MAPPING GROUND VIDEO TO AERIAL DEM'S [*]

Amit K. Agrawal and Chandra Shekhar
Center for Automation Research
University of Maryland
College Park, MD 20742
and
Philip David and Jeff DeHart
Army Research Laboratory
Adelphi, MD 20783-1197

ABSTRACT

We present an approach for registering an aerial Digital Elevation Model (DEM) with a color intensity image obtained using a camera mounted on a mobile robot. An approximate measurement of the camera pose is obtained using auxiliary sensors on-board the robot. The DEM is transformed into a depth map in the camera's coordinate system using this initial pose. The problem is now simplified to the alignment of two images, one containing intensity information, and the other, depth. Region boundaries in the intensity image are matched with discontinuities in the depth map using a robust directed Hausdorff distance. This cost function is minimized with respect to the six parameters defining the camera pose. Due to the highly non-linear nature of cost function with multiple local minima, a stochastic algorithm based on the downhill simplex principle is employed for minimization. Results on real data are presented.

Index Terms: *Digital Elevation Map, Mobile Robot, Video, Registration, Depth Map, Simulated Annealing, Simplex algorithm, Hausdorff Distance, Color Segmentation*

INTRODUCTION

There has been considerable interest recently on using autonomous mobile robots in surveillance. The ability to send mobile, sensor-equipped robots into environments that are potentially hazardous to humans is of vital importance in a number of scenarios (e.g. nuclear/biological/chemical contamination). There is an urgent need for robust, real-time algorithms for exploiting the data collected by the sensors mounted on the robots in order to improve the operators awareness of the scene. The operator's control station often has access to high-resolution (e.g., 1 meter) elevation data of the environment in which the robots are operating. In such a situation, it would be very useful to be able to integrate video from the robots with elevation data to provide the operator with a more accurate picture of the environment.

This is essentially a registration problem, posing the following challenges: (a) Since the DEM and the video are obtained from completely different viewpoints, and have different dimensionalities, it is not possible to use simple techniques such as image correlation. (b) Typically, the pose of the robots video sensor is known only approximately, which means that the search space for video-DEM alignment is very large.

Although there is very little published literature on video-DEM registration, a number of researchers have worked on related problems such as 3D modeling and multisensor registration. Zisserman et al. [Fitzgibbon and Zisserman, 98] have worked on the automatic construction of 3D models of a scene from a sequence of closely spaced 2D images. Pollefeys et al. [Koch, Pollefeys and Gool, 2000] reconstruct realistic surface of 3D scenes from uncalibrated image sequences. Both these methods obtain depth estimates that depend on image texture and camera geometry, and do not use range data. Fruh and Zakhor [Fruh and Zakhor,

[*] Prepared through collaborative participation in the Advanced Decision Architectures Consortium sponsored by the U. S. Army Research Laboratory under the Collaborative Technology Alliance Program, Cooperative Agreement DAAD19-01-2-0009. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation thereon.

2002] generate textured 3D buildings facade meshes from laser scans and camera images. Stamos and Allen [Stamos and Allen, 2000] build 3D model from range data using a volumetric set intersection method and then identify planar regions. Line features from planar intersections are used for registration with 2D image lines. Elstrom uses a stereo-based method for registration of color and LADAR (Light Amplitude Detection and Ranging) images by extracting corresponding points and estimating the rotation and translation from them. Li and Manjunath [Besl and Manjunath, 1995] use an elastic contour matching scheme for registering multi-sensor images assuming a 2-D affine transformation between corresponding points in two images.

In our approach, an approximate measurement of the camera pose is obtained using auxiliary sensors onboard the robot. The DEM is transformed into a depth map in the camera's coordinate system using this initial pose. The problem is now simplified to the alignment of two images, one containing intensity information, and the other, depth or in other words obtaining the relative translation and rotation between the two views. The alignment step involves feature extraction, matching and camera pose correction. Based on the corrected camera pose, it is then possible to map the video texture to the depth map. Figure 1 shows the block diagram for the entire registration process. The rest of this paper is organized as follows. We first discuss the geometry of the problem and then outline the various steps in the registration procedure. Experimental results on real data are then presented followed by conclusions in the end.

GEOMETRY

We work in projective 2- and 3- space, representing points in homogeneous coordinates. A 3D point **X** is represented as $(x, y, z, 1)^T$ and a 2-D point **x** is represented as $(x, y, 1)^T$. Mobile robots used for autonomous navigation and video surveillance are usually equipped with a variety of secondary sensors that provide *metadata* in the form of measurements of the position and orientation of the imaging sensor.

With the geometry shown in Figure 2, a 3D point X_{wc} in world coordinate system can be represented in the

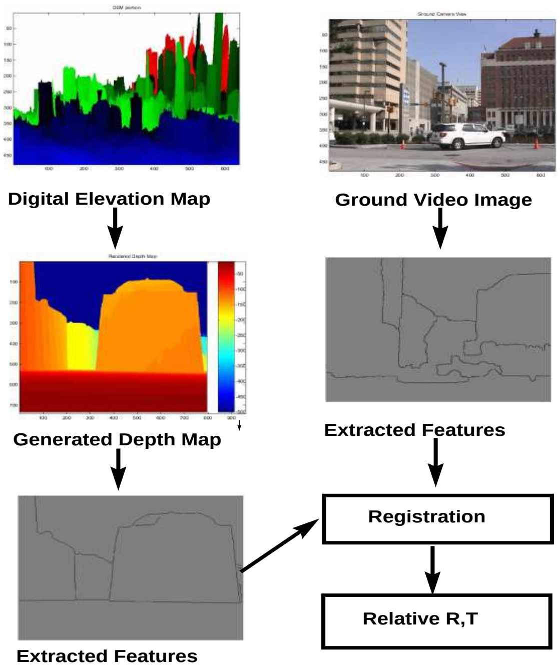

Figure 1: Block diagram of the registration process

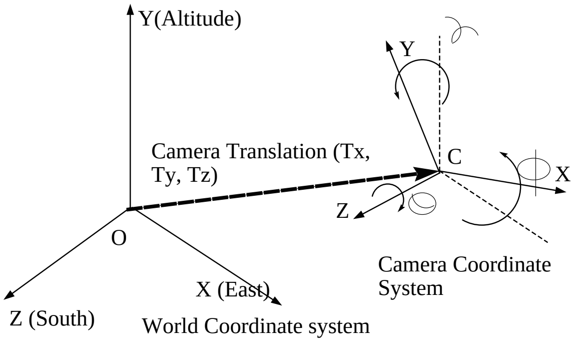

Figure 2: Geometric relationship between the world (X_w, Y_w, Z_w) and camera (X_c, Y_c, Z_c) coordinate systems. O denotes the center of the world coordinate system and C, the location of camera center denoted (T_x, T_y, T_z) in the world coordinate system specifies the translation of the camera. The *pan* angle or the heading direction, ϕ, is measured counter-clockwise with respect to north. The *tilt* angle γ is measured with respect to the vertical Y axis. The *roll* angle θ is measured clockwise with respect to the optic axis of the camera.

camera coordinate system as X_{cc}, given by

$$X_{wc} = R * X_{cc} + T \tag{1}$$

where $T = (T_x, T_y, T_z)^T$ and *R* denotes the 3*3 rotation matrix.

$$R = \begin{bmatrix} c_\phi.c_\theta + s_\gamma.s_\phi.s_\theta & -c_\phi.s_\theta + s_\gamma.s_\phi.c_\theta & c_\gamma.s_\phi \\ c_\gamma.s_\theta & c_\gamma.c_\theta & -s_\gamma \\ -s_\phi.c_\theta + s_\gamma.c_\phi.s_\theta & s_\phi.s_\theta + s_\gamma.c_\phi.c_\theta & c_\gamma.c_\phi \end{bmatrix} \tag{2}$$

where $s_\theta = \sin(\theta)$, $c_\theta = \cos(\theta)$, $s_\phi = \sin(\phi)$, $c_\phi = \cos(\phi)$, $s_\gamma = \sin(\gamma)$ and $c_\gamma = \cos(\gamma)$.
Based on these measurements, it is possible to compute a 3*4 projection matrix mapping (in homogeneous coordinates) points in the 3D world to points in the image. The camera mapping from 3D to 2D is given by perspective projection equation

$$x = PX \tag{3}$$

where P is the 3*4 projection matrix. Given P and the depth Z at each pixel **x** in the image, the corresponding 3D point **X** can be obtained using equation 3. The projection matrix P can be decomposed as

$$P = K[R|T] \tag{4}$$

where K is a 3*3 upper triangular matrix specifying the internal camera calibration parameters.

$$K = \begin{bmatrix} f_x & \alpha & p_x \\ 0 & f_y & p_y \\ 0 & 0 & 1 \end{bmatrix} \tag{5}$$

where f_x, f_y are the focal lengths in the x and y directions, α is the skew parameter, and (p_x, p_y) is the principal point location. In most cameras, it is not unreasonable to assume $f_x = f_y = f$ and $\alpha = p_x = p_y = 0$. Since the camera is known a priori, it may be calibrated off-line to find f and the other components of K.
The registration process can be viewed as determining the relative rotation and translation between the depth map and intensity image, or determining the absolute pose of the camera in the world coordinate frame. Let the camera pose for one of the views be R_1, T_1 in world coordinate frame and the relative rotation and translation for registration be (R_c, T_c), then the camera calibration parameters for the second view in the world coordinate system are

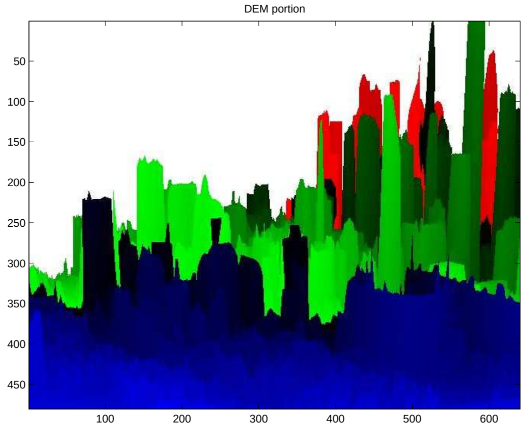

Figure 3: Sample DEM

$$R_2 = R_1 * R_c \tag{6}$$
$$T_2 = R_1 * T_c + T_1 \tag{7}$$

Note that the rotation and translation for 3D points between the two views will be given by

$$R_{3D} = R_c^{-1} \tag{8}$$
$$t_{3D} = -R_c^{-1} * T_c \tag{9}$$

CREATING A DEPTH MAP FROM DEM

The input data to our system consists of a 1m-resolution DEM of downtown Baltimore and ground video captured using an uncalibrated tripod-mounted digital camcorder. The metadata in our current set-up consists of hand-held GPS and compass measurements, which are not synchronized with the video. Our future experiments will use video data and synchronized metadata from an actual mobile robot. The DEM gives the height (Y coordinate) of 3D points at each grid point of a rectangular grid, whose dimensions define the X and Z axes. Figure 3 shows a portion of the DEM. The DEM, which is a rectangular grid of height values, is mapped to the camera coordinate system to create the depth map. Each point in the rectangular grid can be considered to be connected to its four nearest neighbors (to the north, south, east, and west). A ”quad mesh” is built from this underlying four-connected grid by joining each 3D point with its

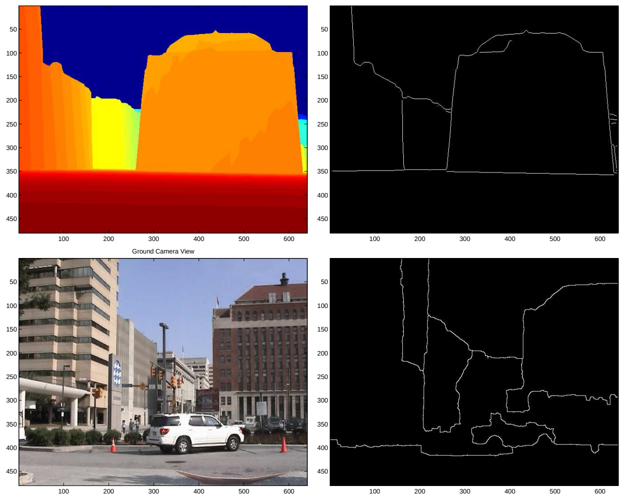

Figure 4: Top: Depth map rendered from DEM and its extracted edges, Bottom: Video frame and its Color segmented region boundaries

neighbors so as to form four triangular patches. Points on the edge of the surface have fewer than four neighbors, giving rise to fewer triangular patches.

The quad-mesh can be rendered from any viewpoint to get a depth map, in which each pixel represents the depth of the corresponding 3D point in the camera coordinate system. Figure 4 (top left) shows the depth map as a color image. Each pixel value represents the distance (Z) along the camera's principal axis.

REGISTRATION

The registration procedure determines the camera pose that brings the two edge maps into maximal alignment. It is based on matching discontinuities in the depth map and in the intensity image.

EDGE EXTRACTION

The DEM is pre-processed to remove low altitude points due to bushes, vehicles etc. The heights of all 3D points below a threshold (5m in our experiments) are set to zero. The data are then median filtered to remove speckle noise. Edges in the depth map correspond to large depth discontinuities. They are extracted from depth map using Canny's edge detector. Figure 4 (top right) shows the extracted edges from the depth map. The 3D locations in the camera coordinate system corresponding to the ”edgels” can be obtained using perspective projection equations, assuming camera calibration is available. In outdoor environments, it is reasonable to assume that depth discontinuities will be associated with color discontinuities. The intensity image is color segmented using the method described in [Deng, Manjunath and Shin, 1999], extracting only coarse region boundaries which are likely to correspond to depth changes. Figure 4 also shows a frame from the ground video and segmented region boundaries for the same.

COST FUNCTION

After feature extraction, the registration problem can be posed as the determination of the six-parameter rigid 3D transformation that best aligns the projected feature points from the depth map with the color region boundaries in the intensity image. For any camera translation and rotation (R_c, T_c), the 3D points are transformed according to equations (8,9). Their projections in the image plane taken are then matched with the color image boundaries, using a robust version of the Hausdorff distance (HD) [Huttenlocher, Klanderman and Rucklidge, 1993], which measures the extent to which each point of a ”model” set lies near some point of an image set. The Hausdorff distance can be efficiently computed from the distance transform of the segmented intensity image. Let **S3D** denote the set which contains the projection of 3D feature points and **S2D** denotes the set containing the pixels lying on the segmented color image boundaries. Then the cost function is defined as [Sim, Kwon, and Park, 1999]

$$E(S3D, S2D) = \frac{\sum_{i=1}^{H} d_{S2d}(a)_i}{H} \tag{10}$$

where $d_{S2d}(a)$ represents the minimum distance value at a projected 3D point to any pixel in S_{2D}, $H = h * N_{S3D}$, N_{S3D} represents the number of 3D points obtained from feature extraction, and $d_{S2d}(a)_i$ represents the *ith* distance value in the sorted sequence $d_{S2d}(x)_1 <= d_{S2d}(x)_1 ... <= d_{S2d}(x)_{N_{S3D}}$. The cost function is minimized based on the distance values that are left after outliers have been filtered out. The parameter *h* was set to be 0.6 in our experiments.

Note that here we do not use the symmetric Hausdorff measure as it will involve computing the distance transform at each iteration for the projected 3D points, and hence will be highly computationally intensive. However, this measure can be used to assess the accuracy of the final solution obtained.

OPTIMIZATION

The cost function as defined in previous section is highly nonlinear with multiple local minima. Since the Hausdorff distance does not have an analytical form, iterative techniques based on gradient descent would require numerical derivatives. Instead, we use the downhill simplex method [Nelder and Mead] which requires only function evaluations, not derivatives. To avoid being trapped in local minima, simulated annealing [Geman and Geman, 1984] principles are incorporated.

EXPERIMENTS AND RESULTS

A 3D mesh is generated form the DEM data in *OpenGL*. Depth maps are rendered using the *z buffer* of OpenGL which gives the depth at each image pixel in the camera coordinate system. Since we currently do not have access to ground truth, we ininitially establish the validity of the approach by registering two depth maps with each other, treating the first as a regular depth map and the second as an intensity image. Figure 5 (top) shows an overlay of projected 3D feature points and intensity edge points) of two depth maps before and after registration. As seen from the image, in spite of a large inintial displacement of about 50 percent of the image size (corresponding to an error in the pan angle of 15 degrees), the two depth maps could be successfully registered using the proposed technique. Figure 6 shows the evolution of the transformation parameters and the cost function as a function of the annealing iterations. The approach is then applied to the problem of registering a depth map with an intensity image. Figure 5 (bottom) shows the overlay of feature points before and after registration for a typical example. Notice that the many pixels in color image boundaries do not have corresponding edge points in the depth map, emphasizing the need for robust measures.

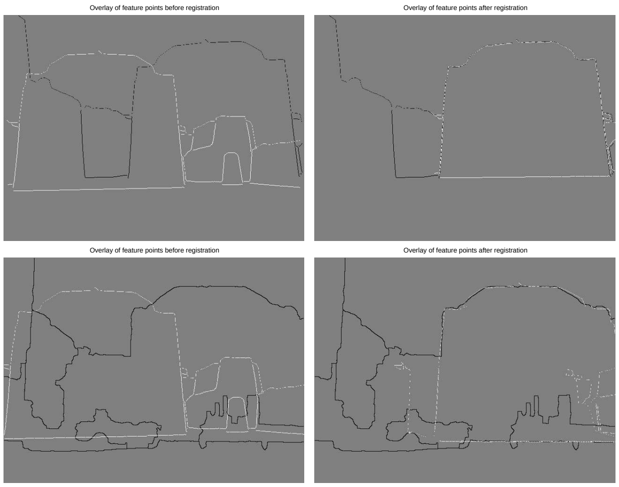
Overlay of feature points before registration
Overlay of feature points after registration
Overlay of feature points before registration
Overlay of feature points after registration

Figure 5: Overlay of feature points before and after registration. Top: Depth map with Depth map, Bottom: Depth map with Intensity image

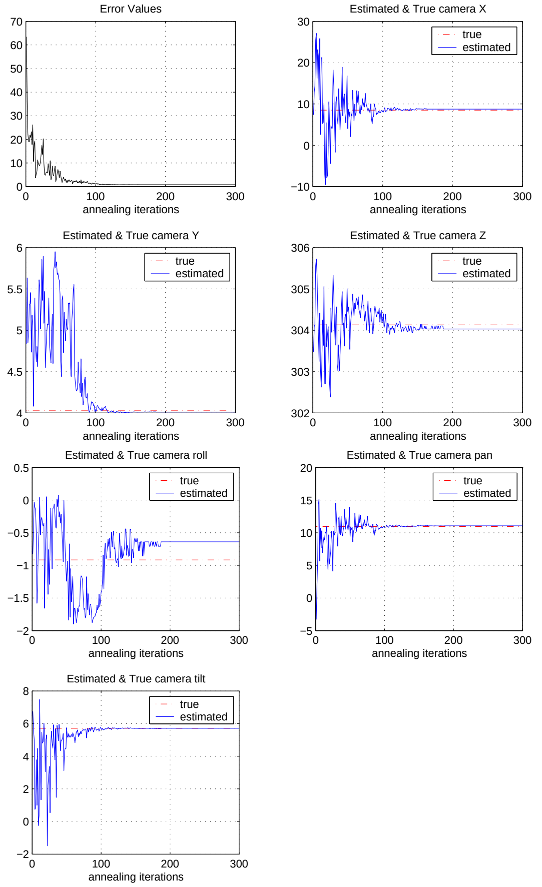

Figure 6: Evolution of Translation and Rotation parameters with annealing iterations

CONCLUSIONS

As a first step towards video-DEM registration for mobile robots, we have presented an approach for registering a DEM to an intensity image. The problem is challenging due to the disparate nature of the two types of data, and the large search space for transformation parameters. The task is made manageable by determining an initial camera pose from the metadata provided by auxiliary sensors. The complex nonlinear nature of the resulting optimization problem is tackled using a stochastic minimization strategy. Preliminary results of the approach are promising; however some problems remain and will be addressed in future work. The approach will also be extended to deal efficiently with a video sequence rather than a single image.

ACKNOWLEDGEMENTS

We are grateful to Larry Tokarcik of ARL for providing us with data and participating in technical discussions. We are also grateful to Prof. Rama Chellappa for his valuable feedbacks and suggestions.

DISCLAIMER

The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Laboratory or the U. S. Government.

REFERENCES

Fitzgibbon, W. A. and Zisserman, A., ”Automatic 3D model acquisition and generation of new images from video sequences”, *Proc. of European Signal Processing Conf. (EUSIPCO ’98), Rhodes, Greece*, pp. 1261-1269, 1998.

Koch, R., Pollefeys, M. and Gool, L.V., ”Realistic Surface Reconstruction of 3D Scenes from Uncalibrated Image Sequences”,*Journal Visualization and Computer Animation*, Vol. 11, pp. 115-127, 2000.

Fruh, C. and Zakhor, A., ”Data Processing Algorithms for Generating Textured 3D Building Faade Meshes From Laser Scans and Camera Images”, *Proc. 3D Data Processing, Visualization and Transmission 2002*, Padua, Italy, pp. 834 - 847, June 2002.

Stamos, I and Allen, P.K. , “3-D Model Construction Using Range and Image Data”, *CVPR*, June 13-15, 2000.

Elstrom, M.D , “A Stereo-Based Technique for the Registration of COLOR and LADAR Images”, *M.S. thesis*, University of Tennessee, Knoxville

Besl, Li, H. and Manjunath, B.S. , “A Contour based Approach to Multisensor Image registration”, *IEEE Transactions on Image Processing*, Vol. 4, No. 3, March 1995.

Besl, P.K. and McKay, N.D. , “A Method for Registration of 3-D Shapes”, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 14, pp. 239-256, February 1992

Deng, Y., Manjunath, B.S., and Shin, H., ”Color image segmentation”, *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 1999.

Huttenlocher, D.P., Klanderman, G.A., Rucklidge W.J., ”Comparing Images using the Hausdorff Distance”, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 15, pp. 850-863, Sept. 1993.

Sim, D.G., Kwon, O.K., and Park, R.H., ”Object Matching Algorithms Using Robuts Hausdorff Distance Measures”, *IEEE Transactions on Image Processing*, Vol. 8, NO. 3, March 1999.

Nelder, J.A. and Mead, R., ”A Simplex Method for Function Minimization”, *Computer Journal*, Vol. 7, pp. 308-313.

Geman, S. and Geman, D., ”Stochastic relaxation, Gibbs distribution, and Bayesian Restoration of Images”, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 6, No. 6, pp. 721-741, 1984.