Homography-Based 3D Scene Analysis of Video Sequences *

Mei Han **Takeo Kanade**
meihan@cs.cmu.edu tk@cs.cmu.edu
Robotics Institute
Cargenie Mellon University
Pittsburgh, PA 15213

Abstract

We propose a framework to recover projective depth based on image homography and discuss its application to scene analysis of video sequences. We describe a robust homography algorithm which incorporates contrast/brightness adjustment and robust estimation into image registration. We present a camera motion solver to obtain the ego-motion and the real/virtual plane position from homography. We then apply the Levenburg-Marquardt method to generate a dense projective depth map. We also discuss temporal integration over video sequences. Finally we present the results of applying the homography-based video analysis to motion detection.

1 Introduction

Temporal information redundancy of video sequences allows us to use efficient, incremental methods which perform temporal integration of information for gradual refinement.

Approaches handling 3D scene analysis of video sequences with camera motion can be classified into two categories: algorithms which use 2D transformation or model fitting, and algorithms which use 3D geometry analysis. Video sequences of our interest are taken from a moving airborne platform where the ego-motion is complex and the scene is relatively distant but not necessarily flat; therefore, an integration of 2D and 3D algorithms is more appropriate.

The layered approach [Baker et al., 1998] has advantages in dealing with this kind of scenario, but layer segmentation remains a problem. Approaches of structure from motion are mostly feature-based and cannot provide dense depth maps. The flow-based method [Xiong and Shafer, 1995] recovers dense shape via the Kalman Filter, but image correspondences are required. Combining 3D geometry into 2D constraints is widely used in motion detection and segmentation [Irani and Anandan, 1997, Shashua and Werman, 1995]. The plane plus parallax method contributes a great deal to ego-motion computation [Irani et al., 1994], parallax geometry analysis [Irani and Anandan, 1996] and applications to video indexing [Irani and Anandan, 1998].

Our framework first calculates image homography between consecutive images since the camera-to-scene distance is relatively large and therefore the first-order approximation of the scene can be planar. Section 2 describes three components to achieve robust homography: contrast/brightness adjustment, progressive complexity of transformation and robust estimation. Based on the homography, a camera motion solver is presented in Section 3 to compute ego-motion and plane equation, then optimization can be performed to recover the dense projective depth map of the environment. Temporal integration is performed over video sequences to refine the projective depth. The results of applying the homography-based video analysis to motion detection are discussed in Section 4.

* This work was funded by DARPA under Contract Number DAAB07-97-C-J031

2 Robust Homography

Video sequences from a camera with ego-motions, especially the sequences taken from a moving airborne platform, usually include lighting and environmental changes. Contrast and brightness adjustment is very critical in image registration. Homography between images is based on the assumption that either the scene is planar or the camera is only undergoing rotation and/or zooms; however, many video sequences are taken with no restriction of camera motion and without dominant planes. Therefore, it is necessary to use statistical techniques to obtain robust homography. We incorporate contrast/brightness adjustment and robust estimation into image registration to generate dominant homographies for complex environments.

2.1 Homography with Image Intensity Adjustment

Homography defines the relationship between two images by an eight-parameter perspective transformation:

$$\mathbf{x}' = \begin{bmatrix} u' \\ v' \\ 1 \end{bmatrix} \cong P\mathbf{x} = \begin{bmatrix} p_0 & p_1 & p_2 \\ p_3 & p_4 & p_5 \\ p_6 & p_7 & p_8 \end{bmatrix} \begin{bmatrix} u \\ v \\ 1 \end{bmatrix}$$

where $\mathbf{x} = (u, v, 1)^{\mathrm{T}}$ and $\mathbf{x}' = (u', v', 1)^{\mathrm{T}}$ are homogeneous coordinates, and $\cong$ indicates equality up to scale. Szeliski and Shum [1997] gave a simple solution for the transformation on which we design our registration algorithm.

Due to the difference of viewpoints and change of lighting, video sequences may have different intensity levels from frame to frame. We model the change between images as a linear transformation [Lucas and Kanade, 1981]:

$$I_0(\mathbf{x}) = \alpha\, I_1(\mathbf{x}') + \beta$$

where α stands for contrast change, β for brightness change, $\mathbf{x}$ and $\mathbf{x}'$ for corresponding pixels in two images. Combining this with the general homography computation, we obtain

$$E(D; \alpha, \beta) = \sum_i \left[I_0(\mathbf{x}_i) - \alpha \tilde{I}_1(\mathbf{x}'_i) - \beta \right]^2$$

where $\tilde{I}_1$ is the warped image of I_1 by P, D is the incremental update for P: $(\mathrm{I} + D)P \Rightarrow P$, and $\mathbf{x}' \cong (I + D)P\,\mathbf{x}$ is calculated by updating the transformation. Through this representation, we can minimize the error metric using a symmetric positive definite (SPD) solver such as *Cholesky* decomposition which is time efficient.

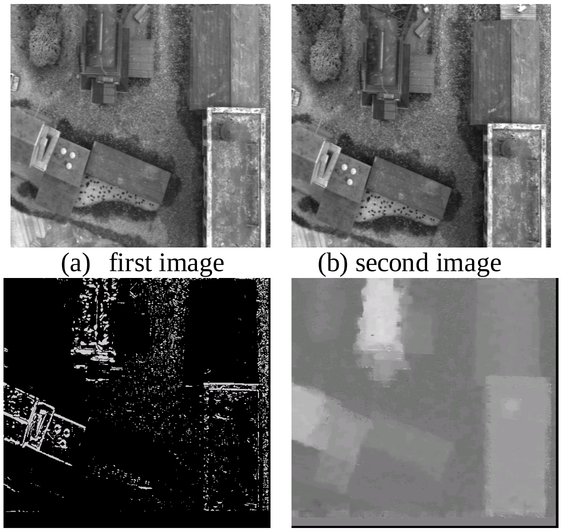
2.2 Progressive Transformation Complexity

Homography is computed hierarchically where estimates from coarser levels of the pyramid are used to initialize the registration at finer levels [Anandan, 1989, Bergen et al., 1992]. To decrease the likelihood of the minimization process converging into local minima, and to improve registration speed, we use different transformations with progressive complexity; translation (2 parameters) at the coarsest level, then scaled rotation plus translation (4 parameters), affine (6 parameters), and perspective (8 parameters). The progressive method improves the robustness and stability of homography computation.

2.3 Robust Estimation

To deal with scenes without dominant planes and/or with a certain percentage of textureless areas, robust estimation is used to compute homography. The random sample consensus paradigm (RANSAC) [Fischler and Bolles, 1981] is an early example of robust estimation. Geometric statistics were also explored in motion problems [Torr and Murray, 1997, Kanatani, 1997]. We apply the RANSAC scheme to homography computation by randomly choosing a small subset of the images to obtain an initial homography solution, i.e., the subset defines a real/virtual plane, and then identifying the outliers, which are the points not lying on the plane. The process is repeated enough times on different subsets and the best solution is the homography which maximizes the number of points lying on the plane. Points which are not identified as outliers are combined to obtain a final homography.

The three components (image intensity adjustment, progressive transformation complexity, and robust estimation) are used in combination to achieve robust homography. Figure 1(a) and (b) give two aerial images taken under different lighting conditions. Robust estimation randomly chooses 20 subsets, each of which is equal to 5 percent of the whole image. Each subset generates a homography. The best homography has the largest support area in the image; this area is used to compute the final homography. In this example, the support area for the final homography consists of the tops of several short buildings rather than the real ground because the ground is not actually flat. White dots in Figure 1(c) show the outliers of the final homography. It can be seen that they correspond to tops of tall buildings (closer than the dominant plane) and part of the ground (farther than the plane).



(a) first image (b) second image

(c) outliers of robust estimation (d) projective depth (darker denotes farther)

Figure 1: Robust homography and projective depth

3 Recovery of Projective Depth

3.1 Projective Depth and Homography

Let $\mathbf{x} = (u, v, 1)^{\mathrm{T}}$ and $\mathbf{x}' = (u', v', 1)^{\mathrm{T}}$ denote homogeneous coordinates of corresponding pixels in two images; the corresponding scene point can be represented by homogeneous coordinate $(u, v, f, w)^{\mathrm{T}}$ in the 3D coordinate system of the first image and $\mathbf{p} = (u/w, v/w, f/w)^{\mathrm{T}}$, where w is the projective depth of point $\mathbf{p}$ [Szeliski, 1996]. $\mathbf{p}'$ denotes the same scene point with respect to the second image coordinate system,

$$\mathbf{p}' = R\,\mathbf{p} + T'$$

where R represents the rotation between the two image coordinate systems and T' represents the 3D translation between the two views expressed in the second image coordinate system. By using

$$V = \begin{bmatrix} f & 0 & 0 \\ 0 & f & 0 \\ 0 & 0 & 1 \end{bmatrix} \quad \text{and} \quad V' = \begin{bmatrix} f' & 0 & 0 \\ 0 & f' & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

to represent the projections of two images, we obtain

$$\begin{aligned} \mathbf{x}' &\cong V'\mathbf{p}' \\ &= V'R\,\mathbf{p} + V'T' \\ &\cong V'RV^{-1}\mathbf{x} + \frac{w}{f}V'T' \end{aligned}$$

Each 3D planar surface can be represented by a 3-vector (a, b, c), which is the scaled normal direction whose size denotes the inverse of the distance to the plane from the origin. If $\mathbf{p}$ is on the plane, that is, if

$$(a, b, c)\ \mathbf{p} = 1$$

we can get

$$(a, b, c)V^{-1}\mathbf{x} = \frac{w}{f}$$

So

$$\begin{aligned} \mathbf{x}' &\cong V'(R + T'(a, b, c))V^{-1}\mathbf{x} \\ &\cong P\,\mathbf{x} \end{aligned}$$

where P is the homography we obtain from the two images.

3.2 Camera Motion Solver

Robust image registration gives an accurate estimation of dominant homography between two images. The support region (non-outliers of RANSAC) corresponds to a real or virtual planar surface in the scene. Given focal lengths (refer to Section 4 for recovery of unknown focal lengths from video sequences), the camera motion and plane equation can be solved directly by the following equation:

$$P \cong V'(R + T'(a, b, c))V^{-1}$$

R can be expressed by Euler angles which have 3 variables, T' and plane distance are up to scale; therefore, they have 5 variables. Since the Euler representation of R is non-linear, the Levenberg-Marquardt method is used to solve the above equation. As the number of variables (8 parameters) is small, the optimization process is rapid.

3.3 Projective Depth Solver

The camera motion solver provides the rotation and translation of two image coordinate systems, that is, we have

$$\mathbf{x}' \cong M\mathbf{x} + w\mathbf{t}$$

where $M = V' R V^{-1}$ and $\mathbf{t} = \frac{1}{f} V' T'$ are known. The Levenberg-Marquardt method is used here to minimize:

$$E(w_i) = \sum_i \left[I_0(\mathbf{x}_i) - \alpha \tilde{I}_1(M\mathbf{x}_i + w_i \mathbf{t}) - \beta \right]^2$$

Assuming that the projective depths of different pixels are independent, we get the diagonal Hessian matrix which makes the optimization process more efficient.

The hierarchical framework used in homography computation is also applied here. To decrease the possibility of converging to local minima and to improve the efficiency, we use patch-based depth recovery and local search. The image is divided into small patches. Each patch shares the same depth while the patch Jacobian is the sum of the Jacobian of each pixel in the patch. When patch displacement exceeds a certain scale, even the multilevel depth recovery fails. To overcome this problem, local search is performed at each patch for subpixel displacement. This displacement is used to solve w_i directly and the solution is incorporated into the optimization as an initial value.

Figure 1(d) gives the result of projective depth recovery from only the two images in Figure 1(a) and (b). The patch size is 2×2 pixels and local search area is 7×7 pixels.

4 Temporal Integration in Video Sequences

A video sequence stores a large amount of redundant information of scenes as the temporal consistency. We use temporal integration over video sequences to refine the projective depth and apply it to motion detection.

4.1 Depth Integration

From each pair of images, we recover the projective depth represented in the first image coordinate system. It is necessary to propagate this depth representation to the second coordinate system so that temporal integration can be performed on the recovered depth. From

$$\mathbf{x}' \cong V' R V^{-1} \mathbf{x} + \frac{w}{f} V' T'$$

we get

$$\mathbf{x} \cong V\, R^{-1}\, V'^{-1} \mathbf{x}' + \frac{w'}{f'} V(-R^{-1}\, T')$$

Representing $\mathbf{x}$ and $\mathbf{x}'$ with scale

$$k' \mathbf{x}' = V' R V^{-1} \mathbf{x} + \frac{w}{f} V' T'$$

$$k\, \mathbf{x} = V\, R^{-1}\, V'^{-1} \mathbf{x}' + \frac{w'}{f'} V(-R^{-1}\, T')$$

we obtain

$$(k'k - 1)\mathbf{x}' = (\frac{wk}{f} - \frac{w'}{f'}) V' T'$$

$V'T'$ is the camera motion which is the same for all pixels; therefore,

$$k'k = 1 \quad and \quad w' = \frac{f'}{f} kw = \frac{f'}{f} \frac{w}{k'}$$

In this way, we represent the depth in the second image coordinate system and then we can refine this depth by the next pair of images consisting of the second and the third images. This process is repeated over the entire video sequence.

4.2 Plane Integration

The first pair of images gives a plane equation from the dominant homography. The plane equation is actually up to scale with the translation parameters. This is the reason why the same scale must be maintained for the same plane in the succeeding pairs in order to refine the current depth. We need to propagate the plane equation representation from the first image coordinate system to the second one.

Let $N = (a, b, c)$ and $N' = (a', b', c')$ denote the equations of the same plane in two coordinate systems. Since they are scaled normal directions,

$$N'^{\mathrm{T}} = \lambda\, R\, N^{\mathrm{T}}$$

where R is the rotation between two coordinate systems and λ is the scale between two normal directions. For point $\mathbf{p} = (x, y, z)^T$ expressed in the first coordinate system, we get

$$N\,\mathbf{p} = 1 \ and \ N'(R\,\mathbf{p} + T') = 1$$

$$\Rightarrow N'\, R\, \mathbf{p} - 1 = -N' T'$$

$$\Rightarrow \lambda\, N\, \mathbf{p} - 1 = -\lambda\, N\, R^T T'$$

$$\Rightarrow 1 - \frac{1}{\lambda} = -N\, R^T T'$$

$$\Rightarrow \lambda = \frac{1}{1 + N\, R^{\mathrm{T}} T'}$$

λ tells the position of the plane in the second coordinate system from propagation so that we can adjust the scale of the next camera motion solver to maintain the plane at the same position.

4.3 Recovery of Focal Length

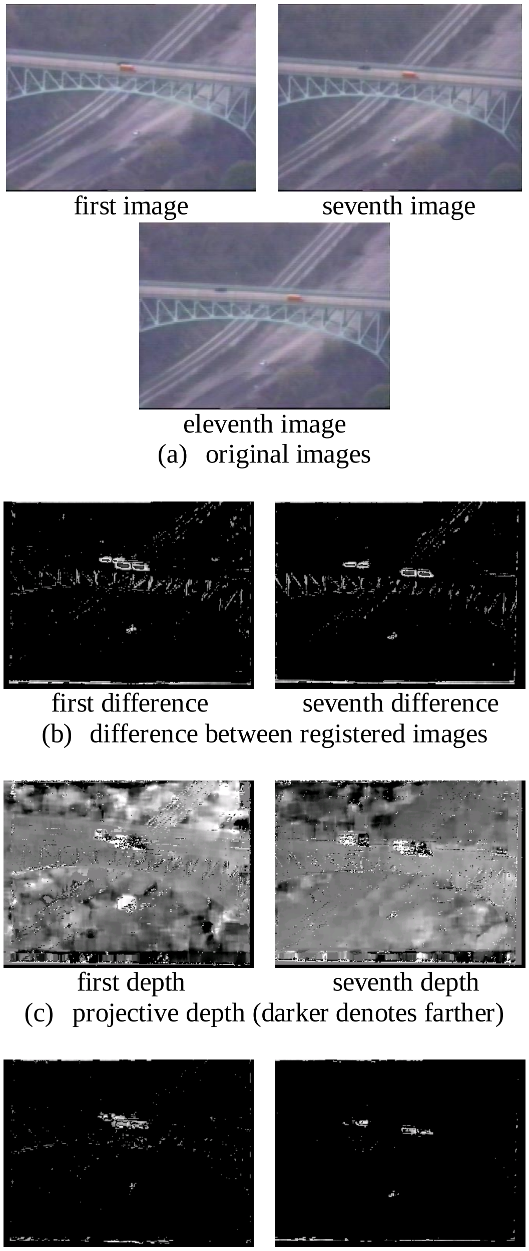
The tutorial [Mohr and Triggs, 1996] summarizes the projective geometry approaches in structure from motion, concluding that when internal parameters are constant three images are enough to recover the Euclidean shape. Pollefeys et al. [1998] demonstrated that if the skew parameter equals zero, even with varying internal parameters three images are sufficient to recover Euclidean shape. In our work, we assume other internal parameters as known except the focal length.

Each homography has 8 parameters which include information of rotation (3 parameters) and translation (3 parameters) of consecutive images. Given the initial values of the first two focal lengths, we can obtain the dominant plane equation from the camera motion solver. The plane equation is propagated to the following images and can then be used to solve focal lengths (2 parameters) from homography in the same way as solving camera motion.

4.4 Application to Motion Detection

Detecting moving objects in a video sequence taken from moving camera is an important task in video sequence analysis. Some algorithms work well in 2D situations when the scene can be approximated by a flat surface and/or when the camera is undergoing only rotations and zooms; some apply to the scene when large depth variations are present. [Irani and Anandan, 1997] discusses a unified approach handling both 2D and 3D scenes. Our goal is motion detection in aerial images while the camera experiences complex ego-motion and the scene can neither be classified as flat surface nor provide significant depth variations.

Figure 2(a) shows three images in the video sequence provided by the Video Surveillance and Monitoring (VSAM) project of CMU. The sequence was taken from an airplane flying above a bridge; two cars were moving on the bridge and one car was moving on the road which is far below the bridge. We first obtained homography to register consecutive images in the video sequence. Figure 2(b) gives the difference images between consecutive registered images. White dots indicate differences which are actually the outliers of homographies; we can observe that the ground below the bridge was selected as dominant plane by robust estimation. Also, we can see that both motion (moving cars) and parallax (the bridge) appear in the difference images. Based on the homographies, we recovered projective depth by


first image seventh image

eleventh image

(a) original images

first difference seventh difference

(b) difference between registered images

first depth seventh depth

(c) projective depth (darker denotes farther)

first difference after depth compensation seventh difference after depth compensation

(d) motion detection (difference between registered images after depth compensation)

Figure 2: Motion Detection

temporal integration over 7 images and use that to register consecutive images again. Figure 2(c) shows the recovered depth. It can be seen that the projective depth was improved over sequences; the projective depth for the seventh image shows the scene structure including the bridge in front and the road along the gully. New difference images (Figure 2(d)) were generated between registered images with depth compensation. We can see that differences due to the depth are cleaned up and white dots represent the motion only. Cars on the bridge and on the road below are detected and tracked correctly. However, in a situation where motion of the object always satisfies the epipolar constraints, the object is classified as a stationary rigid body.

5 Conclusion

We have presented a framework for homography based projective depth recovery and its application to motion detection. We described a robust homography algorithm which incorporates image contrast/brightness adjustment and robust estimation into image registration. Based on the homography between two images, our camera motion solver gives the solution of ego-motion and plane equation; the solution is refined to generate projective depth for each pixel by the Levenburg-Marquardt method. We also discussed temporal integration of projective depth recovery and its application to motion detection.

The encouraging temporal integration results motivate us to expand this work to include spatial integration as well. Other application tasks such as 3D mosaicking, background model recovery and video editing are promising areas to explore.

Acknowledgements

We would like to thank Bob Collins, Alan Lipton, Teck Khim and Wei Hua for fruitful discussions.

References

[Anandan, 1989] P. Anandan. “A Computational Framework and an Algorithm for the Measurement of Visual Motion,” *IJCV* 2(3), pp. 283-310, 1989.

[Baker et al., 1998] S. Baker, R. Szeliski and P. Anandan. “A Layered Approach to Stereo Reconstruction,” *Proc. of CVPR’98*, pp. 434-441, 1998.

[Bergen et al., 1992] J. R. Bergen, P. Anandan, K. J. Hanna and R. Hingorani. “Hierarchical Model-based Motion Estimation,” *Proc. of ECCV’92*, pp. 237-252, 1992.

[Fischler and Bolles, 1981] M.A. Fischler and R.C. Bolles. “Random Sample Consensus: A Paradigm for Model Fitting with Application to Image Analysis and Automated Cartography,” *Commun. Assoc. Comp. Mach*, vol. 24, 1981.

[Irani et al., 1994] M.Irani, B.Rousso and S. Peleg. “Recovery of Ego-Motion Using Image Stabilization,” *Proc. of CVPR’94*, pp. 454-460, 1994.

[Irani and Anandan, 1996] M. Irani, P. Anandan. “Parallax Geometry of Pairs of Points for 3D Scene Analysis,” *Proc. of ECCV*, 1996.

[Irani and Anandan, 1997] M. Irani and P. Anandan. “A Unified Approach to Moving Object Detection in 2D and 3D Scenes,” *PAMI*(20), No. 6, pp. 577-589, June 1997.

[Irani and Anandan, 1998] M. Irani and P. Anandan, “Video Indexing Based on Mosaic Representations,” *Proc. of IEEE* (86), No. 5, pp. 905-921, May 1998.

[Kanatani, 1997], K. Kanatani. *Introduction to Statistical Optimization for Geometric Computation* (*Lecture Note*), 1997.

[Lucas and Kanade, 1981] B. D. Lucas, T. Kanade. “An Iterative Image Registration Technique with an Application to Stereo Vision,” *Proc. of IUW*, pp. 121-130, 1981.

[Mohr and Triggs, 1996] R. Mohr and B. Triggs. *Projective Geometry for Image Analysis*. A Tutorial given at ISPRS, 1996.

[Pollefeys et al., 1998] M. Pollefeys, R. Koch and L. V. Gool. “Self-Calibration and Metric Reconstruction in spite of Varying and Unknown Internal Camera Parameters,” *Proc. of ICCV*, pp. 90-95, 1998.

[Shashua and Werman, 1995] A. Shashua, M. Werman. “Trilinearity of Three Perspective Views and its Associated Tensor,” *Proc. of ICCV’95*, pp. 920-925, 1995.

[Szeliski, 1996] R. Szeliski. “Video Mosaics for Virtual Environments,” *IEEE Computer Graphics and Applications*, pp. 22-30, March 1996.

[Szeliski and Shum, 1997] R. Szeliski and H. -Y. Shum. “Creating Full View Panormic Image Mosaics and Texture-mapped Models,” *SIGGRAPH’97*, pp. 251-258, August 1997.

[Torr and Murray, 1997] P.H.S. Torr, D.W. Murray. “The Development and Comparison of Robust Methods for Estimating the Fundamental Matrix,” *IJCV*(24), No. 3, pp. 271-300, 1997.

[Xiong and Shafer, 1995] Y. Xiong and S. A. Shafer. “Dense Structure From A Dense Optical Flow Sequence,” *Proc. of ISCV’95*, 1995.