SEAFLOOR VIDEO MAPPING: MODELING, ALGORITHMS, APPARATUS

Y. Rzhanov, L. Huff, G. R. Cutter

Center for Coastal and Ocean Mapping (C-COM)
University of New Hampshire
Durham 03824, USA
email: yuri.rzhanov@unh.edu, lhuff@cisunix.unh.edu, gcutter@cisunix.unh.edu

ABSTRACT

This paper discusses technique used for construction of high-resolution image mosaic from a video sequence and the synchronously logged camera attitude information. It allows one to infer geometric characteristics of the imaged terrain and hence improve mosaic quality and reduce computational burden. The technique is demonstrated using numerical modeling and is applied to video data collected on Rainsford Island, Mass.

Calculation of the transformation relating consecutive image frames is an essential operation affecting reliability of the whole mosaicing process. Improvements to the algorithm are suggested, which significantly decrease the possibility of convergence to an inappropriate solution.

1. INTRODUCTION

In the previous publications [1, 2] we have discussed potential use of auxiliary sensors (namely, attitude sensor and GPS receiver) for improvement of quality of image mosaics, in particular, for underwater imagery. These mosaics have important applications for visual survey of large areas or objects under the conditions of poor visibility. Our primary interest lies in area of benthic habitat mapping (e.g. [3]), although pipeline inspection, sediment characterization, archaeological and forensic site mapping could also be potential applications of this technique, to name a few. Currently video mosaicing is the only mapping method which provides resolution better than 1 cm per pixel and rapid continuous coverage of large areas.

It seems obvious that accurate measurement of camera position and orientation can greatly facilitate the construction process of a synthetic mosaiced image. The difficulty however is that existing low-cost sensors are not accurate enough to be directly used in the mosaicing process. Indeed, a GPS receiver may provide information about location of the antenna with an accuracy of 10-20 cm. If the camera is underwater and antenna is on the surface, the accuracy is reduced to 1-2 meters, yet the typical resolution of imagery is less than 1 cm. Even reasonably accurate measurements of the camera Euler angles are not sufficient to position a pixel to within 1 cm. For example, 0.5 degree error in tilt measurement easily translates into 2.5 cm error for 30 degree camera pitch (for typical imaging distances).

Assumption of flat horizontal (FH) terrain allows to derive unique functional relationship between position and orientation of the camera and elements of transformation (homography) used to create a mosaiced image. If terrain is indeed described well by the FH model, resulting mosaic is of good quality and camera motion can be recovered; difference between homographies calculated from the camera vectors and ones obtained from image co-registration procedures stays small. This difference going up indicates that the FH model becomes inapplicable. More sophisticated models, however, cannot solve for camera motion and terrain characteristics simultaneously - solution is not unique. By employing sensor measurements we eliminate ambiguity and infer some local topographic characteristics of the terrain.

2. MODELING OF THE ACQUISITION AND RE-PROJECTION PROCESSES

To understand the mosaicing process in details, a model of video acquisition was developed. The modelled imaged surface is described in terms of digital elevation model (DEM) in association with a raster image, draped over the topography. Both elevation and luminance (or color) values are defined on a regular grid and are assumed to change monotonically in between. The DEM is with respect to a ”zero level plane”, which is normal to Y axis of the associated system of coordinates and passes through point $Y = H$. Camera orientation and position is described by 6 parameters, combined in an ”Euler vector”, $\vec{E}$, components of which represent pitch, roll and yaw of the camera, and translation vector $\vec{S}$ determining offset with respect to the center of origin, $\vec{O}$. In case of zero Euler vector, ideal pinhole camera is oriented along axis X (”North”) and is looking at the terrain vertically down (along -Y direction). ”West” direction then corresponds to Z axis. The acquired image (frame) lies in XZ-plane, ”up” direction being ”North”, and ”left” being ”West”. (The described system is widely used in graphic display field.)

We also need to introduce number of pixels in the frame, Nx (columns) and Ny (rows), and camera field of view in one of the directions, say, horizontal, FoV. Then the normalized focal distance, F, is equal to $\frac{1}{2\tan FoV/2}$.

In case of zero Euler vector, flat horizontal terrain and $H = F$, focal plane coincides with the imaged surface, and the mapping between grid points of the terrain and image pixels is determined by a unit homography. Decrease in H results in zooming in the imaged surface, so we can introduce zoom ratio $Z = \frac{F}{H}$. Shift of the camera in the plane parallel to the terrain leads to the corresponding shift of the imaged area; shifts are convenient to normalize by number of pixels in the horizontal direction (two images shifted horizontally with 50 percent overlap have normalized shift $D_x = 0.5$).

If the terrain is described by the FH model, geometric considerations provide a unique relationship between camera parameters $(\vec{E}, Z, D_x, D_y)$ and coefficients of the rectification homography R, *i.e.* transformation between the coordinates of terrain W and pixels of the frame I (see Appendix). Re-projection process - mapping of pixels onto the terrain space - can be denoted symbolically as $W = R \times I$. (Note that the normalized shifts in image space D_x and D_y correspond to shifts in Z- and X- directions in the real space.)

Two frames I_0 and I_1 imaging approximately the same area could be co-registered, that is, a homography T_{01} could be found, mapping frame I_1 onto the image space of I_1, thus combining them in a mosaic. Denote this symbolically as $I_0 = T_{01} \times I_1$. If rectification (world) homography is known for frame I_0, then the world homography for frame I_1 is $R_1 = R_0 \times T_{01}$, as $W = R_0 \times I_0 = (R_0 \times T_{01}) \times I_1 \equiv R_1 \times I_1$. Other frames, related to I_0 through a chain of relative homographies, $T_{j,j+1}$, could be added to the global mosaic in the same way: $W = R_0 \times \prod_{j=0}^{k-1} T_{j,j+1} \times I_k$. The latter formula shows that if the rectification homography for the initial frame, R_0, was estimated inaccurately, this would affect global mosaic as a whole, while numerical errors accumulated in co-registration procedures of finding relative homographies, $T_{j,j+1}$, may lead to local distortions of the mosaic.

If residual errors are negligible, rectification homographies for frames in the acquired sequence can be directly used for construction of the global mosaic. Relative homographies guarantee optimal pairwise merging of the frames - rectification homographies guarantee optimal merging of all frames on a common image space (terrain map).

Modeling of video acquisition process with non-flat terrain shows that the above statements do not hold true. Relative homographies found using one of the co-registration procedures (typically, based on either feature tracking or optimization) may not correspond to any set of vectors describing the camera, as the model, used to relate these parameters to the elements of homography, fails. We suggest a more sophisticated model, which describes the terrain elevation in terms of 2D low-order polynomials, with coefficients being model variables, as well as the camera vectors $\vec{E}$ and $\vec{S}$.

Unlike the case with the flat horizontal terrain, the solution of the above problem is not unique. Indeed, flat tilted terrain and vertical camera case is indistinguishable from the case when the terrain is horizontal and the camera is tilted. In this situation we obtain the camera tile values from the sensor measurements. Sensor inaccuracies will then result in inaccurate determination of the terrain topography.

3. PROCESSING ALGORITHM IMPROVEMENTS

Robust and accurate calculation of relative homographies is one of the most important parts of the mosaicing process. We are employing optimization technique based on the so called brightness constancy constraint. (It should be noted that depending on visibility and lighting conditions acquired images may need some pre-processing - filtering [4, 5] or de-trending [1].) Although consecutive frames typically have much in common, successful optimization heavily depends on the initial guess used in iterative procedure. Our strategy is to set certain threshold for the average per-pixel error (optimization is considered to be successful if the final error is below this threshold) and to run optimization for several initial guesses (this stage can be easily parallelized). Two common candidates for initial guesses are: (a) successfully found transformation for the previous pair of frames, and (b) unit transformation. In approximately 2 percent of cases these guesses led to a non-global minimum with high residual error. However in these cases estimates of camera Euler angles obtained from the sensor may be used to provide the initial guess for a relative homography. The algorithm is as follows:

1. Using Euler angles for both frames I_1 and I_2 corresponding approximate rectification homographies R_1^A and R_2^A are calculated. Note that translations D_x and D_y remain zero for these estimates.

2. Frames are re-projected onto flat horizontal plane using these homographies:

$$I_k^A = R_k^A \times I_k, k = 1, 2$$

3. Featureless frequency domain-based technique [6, 1] is used to estimate rigid affine transformation between

re-projected images:

$$I_1^A = A_{12} \times I_2^A$$

Note that this method is non-iterative and highly robust.

4. Initial guess for the relative homography T_{12}^{init} is obtained from the above transformations:

$$I_1 = \left(R_1^A\right)^{-1} \times A_{12} \times R_2^A \times I_2 \equiv T_{12}^{init} \times I_2$$

$$T_{12}^{init} = \left(R_1^A\right)^{-1} \times A_{12} \times R_2^A$$

From our experience, this technique, which is computationally intensive, provides optimal choice of initial guess for optimization procedure.

Some failures in finding relative homography were associated with non-stationary content of the frames - objects moving across the camera field of view. The technique outlined above was still able to provide a reasonable initial guess provided the moving objects did not occupy more than 30 percent of the frame. In this case, however, co-registration results in average per-pixel error being significantly higher than the pre-set threshold, and each situation requires manual intervention. Once it was established that co-registration procedure did not fail, the regions with high local pixel error were marked and in creation of the final mosaic only one instance of input video data was used, to decrease blurring effects.

4. APPARATUS AND RESULTS

A consumer grade Sony digital video camera is connected to a laptop computer by a Control-L device, allowing constant monitoring of timecode (which uniquely identifies recorded frame) more than 30 times per second. The NMEA output messages from the attitude sensor and GPS receiver are received on serial inputs of the computer, typically coming 5-8 times per second. The monitoring program logs every sensor message, synchronously with the timecode, GPS message and internal CPU time. Records from the log file are used in post-processing, where frames corresponding to neighboring records are co-registered - this results in a chain of relative homographies relating any two frames in a sequence. The equipment were arranged on a cross arm atop a pole; the GPS antenna was mounted in the center of the arm, the sensor and the camera - at opposite sides of it.

Attitude sensor measurements were found to contain spikes that were not related to actual sensor orientation. To minimize their influence, we have smoothed the measurements, using *sinc* weighting function over 1 sec time span. Similar smoothing was applied to GPS position measurements, latitude and longitude readings being smoothed independently, as mosaiced areas are relatively small, typically less than 100 meters.

For the chosen initial frame it is assumed that sensor measurements do not contain errors, and that at that moment the camera was located in the center of origin. This allows calculation of the rectification homography for the initial frame and, through the chain of relative transformations, rectification homographies R_k^{exp} for all consecutive frames. Employing the simplified model discussed above, camera vectors are found for each rectification homography. Each pair of camera vectors corresponds to some ”model” rectification homography R_k^{model}, and difference between R_k^{exp} and R_k^{model} indicates the deviation from the flat horizontal model.

Figure 1 shows a comparison between the measured value of yaw (dotted line) and yaw calculated from the homographies (solid line) as functions of the frame number. Measured and calculated values differ only in the regions where homographies' error $E_k = \parallel R_k^{exp} - R_k^{model} \parallel$ is significant (Figure 2). Frames with the large error are associated with images of terrain that cannot be described by the FH model.

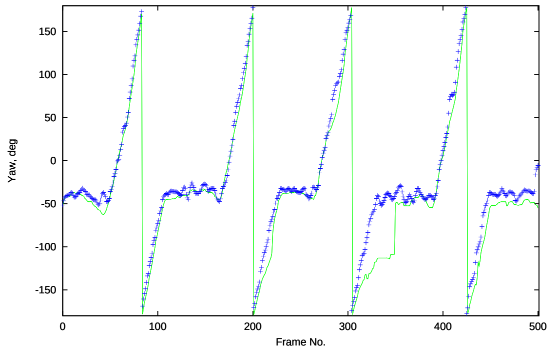

Fig. 1. Yaw of the camera: measured by the sensor and calculated. The survey procedure consisted of moving in a straight line, and periodic rotation of the pole.

Local deviations from the flat terrain (with non-flat area much less than total area of the frame) do not significantly affect the calculated relative homography, but results in a local mismatch between the pixel values of the co-registered frames. This local per-pixel luminance difference may be processed using any ”shape-from-stereo” technique, thus obtaining information about shape of the imaged surface. For our purpose, however, it is sufficient to mark these mismatch areas at the pre-processing stage, and, when the final mosaic is created, to use a single instance of the input video data to fill them in, as opposed to weighted average, typically used for ”feathering”.

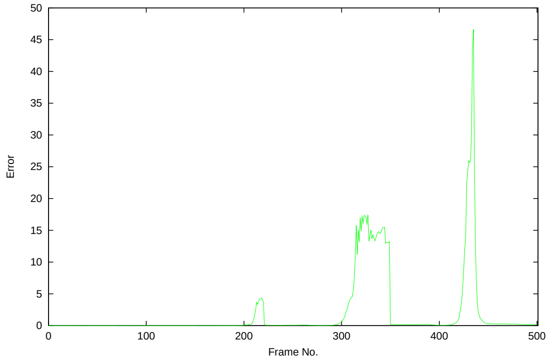

Fig. 2. Error indicating failure of the flat horizontal model.

5. CONCLUSIONS

This paper outlines ways to use data from an attitude sensor attached to a video camera to simplify processing of acquired image sequences and to facilitate construction of a global mosaic - geo-coded image of the underwater surface. The method has been applied to inter-tidal marine environments and provides a cost-effective alternative to previously reported methods [5]. The technique shows much promise for gathering new types of information from the seabed by using video imaging as opposed to traditional acoustic imaging.

Appendix

Let the Euler vector $\vec{E} = (\theta, \phi, \psi)$, and denote $c\tau \equiv \cos\tau$, $s\tau \equiv \sin\tau$, where $\tau = \theta, \phi, \psi$. It is customary to write homography T as:

$$T = \begin{pmatrix} p_0 & p_1 & p_2 \\ p_3 & p_4 & p_5 \\ p_6 & p_7 & 1 \end{pmatrix}$$

where homography elements can be expressed as functions of camera parameters:

$$\begin{aligned}
p_0 &= U[(s\theta s\psi s\phi - c\psi c\phi)/Z - (D_x/F)(s\theta s\psi c\phi + c\psi s\phi)] \\
p_1 &= U[(-s\theta c\psi s\phi - s\psi c\phi)/Z + (D_x/F)(s\theta c\psi c\phi - s\psi s\phi)] \\
p_2 &= U[F(c\theta s\phi)/Z - D_x(c\theta c\phi)] \\
p_3 &= U[(c\theta s\psi)/Z - (D_y/F)(s\theta s\psi c\phi + c\psi s\phi)] \\
p_4 &= U[(-c\theta c\psi)/Z + (D_y/F)(s\theta c\psi c\phi - s\psi s\phi)] \\
p_5 &= U[F(-s\theta)/Z - D_y(c\theta c\phi)] \\
p_6 &= -U(s\theta s\psi c\phi + c\psi s\phi)/F \\
p_7 &= U(s\theta c\psi c\phi - s\psi s\phi)/F
\end{aligned}$$

where $U = -c\theta c\phi$.

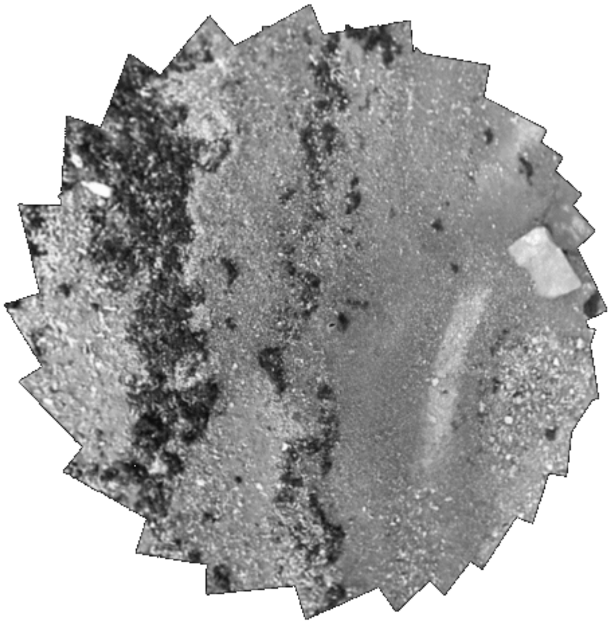

Fig. 3. Example of video mosaic of intertidal marine environment.

6. REFERENCES

[1] Y. Rzhanov, L. M. Linnett, R. Forbes, ”Underwater mosaicing for seabed mapping.” 2000 International conference on Image Processing, ICIP2000, Vancouver, Canada, vol. 1, pp. 224-227, 2000.

[2] Y. Rzhanov, G. R. Cutter, L. Huff, ”Sensor-assisted video mosaicing for seafloor mapping.” 2001 International conference on Image Processing, ICIP2001, Thessaloniki, Greece, vol. 2, pp. 411-414, 2001.

[3] V. E. Kostylev, B. J. Todd, G. B. Fader, *et al.* ”Benthic habitat mapping on the Scotian Shelf based on multibeam bathymetry, surficial geology and sea floor photographs.” MARINE ECOLOGY- PROGRESS SERIES, v. 219, pp. 121-137, 2001.

[4] R. L. Marks, S. M. Rock, M. J. Lee, ”Real-time video mosaicking of the ocean floor.” IEEE Journal of Oceanic Engineering, vol. 20, pp. 229-241, 1995.

[5] H. Singh, L. Whitcomb, D. Yoerger, O. Pizarro, ”Microbathymetric mapping from underwater vehicles in the deep ocean.” Computer Vision and Image Understanding, vol. 79, pp. 143-161, 2000.

[6] B. S. Reddy and B. N. Chatterji, ”An FFT-based technique for translation, rotation and scale-invariant image registration.” IEEE Transactions on Image Processing, vol. 5, pp. 1266-1271, 1996.