} \arrow[r, "{}^{2}T_1"] \arrow[d, "P"] & \{cam_2\} \arrow[d, "P"] \\
\{img_1\} \arrow[d, "H"] & \{img_2\} \arrow[d, "H"] \\
\{bev_1\} \arrow[r, "{}^{2}H_1"] & \{bev_2\}
\end{tikzcd}$$

Fig. 1. Mosaicking ground planes: Stereo camera, Image and BEV coordinate systems.

robots motion on the ground plane allows searching for 2D, instead of 3D, registrations. Hence we follow a 2D ICP methodology, but we take a computer vision approach, namely registering clouds of points using the distance transform [9].

Stereo cameras allow selecting ground-plane points, registering them and then building the ground plane mosaic. Stereo reconstruction is therefore an advantage, however some specific issues arise about its use. For example, the discrete nature of the imaging process, and the variable imaging of objects and occlusions due to robot motion, imply uncertainties on the 3D reconstruction. Hence, the registration of 3D data propagates also some intrinsic uncertainty. The selection of ground-plane data, is convenient for complexity reduction, however a question of the sparsity of data arises. In our work we investigate robust methodologies to deal with these issues, and in particular we investigate whether resorting to the raw image data can help minimizing error propagation.

The paper is structured as follows: Sec.2 details the mosaicking problem and introduces our approach to solve it; Sec.3 shows how we build the orthographic views of the ground plane; Sec.4 details the optimization functionals associated to mosaic construction; Sec.5 is the results section; Finally in Sec.6 we draw some conclusions and guidelines for future work.

2 Problem Description

The main objective of our work is mosaicking (mapping) the ground plane considering that it can be cluttered with objects such as furniture. The sensor is a static trinocular-stereo camera mounted on a wheeled mobile robot. The stereo camera gives 3D clouds of points in the camera coordinate system, i.e. a mobile frame changed by the robot motion. See Fig.1

The ground plane constraint implies that the relationships between camera coordinate systems are 2D rigid motions. As in general the camera is not aligned with the floor, i.e. the camera coordinate system does not have two axis parallel to the ground plane, the relationships do not clearly show their 2D nature. In order to make clear the 2D nature of the problem, we define a new coordinate system aligned with the ground plane (three user-selected well-separated ground points are sufficient for this purpose).

Commercial stereo cameras give dense reconstructions. For example, for each image feature, such as a corner or an edge point, there are usually about 20 to 30 reconstructed 3D points (the exact number depend on the size of the correlation windows). Considering standard registration methods as Iterated Closest Point (ICP, [4]), the large clouds of 3D points imply large computational costs. Hence, we choose to work with a subset of the data, namely by selecting just points of the ground plane. The 2D clouds of points can therefore be registered with a 2D ICP method.

Noting that each 3D cloud of points results from stereo images registration, the process of registering consecutive clouds of points has some error propagated from the cloud reconstruction. In order to minimize the error propagation, we add a fine tuning image-based registration process after the initial registration by a 2D ICP method. The image-based registration is a 2D rigid transformation in Bird's Eye Views (BEV), i.e. orthographic images of the ground plane. BEV images can be obtained also knowing some ground points and the projection geometry. To maintain consistent units system, despite having metric values in the 3D clouds of points, we choose to process both the 2D ICP and the image registration in the pixel metric system, i.e. the same as the raw data.

In summary our approach encompasses two main steps: (i) selection of ground points and 2D ICP, (ii) BEV image registration. Despite the 2D methodology notice that the 3D data is a principal component. The 3D sensor allows selecting the ground plane data, which is useful not only for using a faster 2D ICP method but mainly for registering the ground plane images without considering the distracting (biasing) non-ground regions.

3 Obtaining Bird's Eye Views (BEV)

The motion of the robot implies a motion of the trinocular camera which we denote as ${}^2T_1$. The indexes 1 and 2 indicate two consecutive times, and tag also the coordinate systems at the different times, e.g. the camera frames $\{cam_1\}$ and $\{cam_2\}$. The image plane defines new coordinate systems, $\{img_1\}$ and $\{img_2\}$, and the BEV defines another ones, $\{bev_1\}$ and $\{bev_2\}$. See Fig.1.

The projection matrix, P relating $\{cam_i\}$ and $\{img_i\}$ is given by the camera manufacturer or by a standard calibration procedure [10]. In this section we are mainly concerned with obtaining the homography, H relating the image plane with the BEV.

The BEV dewarping, H is defined by back-projecting to the ground plane four image points (appendix A details the back-projection matrix, P^*). The four image points are chosen so to comprehend most of the field of view imaging the ground plane. The region close to the horizon line is discarded due to poor resolution. Scaling is chosen such that it preserves the largest resolution available, i.e. no image-data loss due to sub sampling.

Is interesting to note that the knowledge of the 3D camera-motion, ${}^2T_1$ directly gives the BEV 2D rigid transformation, ${}^2H_1$ (see Fig.1):

$$ {}^2H_1 = H.P.{}^2T_1.P^*.H^{-1} \tag{1} $$

The inverse transformation, i.e. obtaining ${}^2T_1$ from ${}^2H_1$, is also possible since the motion is constrained to the ground plane: a 2D frame is transformed using ${}^2H_1$, and the missing dimension can be recoved e.g. by the relationship of the cross products of the vectors of the frame. In other words, estimating the camera motion in the camera frame is equivalent to estimating motion in the BEV frame.

4 Mosaic Construction

The input data for mosaic creation consists of BEV images, I_t and I_{t+1}, and clouds of ground-points projected in the BEV coordinate system, $\{[u\ v]^T_{t,i}\}$ and $\{[u\ v]^T_{t+1,i}\}$. In this frame, the camera motion is a 2D rigid translation, ${}^2H_1$, which can be represented by three parameters $\mu = [\delta u\ \delta v\ \delta\theta]$. We want to find μ such that the clouds of points match as close as possible:

$$\mu^* = \arg_\mu \min \sum_i \left\| [u\ v]^T_{t+1,j} - Rot(\delta\theta).[u\ v]^T_{t,i} - [\delta u\ \delta v]^T \right\|^2 \tag{2}$$

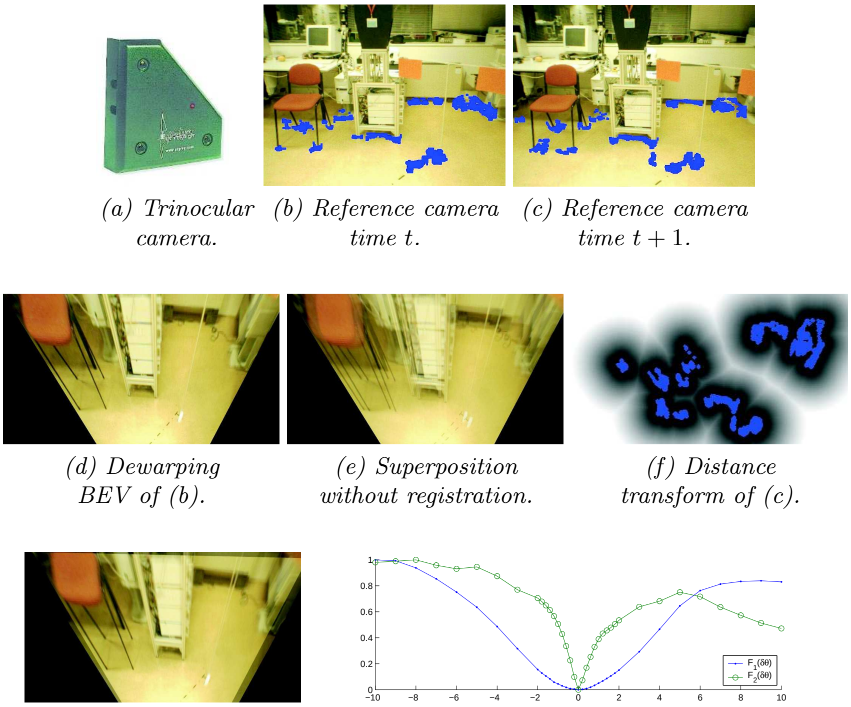
The correspondence between points of the clouds, i.e. finding the index j matching i, is based on the nearest neighbor rule, as with ICP. However in our case the matching is implemented using a distance transform. Using a distance transform allows matching 2D shapes as a 2D lookup-table reading of nearest neighbors and distances to them, instead of a combinatorial search between the clouds of points [9]. In order to smooth the cost functional and deal with the small shape differences (e.g. locally regular patterns generated by dense stereo reconstruction) we read interpolated-distance-values from the distance transform, and in order to deal with large differences (e.g. clouds leaving the field of view) we place a saturation on the distance transform (constant distances imply no influence in the optimization process).

Given the first estimation of the 2D motion and the knowledge of ground points, we can now fine tune the registration using ground plane image data:

$$\mu^* = \arg_\mu \min \sum_i \left\| I_{t+1}(Rot(\delta\theta).[u\ v]^T_{t,i} + [\delta u\ \delta v]^T) - I_t([u\ v]^T_{t,i}) \right\|^2 \tag{3}$$

Despite the fine tuning nature of this process, there is still possible to have regions of one image that got out of the field of view in the next image. The non-matched pixels get comparison values given by the closest matchings in an initial stage. These values are updated in the optimization process only if true matchings become possible, i.e. a new hypothetical 2D rigid motion between BEV images can bring to visibility unmatched points. This allows further smoothing the optimization process for points near the border of the field of view.

Finally, given the 2D rigid motion, the mosaic composition is just an accumulation of images. A growing 2D image buffer is defined such as to hold image points of the ground plane along the robot traveled path.



(a) Trinocular camera. *(b) Reference camera time t.* *(c) Reference camera time $t+1$.*

(d) Dewarping BEV of (b). *(e) Superposition without registration.* *(f) Distance transform of (c).*

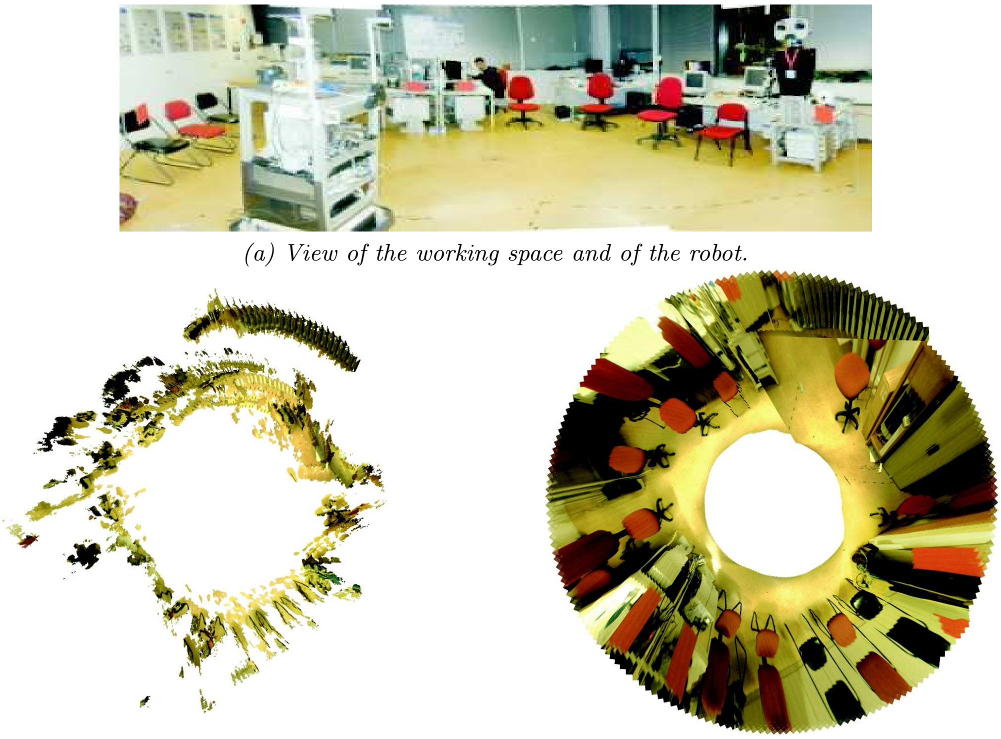
(g) Superposition after registration. *(h) Cost functionals vs perturbation $\delta\theta$ (costs normalized to $[0,1]$, $\delta\theta$ in $[-10^0, 10^0]$).*

Fig. 2. BEV dewarping and registration. (a) Stereo camera. (b) and (c) show reconstructed ground-points (blue-points) in the reference camera of the stereo setup. (d) BEV dewarping of (b). (e) superposition of BEVs without registration (notice the blur). (f) distance transform of the ground points seen in (c). (g) correct superposition of all ground points after registration. (h) comparison of the cost functionals by perturbing $\delta\theta$ about the minimizing point: registration using Eq.2 has a larger convergence region (dots plot) but the image-based registration, Eq.3 is more precise (circles plot).

5 Experimental Results

In our experiments we use the Point Grey's Digiclops trinocular camera (Fig.2a). This stereo camera encompasses three 640×480 color cameras arranged in an L-shape form (top-left, bottom-left and bottom-right), with 10 cm baselines. The robot carrying the camera follows a circular path with a 9.4 meters perimeter. 215 stereo images are acquired along the path.

Figure 2 illustrates the dewarping to BEV images and the registration of the dewarped images. The BEV images are 1568×855. One meter measured in the ground plane is equivalent to 318 pixels the BEV (this calibration information derives directly from the stereo-camera calibration). The registration is illustrated by super-imposing consecutive BEV images after applying to the first image the estimated 2D rigid motion. Notice in particular in Fig.2c the



(a) View of the working space and of the robot.

(b) Ground points used for registration (landmarks)

(c) Mosaic with all imaging data superimposed

Fig. 3. View of the working area (a), mosaic of the ground points chosen as landmarks while registering the sequence of BEV images (b) and a mosaic with all the visual information superimposed (c).

significant shape differences of the clouds of points as compared to Fig.2b, and in Fig.2g the graceful degradation of the superposition for points progressively more distant to the ground plane. Fig.2f shows the distance transform used for matching ground-points. The matching is performed repeatedly in Eq.2 in order to obtain the optimal registration shown in Fig.2g. The existence of local-clusters of points, instead of isolated points, motivates a wider-convergence but less precise registration which can be improved resorting to image data (Eq.3) as shown in figure Fig.2h.

The mosaicking of BEVs shows clearly the precision of the registration process. In particular shows that the image-based registration improves significantly the 2D motion estimation. After one complete circle described by the robot, the 2D ICP registration gives about 2.7 meters localization error (28% error over path length) which is improved to about 0.23 meters (2.3% error over path length) when using image-based registration. This allows obtaining a mosaic closing almost perfectly a circular path (see Fig.3c). Notice that only the ground points are registered in the mosaic, and thus all other points should exhibit artifacts due to parallax.

6 Conclusions

In this paper we proposed a method for creating mosaics of cluttered ground planes. Current stereo cameras provide 3D information and allow selecting ground points reliably. 3D data has been shown to be convenient as it allows selecting ground points, that can be registered and then used for building mosaics of cluttered ground planes.

The input of the mosaicking process consists mainly of many points forming local sparse clouds. This implied using robust registration methods designed for clouds instead of well separated features. We proposed using computer vision techniques such as distance transform to compare shapes and image correlation for fine tuning the registration. Results shown that the distance transform copes well with the sparse nature of the clouds. The saturation of the distance transform is useful for coping with the outliers. 3D reconstructed clouds were found to be useful for an initial registration, but fine tuning required resorting to the original image data.

Future work - The proposed mosaicking method will be used for benchmarking other registration methods based on the same input data (monocular or stereo vision). Combining reconstructed 3D information, accounting for variable local-densities of features and including color information, guaranteeing at the same time good convergence properties, is still a research topic within ICP registration.

As noted in the introduction, we plan to use the mosaics for navigation. The pairwise registration of the 2D ground features, acquired at consecutive time stamps, still suffers the error accumulation problem typical of odometry. However, from the point of view of keeping the robot localised, the image pairwise registration is enough, as the robot can always navigate on the mosaic and roll-back to its starting location by local registration over the mosaic.

A Back-projection to the ground plane

Consider the projection equation in homogeneous coordinates $m \approx PM = [A\ b]M$, where $M = [x\ y\ z\ 1]^T$ is a generic 3D point, $m = [u\ v\ 1]^T$ is the image point projection of M and the sign $\approx$ denotes equality up to a scale factor [11]. We have that the camera projection center is $C = -A^{-1}b$ and the 3D direction associated to m (point at infinity) is $D = A^{-1}m$. Thus the back-projection comes $M = [C; 1] + \alpha[D; 0] = [A^{-1}(\alpha m - b); 1]$ where α is a scaling factor setting the distance from the 3D point to the camera center and ";" denotes vertical stacking of vectors [12].

Representing the ground plane by a normal vector, n and a distance to the coordinate system origin, d, the factor α in the back-projection equation can be computed by enforcing $M_{1:3}^T.n = d$. The back-projection equation can now be arranged to a single 4×3 matrix, P^* converting an image point m to a 3D point M on the ground plane:

$$M \approx P^* m = \begin{bmatrix} (d + b^T A^{-T} n) A^{-1} & -A^{-1} b \\ n^T A^{-1} & 0 \end{bmatrix} \begin{bmatrix} I_3 \\ 0\,0\,1 \end{bmatrix} m. \quad (4)$$

Acknowledgments This work has been partially supported by the FCT Programa Operacional Sociedade de Informação (POSI) in the frame of QCA III, and by the projects Gestinteract POSC/EEA-SRI/61911 /2004 and URUS FP6-EU-IST-045062. The second and third authors were partially supported by the AECI/CYTED Spanish Program under the RIBERO project (Red Iberoamericana de Robtica), and by the ESPOL under the VLIR project, Component 8.

References

1. Davison, A.: Real-time simultaneous localisation and mapping with a single camera. In: IEEE Int. Conf. on Computer Vision. (2003) 1403 – 1410 vol2
2. Karlsson, N., Bernardo, E.D., Ostrowski, J., Goncalves, L., Pirjanian, P., Munich, M.: The vslam algorithm for robust localization and mapping. In: Proc. IEEE Int. Conf. on Robotics and Automation, Barcelona, Spain (2005) 24 – 29
3. Se, S., Lowe, D., Little, J.: Vision-based global localization and mapping for mobile robots. IEEE Trans. on Robotics **21**(3) (2005) 364–375
4. Besl, P.J., McKay, N.D.: A method for registration of 3-d shapes. IEEE Trans. on Pattern Analysis and Mach. Intel. **14**(2) (1992) 239–256
5. Fisher, R.: The iterative closest point algorithm, in cvonline: On-line compendium of computer vision. `http://homepages.inf.ed.ac.uk/rbf/CVonline/LOCAL_COPIES/FISHER/ICP/cvoicp.htm` (2006)
6. Rusinkiewicz, S., Levoy, M.: Efficient variants of the icp algorithm. In: Int. Conf. 3-D Digital Imaging and Modeling. (2001) 145–152
7. Chetverikov, D., Svirko, D., Stepanov, D., Krsek, P.: The trimmed iterative closest point algorithm. In: Int. Conf. on Pattern Recognition. (2002) 545–548vol.3
8. Biber, P., Fleck, S., Strasser, W.: A probabilistic framework for robust and accurate matching of point clouds. In: 26th Pattern Recognition Symposium. (2004)
9. Gavrila, D., Philomin, V.: Real-time object detection for smart vehicles. In: IEEE, Int. Conf. on Computer Vision (ICCV). (1999) 87–93
10. Bouguet, J.: Camera calibration toolbox for matlab. `http://www.vision.caltech.edu/bouguetj/calib_doc/` (2006)
11. Faugeras, O.: Three-Dimensional Computer Vision - A Geometric Viewpoint. MIT Press (1993)
12. Hartley, R., Zisserman, A.: Multiple View Geometry in Computer Vision. Cambridge University Press (2000)