Proceedings of the 2002 IEEE/RSJ
Intl. Conference on Intelligent Robots and Systems
EPFL, Lausanne, Switzerland • October 2002

An Approach to Vision-Based Station Keeping for an Unmanned Underwater Vehicle

Xavier Cufí, Rafael Garcia and Pere Ridao

Computer Vision and Robotics Group
Institute of Informatics and Applications
University of Girona, E.P.S.
17071 Girona, Spain
e-mail: {xcuf,rafa,pere}@eia.udg.es

Abstract

This paper presents an automatic vision-based system for UUV station keeping. The vehicle is equipped with a down-looking camera, which provides images of the sea-floor. The station keeping system is based on a feature-based motion detection algorithm, which exploits standard correlation and explicit textural analysis to solve the correspondence problem. A visual map of the area surveyed by the vehicle is constructed to increase the flexibility of the system, allowing the vehicle to position itself when it has lost the reference image. The testing platform is the URIS underwater vehicle. Experimental results demonstrating the behavior of the system on a real environment are presented.

1. Introduction

Underwater vehicles are an important tool when we aim to inspect man-made underwater structures. Likewise, some of the repairing tasks can be performed by means of Remotely Operated Vehicles (ROVs) without endangering human lives. In this way, a pilot can teleoperate the vehicle from the surface, performing the desired task with little effort. However, maintaining the position of the vehicle within the working area may be a difficult task in the presence of underwater currents, even for experienced pilots. If we add to this factor the risk of inattention due to long survey missions performing this tedious task, we can see the necessity of endowing the vehicle with the capability of detecting motion and correcting its position to maintain station. This is accomplished by equipping the vehicle with a down-looking camera which acquires images of the sea floor. Our testing platform is the URIS underwater vehicle (Figure 1), an Autonomous Underwater Vehicle developed at the University of Girona.

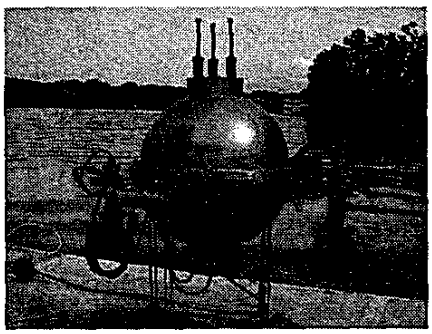

Figure 1: URIS Underwater Vehicle.

Unfortunately, underwater images are difficult to process due to the medium transmission characteristics [1]. These properties provoke a blurring of the elements of the image, limited range and need of artificial light which introduces new properties to the image, such as low contrast, non-uniform illumination, high clutter in the regions of interest and lack of distinct features. Negahdaripour *et al.* [2,3] proposed a station keeping method based on a "direct method" to compute the vehicle motion directly from spatio-temporal image derivatives. This approach requires the motion from one image to the next to be small, therefore a multiresolution scheme should be applied to the images when this assumption is violated. On the other hand, standard region-correlation techniques have been extensively used to search for correspondences between pairs of images [4], allowing the detection of motion. Stanford/MBARI researchers have proposed a correlation-based approach to estimate the motion of the vehicle relative to the sea floor to accomplish station keeping [5]. Although these approaches lead to successful matches in well-contrasted images, in some cases the lack of image features cause the matching procedure to fail. For this reason we propose an approach based on region matching and selective texture analysis

0-7803-7398-7/02/$17.00 ©2002 IEEE

is proposed in this paper. The extensive use of textural operators can highly improve the accuracy of image correspondences, enabling a better motion estimation, which is used for station keeping.

The paper is organized as follows. First, section 2 describes how to detect features in one image and robustly match them in a reference image centered at the hover point where the vehicle should maintain its position. Next, section 3 describes how these features can be used to estimate the motion of the vehicle. Finally, experimental results on real images are presented and analyzed.

2. Feature Matching

2.1 Introduction

In order to *estimate* and *correct* the vehicle motion, we propose a region-based matching method. This method is based on detecting a set of features in the present image, and then find their corresponding matches in the reference image, which is used as hover point, through texture characterization. However, before performing this computation, there is an important aspect which should be taken into account. Due to construction, camera lenses produce a geometric distortion (radial and tangential) in the image-formation process. Moreover, the underwater environment produces ray diffractions at the camera housing interface. For all these reasons, the first step to be achieved in order to perform measurements with our camera consists of the estimation of a number of intrinsic camera parameters to correct lens distortion [6].

Then, the undistorted image can be used to select the adequate features in the present image to be matched in the reference frame. Therefore, the selection of robust features depends, to a large extent, on the technique used to detect correspondences. Normally, small windows containing high frequencies are quite adequate since they are located in the border of different image textures. For this reason, our feature detector searches for small zones presenting high spatial gradient information in more than one direction, as performed by some corner detectors [7,8]. To do this, the image is convolved with two directional high-pass filters (in the x and y directions). The areas with the highest gradient in both directions are selected. When a feature is selected, the algorithm goes on to search for any other selected features in its neighborhood. If a higher-valued feature exists in this neighborhood, only the best feature is selected as an interest point. This avoids the selection of other features in the same neighborhood and ensures a reasonable distribution of the interest points within the image.

After the detector of features has selected the most reliable points of the undistorted image, we can already search for the correspondences of these features in the reference image. Normally, once the vehicle has approached the working area in teleoperated mode, the pilot selects a frame of the sequence to be used as reference image I_R. However, as the vehicle moves, the present image I may have changed its orientation with respect to the original reference image I_R, caused by the yaw motion of the vehicle. In this case, standard correlation methods may not work properly since they are sensitive to considerable rotations of the images. Moreover, a second problem could appear: underwater currents may displace the vehicle, so that no overlap exists between I and I_R, making impossible the estimation of motion. For this reason, at every iteration of the station keeping system, a visual map of the area surveyed by the vehicle is updated. This visual map is known as *mosaic* in the literature [9,10]. Figure 2 shows the visual map (gray area) which covers the underwater terrain surveyed by the vehicle. The initial reference image I_R is shown at the top left corner of the map. As currents take the vehicle away from the hover point, the station keeping system constructs a mosaic image of the surveyed area. At every iteration of the algorithm, a new reference image I' is extracted through bilinear interpolation from the mosaic image at the previous location and orientation of the camera, being its size a little bigger than the original images to maximize overlapping. Since the transformation from the hover point to the present reference image I' is perfectly known, adequate control signals can be generated to correct the position of the vehicle towards the initial reference image I_R.

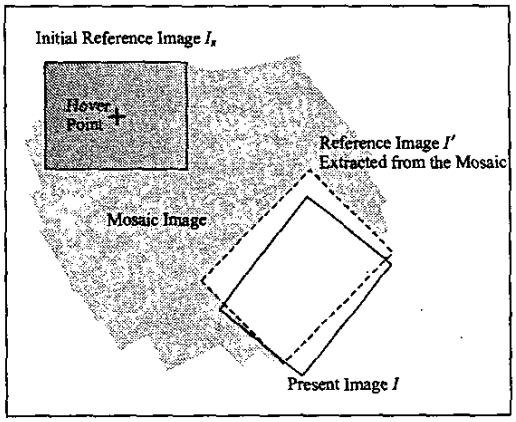

Figure 2: Extraction of the reference image I' from the mosaic image. I' is extracted from the location and orientation of the previous live image. Its size is slightly bigger than that of the captured images, maximizing the overlapping area.

2.2 Similarity Measure

Finding correspondences between images is not an easy task in computer vision, and even less in underwater imaging. On that account we pay special attention to the matching process, carrying out a two step approach, as described in [11]. First, a correlation-based matching strategy is applied to the images [12], selecting a set of candidate matches for a given interest point. Then, a texture characterization of the points is used for selecting the best correspondence. For every interest point in the present image I, a correlation score is computed in reference image I'. This is performed by comparing a small $n \times n$ window centered at the interest point $\mathbf{m} = [x, y]$ with all the possible locations of the feature $\mathbf{m}' = [x', y']$ in the next image. These possible locations of the feature $\mathbf{m}'$ are limited to a window of I', centered at the coordinates of $\mathbf{m}$ in I. The size of this window depends on the motion between consecutive images. Equation (1) is used to compute the correlation score.

$$corr(\mathbf{m}, \mathbf{m}') = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n}\left[I(x+i,y+j)-\overline{I(x,y)}\right]\cdot\left[I'(x'+i,y'+j)-\overline{I'(x',y')}\right]}{n^2\sqrt{\sigma^2(I)\cdot\sigma^2(I')}} \quad (1)$$

where $\overline{I(x,y)}$ is the average of the gray-levels in the $n \times n$ neighborhood, and $\sigma^2(I)$ is the standard deviation of the image I in the $n \times n$ window centered at the interest point $\mathbf{m}$, which is given by:

$$\sigma^2(I) = +\sqrt{\frac{\sum_{i=1}^{n}\sum_{j=1}^{n}\left[I(x,y)\right]^2}{n^2} - \overline{I(x,y)}^2} \quad (2)$$

In this way, it is possible to find in I' a set of possible correspondences $\mathbf{m}_i'$ of every feature $\mathbf{m}$ detected in I. To decide which of these matches is the actual one, the textural characteristics of every feature are analyzed. Given a feature $\mathbf{m}$, a set of statistical-based texture operators is computed on its neighborhood. Different configurations of the energy filters defined by Laws [13], co-occurrence matrix [14] and contrast features [15] of Ojala *et al.* are computed at the location of feature $\mathbf{m}$. These textural features have been chosen for their suitability for underwater imaging and their low cost in terms of computing complexity. A brief explanation of their main characteristics is given below.

Texture energy filters require a pre-filtering of the image with a set of 3×3 masks, giving rise to a subimage of a size $n \times n$ corresponding to the analyzed area. Then, a series of statistical measures have to be computed for every resulting subimage (in our case we have used standard deviation and positive/negative mean):

$$\sigma = +\sqrt{\sum_{i=1}^{n^2}\frac{(c_i-\mu)^2}{n^2}}, \text{ with } \mu = \frac{\sum_{i=1}^{n^2} c_i}{n^2} \quad (3)$$

$$\text{positive mean} = \mu^+ = \frac{\sum_{i=1}^{n^2} c_i}{n^2}, \text{ with } c_i >= 0 \quad (4)$$

$$\text{negative mean} = \mu^- = \frac{\sum_{i=1}^{n^2} c_i}{n^2}, \text{ with } c_i < 0 \quad (5)$$

where n^2 is the size of the vector which stores the neighboring pixels, and c_i is the i^{th} element of this vector.

The *Co-occurrence Matrix* operator defined by Haralik *et al.* in [14] searches the repeated occurrence of pairs of gray-level configurations in the image according to two parameters: the distance and the angle defined by the pair of points. Consider d as the distance between two pixel positions. The immediate neighbors of any pixel can lie on four possible directions: $\theta \in \{0°, 45°, 90°, 135°\}$. The co-occurrence matrix computes the probability of two given gray-levels to appear in the image at a distance d and angle θ. Then, for a given d and θ the rows and columns of the matrix represent the different gray levels of the image, and every position of the matrix corresponds to the frequency of occurrence of that combination of intensities. Like in the case of the Energy Filters, once the co-occurrence matrix has been computed, a set of statistics is computed form the matrix, obtaining the textural characteristics of the image. We have obtained the best results constructing a matrix of 4×4 with distance 1, angle 0 and the entropy statistical measure, as shown in equation (6).

$$\text{Entropy} = \sum_{i,j=0}^{n-1} m_{ij} \cdot \log(m_{ij}) \quad (6)$$

where m_{ij} is the element of row i and column j of the co-occurrence matrix.

Ojala and Pietikäinen [15] proposed a *contrast* feature to be used jointly with *Local Binary Patterns*. The contrast operator consists of performing a gray-scale differentiation in the region which is being considered. The neighboring pixels are compared with the selected point, computing the average of those neighbors with a gray-value higher than that of the center pixel. A second average is computed with the neighbors with an intensity value below the selected pixel. Then, the difference of both averages is

computed. This value is known as *contrast* of the texture.

These three texture operators result in a vector of texture values characterizing every interest point of the present image I, including four measures of the Energy Filter, one of the co-occurrence matrix and another of the contrast measure, giving rise to a 6-component vector which characterizes the texture of every feature. Once this vector of 6 parameters has been computed for a given interest point of the present image I, it is then computed for every candidate match in I'. After a process of normalization, the texture vector of the interest point is compared with the textural properties of all the possible matches by means of the weighted Euclidean distance. A texture similarity measure is then obtained for every possible correspondence.

After this process, every candidate match has two measures of similarity: (i) a block-matching correlation score obtained through equation (1); and (ii) a texture score produced by feature characterization. By averaging these two values, the best correspondence is selected. Therefore, for every interest point in the present image I a unique match is obtained in the reference image I'.

3. Motion Estimation and Station Keeping

Once a set of pairs have been detected in the present and reference images, the station keeping system has to identify the points which describe the dominant motion of the image. A 2D transformation matrix $\mathbf{H}$ which relates the coordinates of a feature in I' with its coordinates in image I is computed:

$$\tilde{\mathbf{m}} = \mathbf{H} \cdot \tilde{\mathbf{m}}' \text{ or } \begin{bmatrix} x_i \\ y_i \\ 1 \end{bmatrix} = \begin{bmatrix} s\cos\theta & -s\sin\theta & t_x \\ s\sin\theta & s\cos\theta & t_y \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x_i' \\ y_i' \\ 1 \end{bmatrix} \quad (7)$$

where $\tilde{\mathbf{m}} = (x_i, y_i, 1)^T$ and $\tilde{\mathbf{m}}' = (x_i', y_i', 1)^T$ denote a correspondence point in images I and I'. The matrix $\mathbf{H}$, which performs this transformation, is known as "homography", and can be computed by Singular Value Decomposition if 2 or more pairs of matches are available. The *similarity* transformation described by matrix $\mathbf{H}$ has 4 degrees of freedom. A more general motion model (e.g. *affine* or *projective*) could be used [16]. However, our vehicle has been designed to be passively stable in pitch and roll (its center of gravity is below the center of buoyancy). For this reason, rolling and pitching motion of the vehicle are very small, and therefore better results are obtained with a similarity motion model.

Although an accurate texture analysis is devoted to the matching procedure, moving objects of the scene (algae, fishes, etc.) could produce some matches describing a motion different from the motion-less sea-floor correspondences which describe the vehicle motion. For this reason, a robust estimation method has to be applied. The *Least Median of Squares* (LMedS) algorithm [17] is used for finding the matrix $\mathbf{H}$ which minimizes the median of the squared residuals M_{err}:

$$M_{err} = \underset{i}{med}\left(d^2\left(\tilde{\mathbf{m}}_i, \mathbf{H}\tilde{\mathbf{m}}_i'\right)\right) + \left(d^2\left(\tilde{\mathbf{m}}_i', \mathbf{H}^{-1}\tilde{\mathbf{m}}_i\right)\right) \quad (8)$$

where $d^2\left(\tilde{\mathbf{m}}_i, \mathbf{H}\tilde{\mathbf{m}}_i'\right)$ is the square Euclidean distance from a point $\tilde{\mathbf{m}}_i$, defined on image I, to the projection on the same image plane of its correspondence $\tilde{\mathbf{m}}_i'$. Hence, the error is defined by the distance of a point to the projection of its correspondence.

Once the pairs of features describing the dominant motion have been selected, a 2D projective transformation matrix relating the coordinates of both images is computed. Initially, the reference image containing the hover point is selected as a base frame. The mosaic coordinate system is placed at the origin of this reference frame. Then, when image I has to be added to the mosaic, a 2D planar transformation ${}^{I'}\mathbf{H}_I$ provides its best fitting with respect to the reference image I', extracted from the mosaic. In order to obtain a global registration from present image I to the mosaic reference frame, the following matrix product has to be performed:

$${}^{I_R}\mathbf{H}_I = {}^{I_R}\mathbf{H}_{I'} \cdot {}^{I'}\mathbf{H}_I \quad (9)$$

where ${}^{I_R}\mathbf{H}_I$ is the homography that produces the coordinates of a point in the mosaic image, from the coordinates of the same point in the present image I. This matrix provides a direct measurement of the vehicle position with respect to the initial reference image I_R.

Finally, the vehicle motion can be recovered. This is done with the aid of an ultrasonic altimeter and the knowledge of the intrinsic parameters of the camera. As the distortion produced by the camera lenses and the ray diffractions at the air/camera-housing/water interfaces has been corrected in the first phase of the station keeping process (correction of lens distortion), the processed images are an ideal projective projection of the ocean floor. That is, the images are an ideal linear projection of the incident light rays, as shown in Figure 3. Therefore, the metric measure Z provided by the altimeter, together with the knowledge of the camera focal length f, can be used to convert the incremental motion estimation from the camera coordinated system (in pixels) to the world reference system (metric information). Applying the geometric law of the perspective relation, the following equation can be obtained:

$$\frac{d}{f} = \frac{D}{Z}, \text{ then } D = \frac{d \cdot Z}{f} \quad (10)$$

When the first image of the sequence is placed in the mosaic, the world coordinate system $\{O_w\}$ is aligned to the XY plane defined by this image, and the initial Z is measured from the altimeter. For every new image, the subsequent homographies provide a 2D estimation of the vehicle motion. Considering the picture illustrated in Figure 3, incremental measure d can be decomposed in d_x and d_y, measured with respect to the coordinate system of the previous image. Therefore, equation (10) can be decomposed in

$$D_x = \frac{d_x \cdot Z}{f}; \quad D_y = \frac{d_y \cdot Z}{f} \tag{11}$$

where (D_x,D_y) are the components of the incremental motion from image I to I', expressed in world coordinates.

Therefore, the 3D position of the vehicle can be obtained from incremental motion (D_x,D_y) and absolute measurement Z.

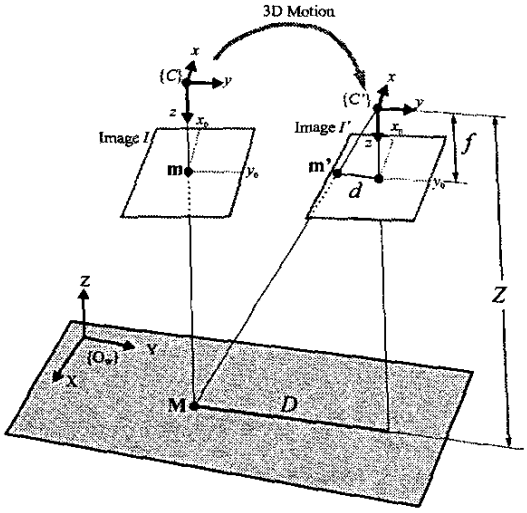

Figure 3: Motion estimation in world (metric) coordinates. The incremental motion d is obtained in pixels from the mosaic. If an estimation of Z is known, *e.g.* from an altimeter sonar, and knowing the camera focal length f, a measure D can be obtained in world coordinates.

4. Experimental Results

We have performed several tests to verify the behavior of our station keeping approach in a real environment. The URIS underwater vehicle has been used for experimental testing of our station keeping system. For the sake of space only one of the sea experiments is reported here. Several tests were performed in Costa Brava (Mediterranean Coast). The trials consisted in teleoperating the vehicle up to the hover point, then the pilot clicks a button to start the station keeping system in teleoperated mode. This means that the sensing system starts to compute its position with respect to the reference image. Teleoperation of the vehicle at this point simulates perturbation of the system. Then, the station keeping system is asked to take control of the vehicle, bringing the robot to the hover point. Figure 4 shows the 3D motion estimation from an initial altitude of 3 meters. The station keeping system is running on an off-board computer.

The reported trial was performed at low depth (5 meters) in a zone with 3D relief, and the vehicle had to keep station 3 meters above the sea-floor. The experiment has been performed on a sunny day. It can be seen in Figure 5 that the waves produce bright spots in the sea floor, which translate into changing image irradiance. However, the station keeping system is able to detect its position within the trial and to conduct the vehicle towards the hover point.

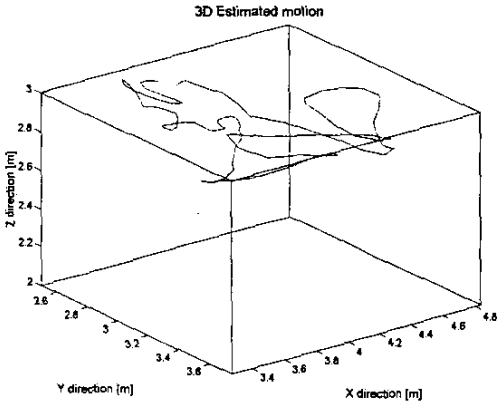

Figure 4: XYZ trajectory. Motion estimation in world (metric) coordinates.

5. Summary

We have described a vision system to automatically maintain the position of the vehicle when it navigates near the sea-floor. A down-looking camera carried by the vehicle provides the images which are processed to estimate the motion of the vehicle. Since processing underwater images is a difficult task, a robust feature-based motion detection technique has been proposed. The exploitation of textural information encoded in the image patches, in addition to classical correlation techniques provides a satisfactory degree of robustness in the estimation of the vehicle position relative to the desired position. Our proposal to solve the correspondence problem is then a mixture of texture-based operators

and gray-level correlation procedures. Data redundancy in the detection of correspondences is the key point to introduce robustness in the matching procedure.

Outdoor experiments have been performed to test the reliability of the system in a real environment. The sea trials have proved the satisfactory performance of the stationkeeping system. The system is able to operate in real time on an off-board computer.

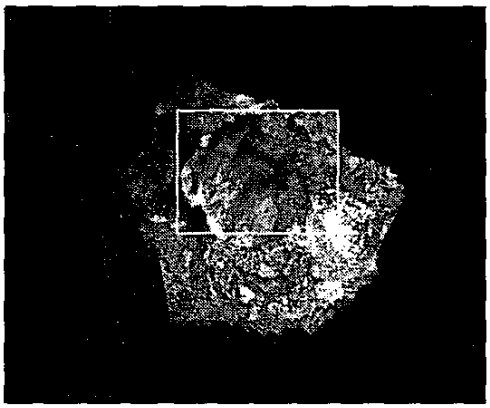

Figure 5: Resulting mosaic image. The white square drawn in the image corresponds to the position of the initial reference image.

6. References

[1] C.J. Funk, S.B. Bryant, P.J. Beckman Jr., "Handbook of underwater imaging system design", Ocean Technology Department, Naval Undersea Center, 1972.

[2] S. Negahdaripour and J. Fox, "Underwater optical station-keeping: improved methods", *Journal of Robotic Systems*, vol. 8, no. 3, pp. 319–338, 1991.

[3] L. Jin, X. Xu, S. Negahdaripour, "A real-time vision-based stationkeeping system for underwater robotics applications", in *Proceedings of the MTS/IEEE OCEANS Conference*, vol. 3, pp. 1076–1081, 1996.

[4] A. Giachetti, "Matching techniques to compute image motion", in *Image and Vision Computing*, no. 18, pp. 247-260, 2000.

[5] R.L. Marks, H.H. Wang, M.J. Lee and S.M. Rock, "Automatic visual station keeping of an underwater robot," in Proc. of IEEE/MTS OCEANS, Vol. 2 , pp. 137-142, 1994.

[6] O.D. Faugeras and G. Toscani, "The calibration problem for stereo," in Proc. of the *IEEE Computer Vision and Pattern Recognition*, pp. 15-20, 1986.

[7] L. Kitchen and A. Rosenfeld, "Gray-Level corner detection," *Pattern Recognition Letters*, vol. 1, no. 2,pp. 95-102, 1982.

[8] C.G. Harris and M.J. Stephens, "A combined corner and edge detector," in Proceedings of the Fourth Alvey Vision Conference, Manchester, pp. 147–151, 1988.

[9] N. Gracias and J. Santos-Victor, "Underwater Video Mosaics as Visual Navigation Maps", *Computer Vision and Image Understanding*, vol. 79, no. 1, pp. 66-91, 2000.

[10] R. Garcia, J. Batlle, X. Cufi, and J. Amat, "Positioning an Underwater Vehicle through Image Mosaicking," in *Proc. IEEE Int. Conf. on Robotics and Automation*, vol. 3, pp. 2779–2784, Seoul, Rep. of Korea, 2001.

[11] R. Garcia, X. Cufi and J. Batlle, "Detection of Matchings in a Sequence of Underwater Images through Texture Analysis," *IEEE Int. Conf. on Image Processing*, vol. 1, pp. 361–364, Thessaloniki, Greece, 2001.

[12] Z. Zhang, R Deriche,. O. Faugeras, Q.T. Luong, "A robust technique for matching two uncalibrated images through the recovery of the unknown epipolar geometry," INRIA RR-2273, 1994.

[13] K.I. Laws, "Textured Image Segmentation," Ph.D. Thesis, Processing Institute, University of Southern California, Los Angeles, 1980.

[14] R.M. Haralick, K. Shanmugan and I. Dinstein. "Textural features for image classification," *IEEE Transactions on Systems, Man and Cybernetics*, vol. 3, no. 6 pp. 610–621, 1973.

[15] T. Ojala and M. Pietikäinen, "Unsupervised texture segmentation using feature distributions," *Pattern Recognition*, vol. 32, pp. 477–486, 1999.

[16] R. Hartley and A. Zisserman, "Multiple View Geometry in Computer Vision," Cambridge University Press, 2000.

[17] P. Rousseeuw and A. Leroy, "Robust Regression and Outlier Detection," John Wiley & Sons, New York, 1987.