Calibration of an Optical-Acoustic Sensor*

Andrea Fusiello and Vittorio Murino
Dipartimento Scientifico e Tecnologico, University of Verona, Italy
{*fusiello,murino*}*@sci.univr.it*

Abstract. In this paper an application is described, where image registration to a model is used in conjunction with registration of acoustic (3-D) data to the same model to perform calibration of an optical/acoustic rig. This allows to register the video image with the depth map acquired by the acoustic device, thereby obtaining a single multisensorial image. Experimental results on both synthetic and real images are reported.

Key words: Multisensor model of image formation, 3-D imaging calibration, acoustical images

1. Introduction

This paper presents a novel technique for calibrating a compound sensor consisting of a video camera and an acoustic three-dimensional (3-D) camera, mounted side by side on a rig. We use the scene itself as a calibration object: both acoustic and optical images are registered to a known model of the observed object, thereby obtaining the relative pose of optical and acoustic cameras. The final result is the registration of the low-resolution depth map with the optical (high resolution) image. The integrated data can be thought as a *multisensorial image*, which can be displayed as is or used to synthesize a 3-D virtual environment [11].

This work is motivated by underwater applications where the integration of several different sensors is crucial for typical tasks like navigation and inspection. Actually, in underwater environments, an optical camera is often insufficient and an acoustic sensor can make up for its shortcomings.

To the best of our knowledge, no previous work addresses this topic. In fact, previous works in (underwater) acoustic 3-D data reconstruction are present in literature, and optical calibration is a well known technique in computer vision. However, the integration of acoustic and optical sensors via geometric calibration were never proposed before. Our work differs from those addressing the joint processing of 3-D and gray-level images [5, 9], since the latter operates on perfectly registered images having the same (lateral) resolution[1], whereas our method relies on very different sensors with different resolution,

*This work is supported by the European Commission under the BRITE-EURAM III project no. BE-2013 VENICE (Virtual Environment Interface by Sensor Integration for Inspection and Manipulation Control in Multifunctional Underwater Vehicles)

[1]The resolution of an imaging device is the solid angle that projects to an image elements. In a range

in which a point-to-point relation is not provided, but is indeed the outcome of our algorithm.

In Sec. 2 the optical and acoustic cameras are described. Sec. 3 outlines the idea of optical/acoustic calibration, and gives an overview of the whole method, which relies on 3-D alignment (Sec. 4) and 2-D alignment (Sec. 5). Experimental results (Sec. 6) on both synthetic and real images show the effectiveness of the proposed approach. More details on this work can be found in [12].

2. Camera models

2.1. Optical camera

A pinhole camera is modeled by its *optical center* C and its *retinal plane* (or *image plane*) $\mathcal{R}$. A 3-D point W is projected into an image point M given by the intersection of $\mathcal{R}$ with the line containing C and W. The line containing C and orthogonal to $\mathcal{R}$ is called the *optical axis* and its intersection with $\mathcal{R}$ is the *principal point*. The distance between C and $\mathcal{R}$ is the *focal distance*.

Let $\mathbf{w} = [x\ y\ z]^\top$ be the coordinates of W in the *model reference frame* and $\mathbf{m} = [u\ v]^\top$ the coordinates of M in the image plane (pixels). The mapping from 3-D coordinates to 2-D coordinates is the *perspective projection*, which is represented by a linear transformation in *homogeneous coordinates*. Let $\tilde{\mathbf{m}} = [u\ v\ 1]^\top$ and $\tilde{\mathbf{w}} = [x\ y\ z\ 1]^\top$ be the homogeneous coordinates of M and W respectively; then the perspective transformation is given by the 3×4 matrix $\tilde{\mathbf{P}}$:

$$\kappa\tilde{\mathbf{m}} = \tilde{\mathbf{P}}\tilde{\mathbf{w}}, \tag{1}$$

where κ is an arbitrary scale factor. The camera is therefore modeled by its *perspective projection matrix* $\tilde{\mathbf{P}}$, which can be decomposed, using the QR factorization, into

$$\tilde{\mathbf{P}} = \mathbf{A}[\mathbf{I}|\mathbf{0}]\mathbf{G}_o. \tag{2}$$

The matrix $\mathbf{A}$ depends on the *intrinsic parameters* only: focal length in pixel, aspect ratio, principal point and skew factor. The camera position and orientation (*extrinsic parameters*), are encoded by the 4×4 matrix $\mathbf{G}_o$ representing (in homogeneous coordinates) the rigid transformation that brings the camera reference frame onto the model reference frame.

Calibration consist in computing the intrinsic and extrinsic parameters of the camera. Given a sufficient number of correspondences between model reference points and image points, it is possible to solve the perspective projection equation (1) for the unknown parameters. In our experiments we used a suitable object to accurately identify reference points, and the calibration algorithm developed by L. Robert [8].

image we distinguish between lateral and range resolution. The latter is the resolution in the depth measuring.

2.2. Acoustic camera

The acoustic camera (called Echoscope [7]) is formed by a two-dimensional array of transducers sensible to signals backscattered from the scene previously insonified by a high-frequency acoustic pulse. The whole set of raw signals is then processed to estimate signals coming from fixed steering directions (called beamsignals) while attenuating those coming from other directions. Assuming that beamsignals represent the responses of a scene from a 2D set of (steering) directions, a 3-D point set can be extracted detecting the time instant (t^*) at which the maximum peak occurs in each beamsignal. Besides, the intensity of the maximum peak can be used to generate another image, registered with the former, representing the reliability of the associated 3-D measures. In other words, the higher the intensity, the safer the 3-D measure associated. Images are formed by 64×64 3-D points ordered such that adjacent points correspond to adjacent beam signals. Their coordinates are expressed in a 3-D reference frame attached to the sensor. The rigid transformation that brings this reference frame onto the model reference frame is by a 4×4 matrix $\mathbf{G}_a$.

3. Optical/Acoustic calibration

Given a rig composed by an optical and an acoustic camera, calibration consists in recovering the unknown rigid transformation that relates the two reference frames attached to the cameras. In this way, given an acoustic image, composed by a set of target points, each with a certain 3-D position, we can project it onto the optical image plane, obtaining a depth map with reference to the image plane. This can be resampled (e.g. by bilinear interpolation) on the pixels grid to obtain a depth map referred to the optical image.

In principle, to calibrate the cameras, a suitable object should be manufactured which gives raise to distinct features both in the acoustic image and optical image. This is very difficult to achieve, mainly because of the different resolution of the two devices. In our approach, we use the scene itself as a calibration object. Knowing the CAD model of the observed objects, we register both acoustic and optical data to the model, thereby obtaining the relative pose of optical and acoustic cameras. The typical object we deal with is an off-shore oil rig, consisting of connected cylindrical pipes.

Let $\mathbf{G}_o$ be the extrinsic parameters matrix of the optical camera:

$$\tilde{\mathbf{w}}_{\text{camera}} = \mathbf{G}_o \tilde{\mathbf{w}}, \tag{3}$$

and let $\mathbf{G}_a$ be the rigid transformation that brings the acoustic camera reference frame onto the model reference frame:

$$\tilde{\mathbf{w}}_{\text{sonar}} = \mathbf{G}_a \tilde{\mathbf{w}}. \tag{4}$$

By composing the two transformations we get: $\tilde{\mathbf{w}}_{\text{camera}} = \mathbf{G}_o \mathbf{G}_a^{-1} \tilde{\mathbf{w}}_{\text{sonar}}$. Hence, the perspective projection matrix that projects the 3-D points expressed in the acoustic

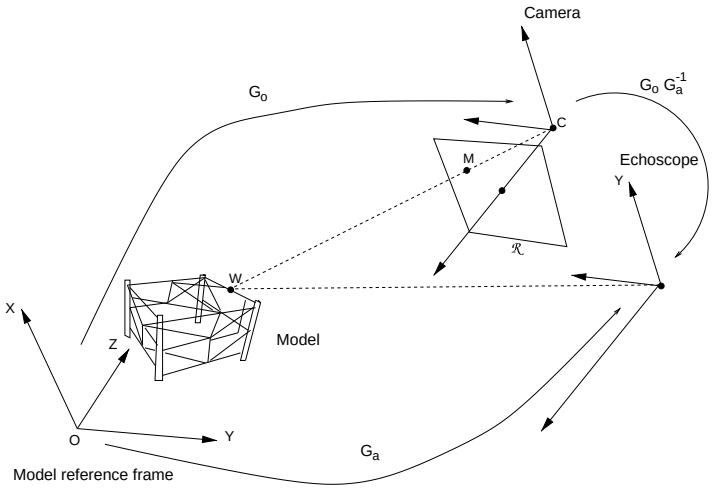

Fig. 1. Optical/acoustic calibration

camera reference frame onto the image plane of the optical camera is given by

$$\tilde{\mathbf{P}}_{\mathrm{oa}} = \mathbf{A}[\mathbf{I}|\mathbf{0}]\mathbf{G}_o\mathbf{G}_a^{-1}. \tag{5}$$

The aim of the optical/acoustic calibration is to obtain $\mathbf{G}_o$ and $\mathbf{G}_a$ by registering the optical image and the acoustic image to the CAD model. These are two distinct problems: (i) the 3-D - 3-D alignment, also known as the 3-D motion problem or absolute orientation problem [6]; (ii) the 3-D - 2-D alignment, also known as the exterior orientation problem [6].

4. 3-D Alignment

In our approach, acoustic data points lying on the surface of cylinders are matched to the underlying object surface model using an iterative least squares technique. Data points are expressed in the acoustic reference frame, whereas the model cylinders are placed in the model reference frame. The sought rigid transformation that links the two reference frames is given by $\mathbf{G}_a$.

4.1. The ICP algorithm

In their paper, Besl and McKay [3] introduce the Iterative Closest Point (ICP) algorithm, a general purpose method for solving the *3-D registration problem*.

Suppose that we are given two sets X and Y corresponding to a single shape, where Y is a set of 3-D points and X is a surface (of a cylinder, in our case). The correspondence

between Y and X is unknown. For each point $\mathbf{y}_i$ from the set Y, there exists at least one point on the surface of X which is closer to $\mathbf{y}_i$ than all other points in X. This is the closest point, $\mathbf{x}_i$. The basic idea behind the ICP algorithm is that, under certain conditions, the point correspondences provided by sets of closest points are a reasonable approximation to the true point correspondence. Besl and McKay [3] proved that if the process of finding closest point sets and then solving the *point set registration problem* [10] is iterated, the solution is guaranteed to converge to a local minimum. In our case, pre-alignment based on the inertial tensor [11] produced a fairly good initial pose, sufficient to achieve convergence to the correct solution. In cases where this fails, the human operator is in charge of providing a rough initial pose estimate by specifying at least three point correspondences.

5. Optical alignment

Optical alignment, that is solving for the camera's pose that best fit a model to some matching image features, is performed using Lowe's algorithm [1].

5.1. Lowe's algorithm

Let us begin by supposing that point correspondences are available and that the intrinsic camera parameters are known. Let $\mathbf{w}_1 \dots \mathbf{w}_N$ be N points of an object model expressed in the model reference frame and $\mathbf{m}_1 \dots \mathbf{m}_N$ be the image points, projections of the $\mathbf{w}_i$. The relation between an object point and an image point, is given by the perspective projection:

$$\kappa \mathbf{A}^{-1} \tilde{\mathbf{m}}_i = [\mathbf{R}|\mathbf{t}] \tilde{\mathbf{w}}_i. \tag{6}$$

derived from (2) by letting $\mathbf{G}_o = \begin{bmatrix} \mathbf{R} & \mathbf{t} \\ \mathbf{0} & 1 \end{bmatrix}$. Let $\tilde{\mathbf{p}}_i = [u_i, v_i, 1]^\top = \mathbf{A}^{-1} \tilde{\mathbf{m}}_i$ be the *normalized image coordinates*. To explicitly enforce orthogonality, $\mathbf{R}$ is parameterized with the three Eulerian angles ϕ, ψ, θ. Let's think of (6) as a mapping $\mathbf{F}_i : R^6 \rightarrow R^2$ from the six parameters space of the rigid transformations to the image coordinates (u_i, v_i). Then (6) is equivalent to

$$\mathbf{p}_i = \mathbf{F}_i(\mathbf{e}). \tag{7}$$

where $\mathbf{e} = [\mathbf{t}, \phi, \psi, \theta]^\top$. This non-linear equation can be solved with Newton's method. $\mathbf{F}_i$ is linear with respect to translation and scaling over the image plane, and approximately linear over a wide range of values of the rotational parameters. Hence, the method is likely to converge to the desired solution for a rather wide range of possible starting positions. In our case the initial pose for the optical camera is assumed to be the same of the acoustic one ($\mathbf{G}_a$), already computed. Given the small displacement between the two cameras, this is usually sufficient to ensure convergence. The method can be easily extended to cope with line correspondences[1].

The case of a smooth boundaries objects, as cylinders, is different. A rim generated by a sharp edge is stable on the object as long as the edge is visible, whereas a rim generated by a smooth surface changes continuously with the viewpoint. However, as noted by Lowe [1], treating the rim as a fixed line in space does not prevent the algorithm to converge, and does not affect the precision of the final alignment.

The matching between image features and model features is computed using an algorithm introduced by Scott and Longuet-Higgins [2] for associating features of two arbitrary patterns.

In our case, model features are segments belonging to the cylinder's *rim*[2], and image features are lines containing the *bounding contour* of the cylinder, extracted automatically from the image. Since the initial pose given by acoustic alignment is fairly close to the true one, this simple matching procedure is sufficient.

6. Experiments and Discussion

In this section experimental results are shown with both synthetic and real images.

Synthetic images were generated using Rayshade, with an option that allows to save the z-buffer, simulating an ideal acoustic 3-D image. The optical camera were displaced with respect to the acoustic one by a lateral translation of 300 mm. Image lines were extracted from the image with Canny filter followed by Hough transform (details in [11]). Figure 2 shows the result of the optical/acoustic alignment.

More interesting are the experiments with real images, of which we show here only one example in Figure 3. A video camera and an acoustic camera was mounted side by side on an underwater remote operated vehicle. The distance between their centers was approximately 300mm. Acoustic raw images are typically quite noisy. For this reason, a low-level processing phase is mandatory. We used the reliability image to discard 3-D points associated with a low intensity, followed by a size filter (more details in [11]) The 3-D registration with ICP performed well, converging to a RMS distance of about 70mm, on the average. As expected, the use of 3-D data provide a very precise object pose estimate. The optical alignment also converged, but the estimated position of the camera's optical center was very uncertain along the direction of the optical axis (the standard deviation was of the order of 600 mm). This points strongly to iterative estimation with Kalman filtering of the unknown but fixed displacement between the two cameras, as the vehicle moves.

Since the optical alignment seems to be the critical part of the process, we tested its sensitivity to noise, by perturbing the intrinsic parameters in the synthetic case. It turns out that the translation component of the displacement is very sensitive to intrinsic

[2]Given a viewpoint, the *rim* of an object is the set of points on the object's surface where the line containing the viewpoint (i.e., the optical ray) is tangent. The projection of the rim in the image is the *bounding contour* of the object.

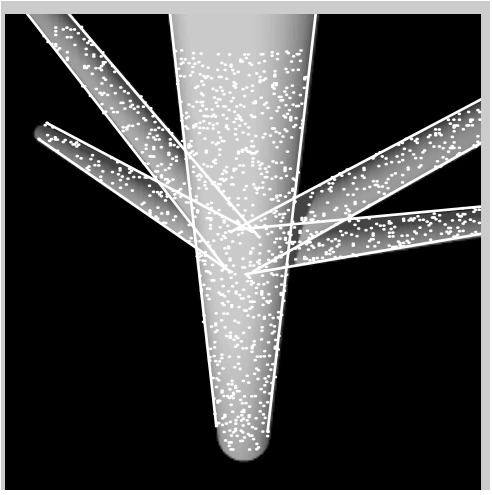

Fig. 2. Synthetic image, with the projected acoustic points (white dots) and the projected model (white lines) superimposed.

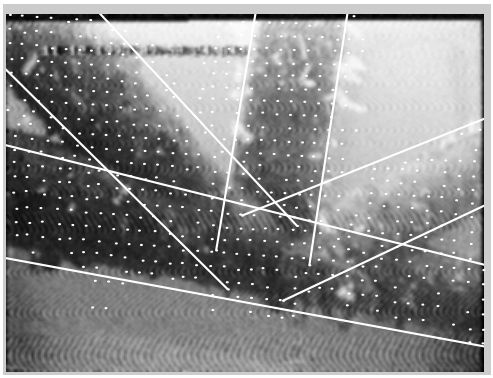

Fig. 3. Real image, with the projected acoustic points (white dots) and the projected model (white lines) superimposed. Some points falls outside the objects because of the low spatial resolution of the acoustic device.

parameters variation, especially along the direction of the optical axis. Even without noise, there was an error of 173 mm in the translation between camera centers. This is in agreement to [4], who noted that, while 2-D data is useful for estimating object motion in planes normal to a camera's optical axis, it is less sensitive to motions which deviate from these planes.

Acknowledgements

R. Hansen of Omnitech A/S is gratefully acknowledged for providing the Echoscope images. R. Giannitrapani developed the acoustic low-level processing modules. A. Verri kindly provided the calibration jig.

References

1991

[1] D.G. Lowe. Fitting parameterized three-dimensional models to images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13(5):441–450, May 1991.

[2] G. Scott and H. Longuet-Higgins. An algorithm for associating the features of two images. In *Proceedings of the Royal Society of London B*, volume 244, pages 21–26, 1991.

1992

[3] P. Besl and N. McKay. A method for registration of 3-D shapes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 14(2):239–256, February 1992.

[4] J. Wang and W. J. Wilson. 3D relative position and orientation estimation using Kalman filter for robot control. In *Proceedings of the IEEE International Conference on Robotics and Automation*, pages 2638–2645, Nice, France, 1992.

1995

[5] A. Jain, B. Gunsel, and E. Panayirci. Reconstruction and boundary estimation of range and intensity images using multiscale MRF representations. *CVGIP: Image Understanding*, 63(2):353–366, March 1995.

[6] R. Jain, R. Kasturi, and B.G. Schunk. *Machine Vision*. Computer Science Series. McGraw-Hill International Editions, 1995.

1996

[7] R. K. Hansen and P. A. Andersen. A 3-D underwater acoustic camera - properties and applications. In P.Tortoli and L.Masotti, editors, *Acoustical Imaging*, pages 607–611. Plenum Press, 1996.

[8] L. Robert. Camera calibration without feature extraction. *Computer Vision, Graphics, and Image Processing*, 63(2):314–325, March 1996.

1997

[9] C.O. Jaynes, A.R. Hanson, E.M. Riseman, and H. Schultz. Building reconstruction from optical and range images. In *CVPR97*, pages 380–386, 1997.

[10] A. Lorusso, D. W. Eggert, and R. B. Fisher. A comparison of four algorithms for estimating 3-D rigid transformations. *Machine Vision and Applications*, 9:272–290, 1997.

1999

[11] A. Fusiello, R. Giannitrapani, V. Isaia, and V. Murino. Virtual environment modeling by integrated optical and acoustic sensing. In *Second International Conference on 3-D Digital Imaging and Modeling (3DIM99)*, pages 437–446, Ottawa, Canada, 4-8 October 1999. IEEE Computer Society Press.

2000

[12] A. Fusiello, and V. Murino. Calibration of an Optical-Acoustic Sensor for Underwater Applications. Research Memo DST/VIPS/RM-01/00, University of Verona, 2000. Available at ftp://vips.sci.univr.it/pub/papers/RM-00-01.ps.gz