To appear in *OCEANS '98 Conference*, 1998

Automatic Mosaic Creation of the Ocean Floor

Nuno Gracias and José Santos-Victor *
Instituto Superior Técnico & Instituto de Sistemas e Robótica
Av. Rovisco Pais,
1096 Lisboa Codex, Portugal

Abstract

In this paper, we describe the automatic creation of video mosaics of the ocean floor, which deals with the problem of image motion estimation in a robust and automatic way.

The motion estimation presented in this work is based on a initial matching of corresponding areas over pairs of images. As the error prone nature of the matching process is a commonly overlooked problem, this paper makes use of robust matching techniques, which can cope with an high percentage of wrong matches.

In our approach, several motion models are established under the projective geometry framework, allowing the creation of high quality mosaics where no assumptions are made on the camera motion. This is an improvement over traditional approaches for underwater mosaicing, usually relying on the camera to be facing the sea floor, so that the image plane is approximately parallel to the floor plane.

Extensive tests were run on underwater image sequences, testifying the good performance of the implemented matching and registration methods. Even with notorious violations of the underlying assumption of static planar scenes, the algorithm can still find the motion parameters as to create mosaics with bearably noticeable misalignments to the human eye.

I. Introduction

In the past few years we have witnessed a significant research effort to increase the autonomy of underwater vehicles. One of the key problems is the need of advanced sensor technologies that provide a better perception of the environment. Among these technologies, computer vision is increasingly being used as a mean for generating suitable representations of the underwater medium, both for aiding a human operator or to be integrated in the vehicle navigation system. In this context, video mosaicing constitutes an important tool for ocean floor exploration, since it can be used in operations such as site exploration, navigation and wreckage visualization. Furthermore, due to the underwater limited visual range, registration of close range images is often the only solution for obtaining large visual areas of the floor.

Traditional approaches for underwater mosaicing rely on the camera to be facing the sea floor, so that the image plane is parallel to the floor plane. Moreover it is usually assumed that image rotation and zoom is small. In our approach, several motion models are established under the projective geometry framework. These models range from simple image translation to the most general projective planar transformation that accounts for the registration of any view of a planar scene. Therefore, this approach allows the creation of high quality mosaics where no assumptions are made on the camera motion.

One of the main problems in motion analysis lies on the difficulty of the matching process between corresponding image areas. Contributing factors to this difficulty include the lack of image texture, object occlusions and acquisition noise, which are frequent in underwater imaging. As the error prone nature of the matching process is a commonly overlooked problem, the work here presented makes use of robust matching techniques, which cope with up an high percentage of wrong matches. As it will be shown here, the association of robust methods and geometrical model based estimation, allows the creation of video mosaics even in the presence of moving objects and non-planar scenes.

Research on automatic mosaic creation for underwater applications has been conducted in the last few years. In [7] a setup is proposed for creating mosaics by taking images at locations whose coordinates are known with high precision. Image merging can thus be performed without image analysis, because the frame-to-frame motion parameters can be computed directly

*Email: {ngracias,jasv}@isr.ist.utl.pt. The work described in this paper has been supported by the Portuguese Foundation for the Science and Technology PRAXIS XXI BD/13772/97, INFANTE Proj. PRAXIS 2/2.1/TPAR/2042/95, and NARVAL Esprit-LTR Proj. 30185

from the camera positions. Marks *et al.* [8] have developed a system for ocean floor mosaic creation in real-time. In their work, a four-parameter semi-rigid motion model is used, and small rotation and zooming on the image frames is assumed. This allows fast processing algorithms, but restricts the scope of applications to the case of a images taken by a camera whose retinal plane is closely parallel to the ocean floor. A common difficulty in underwater mosaicing arises from the presence of 3-D occlusions caused by seabed irregularities. Strategies for dealing with such occlusions are discussed by Tiwari in [12].

Although not dealt with in this paper, an important issue in mosaic creation is the propagation of registration errors over a sequence of images. This problem is tackled and discussed by Fleischer *et al.* in [2, 3] for the case of closed-loop image chains were the effects of the accumulation of small registration errors become apparent.

This paper is organized as follows. Section II. describes the framework under which the geometrical models for the mosaic formation are obtained, and the robust methods used for image motion estimation. In section III., the creation of video mosaics is presented, as being accomplished in two separate stages: registration and rendering. Finally, section IV. presents some results obtained from underwater footage and draws the conclusions on the performance and applicability of the method.

II. Approach

A. Image Motion Model

In this section we will assume the reader to be familiar with the basic concepts and properties of projective geometry. For an in-depth explanation on the subject, refer to [1].

The most commonly used camera model in computer vision is the pinhole model under which the camera performs a linear projective mapping from the projective space $\mathbb{P}^3$ to the projective plane $\mathbb{P}^2$. This mapping can be concisely written as $\widetilde{\mathbf{m}} \doteq P\widetilde{\mathbf{M}}$, where $\widetilde{\mathbf{M}}$ is a 3-D point location expressed in homogeneous coordinates, $\widetilde{\mathbf{m}}$ is its projection in the retinal plane, P is the (3×4) camera projection matrix and the $\doteq$ symbol denotes the equality up to a scale factor.

As we are interested in registering scenes with primarily planar content, we will now focus on 2-D projective transformations whose importance is emphasised by the fact that they can be used as models for image motion with an enormously vast field of application in Computer Vision. It can be easily shown[10, 4] that two different views of the same planar scene in 3-D space are related by a collineation in $\mathbb{P}^2$, represented by a (3×3) matrix[1] defined up to scale and establishing a one-to-one relation between corresponding points over two images. Thus, for a pair of image points of the same 3-D point of a planar scene with homogeneous coordinates $\widetilde{\mathbf{u}}$ and $\widetilde{\mathbf{u}'}$, the collineation T_{2D} relating $\widetilde{\mathbf{u}}_i$ and $\widetilde{\mathbf{u}'_i}$ will impose $\widetilde{\mathbf{u}'} \doteq T_{2D}\widetilde{\mathbf{u}}$.

The computation of a planar collineation requires at least four pairs of corresponding points matched over two images. If more than four correspondences are available, then a least-square estimation can be accomplished. Let T_{2D} be the collineation relating two image planes from which we have a set of n correspondences such that $\widetilde{\mathbf{u}'_i} \doteq T_{2D}\widetilde{\mathbf{u}}_i$, for $i = 1, \ldots, n$. For each pair we will have two linear constraints on the elements of T_{2D}. An homogeneous system of equations can thus be assembled in the form $H.\mathbf{t}_l = 0$, where $\mathbf{t}_l$ is the column vector containing the elements of T_{2D} in a row-wise fashion, and H is a $(2n \times 9)$ matrix. The system can be solved by the means of the Singular Value Decomposition, after imposing an additional constraint of unit norm for $\mathbf{t}_l$, *i.e.*, $||\mathbf{t}_l|| = 1$.

As it is defined up to scale, the most general collineation in $\mathbb{P}^2$ has eight independent parameters. If additional information is available on the camera setup, such has camera motion constraints, then the coordinate transformation $\widetilde{\mathbf{u}'_i} \doteq T_{2D}\widetilde{\mathbf{u}}_i$ might not need the eight independent parameters of the general case to accurately describe the image motion. As an example we can point out the case where the camera is just panning, thus inducing a simple sideways image translation. If we know beforehand which is the simplest model that can explain the data equally well, then there will be no reason for using the most general. Table 1 illustrates some restricted models.

B. Robust Motion Estimation

Model estimation, in the sense of model fitting to noisy data, is employed in computer vision on a large variety of tasks. The most commonly used method is the least-squares mainly due to the ease of implementation and fast computation. The least-squares is optimal when the underlying error distribution of the data is Gaussian[9]. However, in many applications the data are not only noisy, but it also contains *outliers*, *i.e.* data in gross disagreement with the assumed model. Under a least-square framework, outliers can distort the fitting process to the point of making the fitted parameter arbitrary[13].

[1] A collineation in $\mathbb{P}^2$ is also commonly refered to as a planar transformation.

Image Model	Matrix form	p	Domain
Translation and zoom	$T_{2D} = \begin{bmatrix} t_1 & 0 & t_2 \\ 0 & t_1 & t_3 \\ 0 & 0 & t_4 \end{bmatrix}$	3	Image plane is parallel to the planar scene. No rotation but with variable focal length or distance to the scene.
"Semi-Rigid"	$T_{2D} = \begin{bmatrix} t_1 & t_2 & t_3 \\ -t_2 & t_1 & t_4 \\ 0 & 0 & t_5 \end{bmatrix}$	4	Same as above but with rotation and scaling along the image axes.
Affine Transformation	$T_{2D} = \begin{bmatrix} t_1 & t_2 & t_3 \\ t_4 & t_5 & t_6 \\ 0 & 0 & t_7 \end{bmatrix}$	6	Distant scene subtending a small field of view.

Table 1: Some of the possible motion models used for image merging, ordered by the number of free parameters p.

A widely used non-linear minimization method for dealing with outliers is the *least-median-of-squares* [11] (LMedS). The parameters for the planar transformation are estimated by solving

$$\min \operatorname*{med}_i \left(\mathbf{h}_i^T . \mathbf{t}_l\right)^2$$

where $\mathbf{h}_i^T$ is the i^{th} row of the observations matrix H. As pointed out in [9], this minimization problem cannot be reduced to a least-squares based solution. The minimization on the space of all possible solutions is usually impracticable. Therefore it is common practice to use a Monte Carlo technique and to analyze only a randomly sampled subsets of points.

In the work presented on this paper, we have used a two-step variant of LMedS, referred to as MEDSERE[5]. It exhibits a similar breakdown point but requires less random sampling in order to achieve the same degree of outlier rejection[4]. The MEDSERE algorithm comprises two phases of random sampling LMedS. After the first phase, the data set is reduced by selecting the best data points in the sense of the chosen cost function. Next, the reduced data undergoes another random sampling LMedS phase.

III. Mosaic Creation

The creation of video mosaics is accomplished in two stages: registration and rendering. On the registration stage the image motion is estimated, then the individual frames are fitted to a global model of the sequence. The rendering stage deals with the creation of a single mosaic, by applying a temporal operator over the registered and aligned images.

The frame-to-frame motion estimation procedure allows the construction of mosaics by the analysis of consecutive pairs of frames. In the global registration step, the frame-to-mosaic transformation for the last frame is computed by sequentially cascading all the previous inter-frame transformations.

A. Registration

The work presented on this paper evolves around the analysis of point projections and their correspondence between image frames. In order to improve the correspondence finding, a number of points are selected corresponding to image corners or highly textured patches, using a simplified version of the well-known corner detector proposed by Harris and Stephens[6].

For each image I_k, a set of features is extracted and matched directly on the following image I_{k+1}, using a correlation-based matching procedure and resulting in two lists of coordinates of corresponding points. Due to the error prone nature of the matching process, it is likely that a number of point correspondences will not relate to the same 3-D point.

The MEDSERE algorithm is used for the estimation of $T_{k,k+1}$ which relates the coordinate frames of I_k and I_{k+1}. Let ${}^{(k)}\mathbf{u}_i$ be the location of the i^{th} feature extracted from image I_k, and matched with ${}^{(k+1)}\mathbf{u}$ on image I_{k+1}. The criterion to be minimized is the median of sum of the square distances,

$$\begin{aligned} \operatorname*{med}_i (& d^2({}^{(k)}\mathbf{u}_i, T_{k,k+1}\ {}^{(k+1)}\mathbf{u}_i) \\ & + d^2({}^{(k+1)}\mathbf{u}_i, T_{k,k+1}^{-1}\ {}^{(k)}\mathbf{u})) \end{aligned} \tag{1}$$

where $d(\cdot, \cdot)$ stands for the point-to-point Euclidean distance.

After estimating the frame-to-frame motion parameters, these collineations are cascaded to form a global model. The global model takes the form of a global registration, where all frames are mapped into a common, arbitrarily chosen, reference frame. Let $T_{Ref,1}$ be the transformation matrix relating the frames of the chosen reference and the first image frame. The global registration is defined by the set of transformation matrices $\{T_{Ref,k} : k = 1 \ldots N\}$, where for $2 \leq k \leq N$,

$$T_{Ref,k} = T_{Ref,1} \prod_{i=1}^{k-1} T_{i,i+1}$$

B. Rendering

After global registration, the following step consists in merging the images. On ovelapping regions there are more multiple contributions for a single point on the output image, and some method has to be established in order to determine the unique intensity value that will be used. The contributions for the same output point can be thought of as lying on a line which is parallel to the time axis, in a space-time continuum of the globally aligned images. Therefore, the referred method operates on the time domain, thus called a *temporal operator*. Some of the commonly used methods are the use-first, use-last, mean and median. The first two use only a single value from the contributions vector, respectively the first and the last entries of the timely ordered vector. The mean operator takes the average over all the point contributions, and is effective in removing temporal noise inherent in video. Finally, the median operator also removes temporal noise but is particularly effective in removing transient data, such as fast moving objects whose intensity patterns are stationary for less than half the frames. It is therefore adequate for underwater sequences of the seabed, where moving fish or algae are captured.

IV. Results and Discussion

The ocean floor mosaics presented in this paper were created from a number of video sequences where no information was used, other than the images themselves and the most suitable motion model.

An example of a sea bed mosaic is given in Figure 1. It was composed with 101 frames, registered under the semi-rigid model and rendered with the median operator. The original sequence was obtained by a manually controlled underwater vehicle, and depicts a man-made construction. This scene is not planar nor static. The camera is moving along a fracture inside which some rocks can be seen. In the fracture there are noticeable depth variations as opposed to the almost planar surrounding sea bed. Even so, the sea bed is mostly covered with algae and weeds, which provide good features for the matching process, but violate the underlying planar scene assumption. Another assumption violation is due to some moving fish. Figure 2 shows two sub-mosaics in which the motion of the fish can be clearly noticed. Although constructed from the same sequence, these sub-mosaics were rendered using the use-last temporal operator.

Figure 3 presents two views of a mosaic from a sequence of images captured by a surface-driven ROV, on a pipe inspection task. In this example the perspective distortion effects are noticeable, since the image plane of the camera is distinctly not parallel to the sea floor. The most suitable motion model is, therefore, the full planar transformation. The left image was created using the first frame of the sequence as the reference frame. For the right image, a reference frame was chosen as to make the contour lines of the pipe approximately parallel, yielding a top view of the floor.

The presented mosaics illustrate the good performance of the implemented matching and registration methods. Even with notorious violations of the assumed model, the algorithm can still find the motion parameters as to create a mosaic with small misalignments to the human eye.

Acknowledgments: The authors would like to thank Drafinsub, s.r.l. and Gabriel Codina for providing the original image sequences.

References

[1] O. Faugeras. *Three Dimensional Computer Vision.* MIT Press, 1993.

[2] S. Fleischer, R. Marks, S. Rock, and M. Lee. Improved real-time video mosaicking of the ocean floor. In *Proc. of the IEEE OCEANS 95 Conference*, pages 1935–1944, California, USA, October 1995.

[3] S. Fleischer, H. Wang, S. Rock, and M. Lee. Video mosaicking along arbitrary vehicle paths. In *Proc. of the 1996 Symposium on Autonomous Underwater Vehicle Technology*, pages 293–299, California, USA, June 1996.

[4] N. Gracias. Application of robust estimation to computer vision: Video mosaics and 3–D reconstruction. Master's thesis, http://www.isr.ist.utl.pt/labs/vislab/thesis, Lisbon, Portugal, April 1998.

[5] N. Gracias and J. Santos-Victor. Robust estimation of the fundamental matrix and stereo correspondences. In *Proc. of the International Sym-*

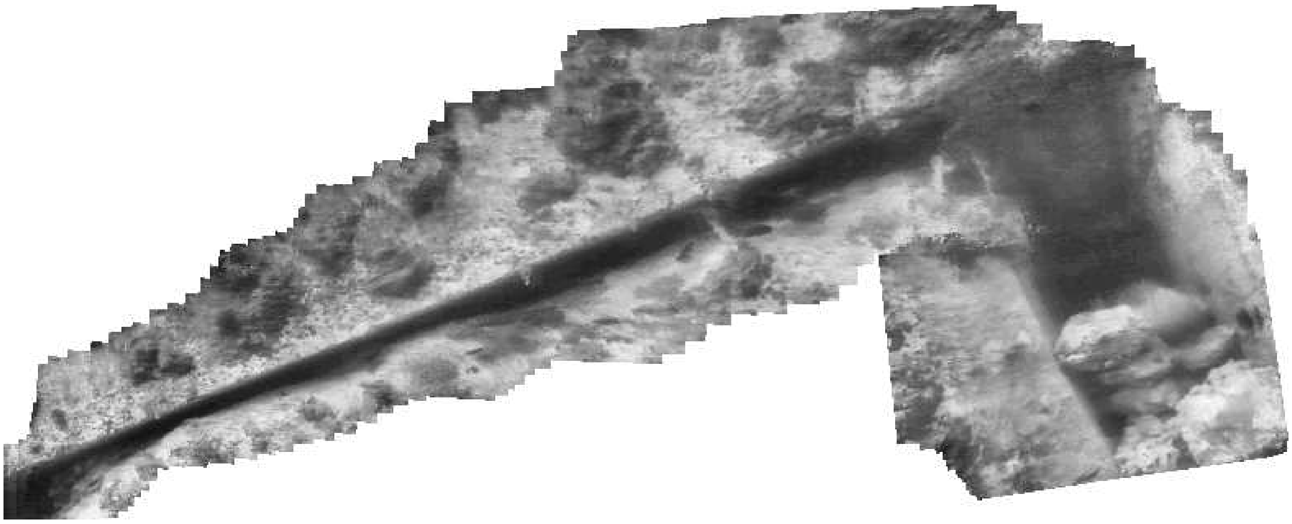

Figure 1: Sea bed mosaic example. The images were registered using the semi-rigid motion model and rendered using the median operator.

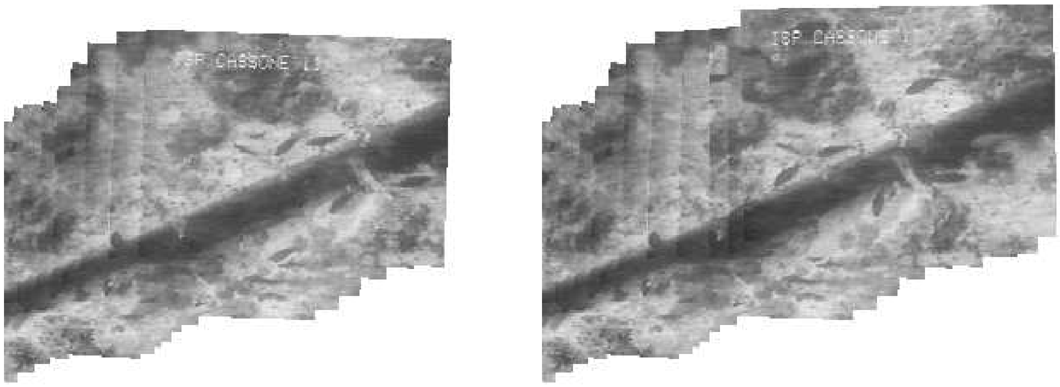

Figure 2: Example of mosaic creation where the static scene assumption is violated by the presence of moving fish.

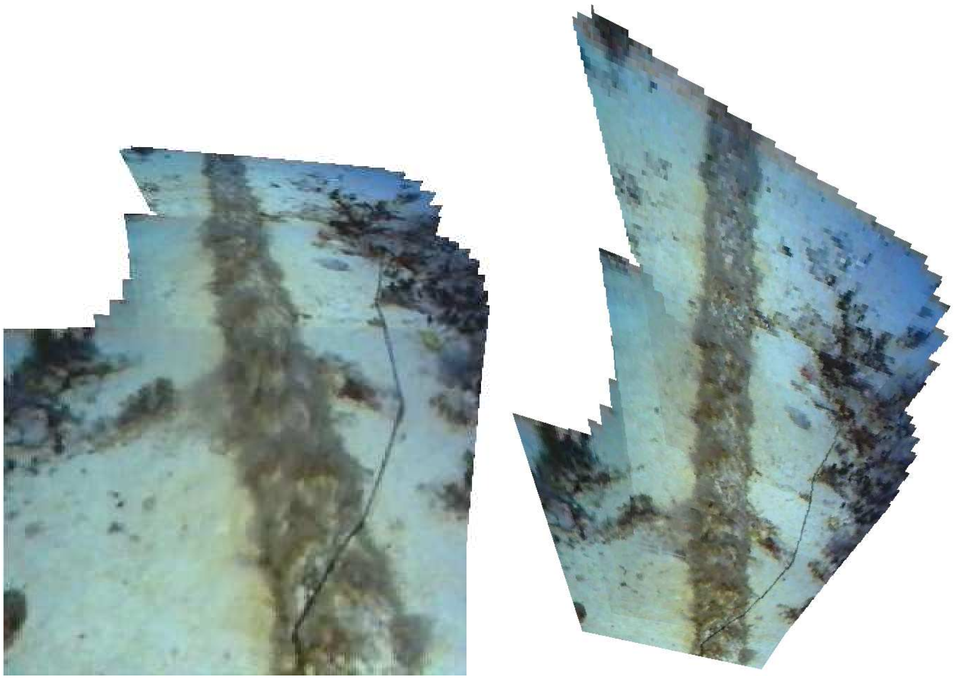

Figure 3: Underwater pipe mosaic example. For the image registration, the full planar tranformation model was used. The images were registered with the use-last operator (left). A useful reference frame can be chosen in order to have a better perception of the sea floor (right).

posium on Intelligent Robotic Systems, Stockholm, Sweden, July 1997.

[6] C. Harris. Determination of ego-motion from matched points. In *Proceedings Alvey Conference*, Cambridge, UK, 1987.

[7] R. Haywood. Acquisition of a micro scale photographic survey using an autonomous submersible. In *Proc. of the OCEANS 86 Conference*, New York NY, USA, 1986.

[8] R. Marks, S. Rock, and M. Lee. Real-Time video mosaicking of the ocean floor. *IEEE Journal of Oceanic Engineering*, 20(3):229–241, July 1995.

[9] P. Meer, D. Mintz, A. Rosenfeld, and D. Kim. Robust regression methods for computer vision: a review. *Int. Journal of Computer Vision*, 6(1):59–70, 1991.

[10] R. Mohr and B. Triggs. Projective geometry for image analysis. Tutorial given at International Symposium of Photogrammetry and Remote Sensing, Vienna, Austria, July 1996.

[11] P. Rousseeuw and A. Leroy. *Robust Regression and Outlier Detection*. John Wiley & Sons, 1987.

[12] S. Tiwary. Mosaicking of the ocean floor in the presence of three-dimensional occlusions in visual and side-scan sonar images. In *Proc. of the 1996 Symposium on Autonomous Underwater Vehicle Technology*, pages 308–314, California, USA, June 1996.

[13] P. Torr and D. Murray. The development and comparision of robust methods for estimating the fundamental matrix. *International Journal of Computer Vision*, 24(3):271–300, September/October 1997.