# HIERARCHICAL REGION-BASED IMAGE REGISTRATION IN SCALE SPACE

*Nan Jiang, Jennie Si, Glen Abousleman*

Department of Electrical Engineering
Arizona State University
Tempe, AZ 85281

**ABSTRACT**

This paper presents a new method for image registration for real natural scenes. The method is based on the observation that most natural scenes are actually 3-D. If the assumption of weak perspective is violated, the error in image registration induced by parallax is increased, leading artifacts or blurs in image mosaic. Our method first applies the affine-invariant point detector in scale space. After clustering feature point pairs, an initial global transformation is formed based on majority correspondence. The global transformation is evaluated in each region at certain scale determined by inliers. The global model is optimized for local registration by minimizing Least Square Error. This method is more robust than standard image registration algorithms on images subject to uncalibrated camera motion.

## 1. INTRODUCTION

The common approach to video mosaic is sequential registration [1]. Pairwise images are registered, i.e. the second image to the first image, the third to the second and so on. The images are aligned to a common reference frame based on the concatenation property of homographies. However, registration error also propagates along the sequence, and eventually the accumulated error becomes visible, bringing in artifacts or blurring in the panorama. There are many sources of registration error, such as the white noise, moving objects, lens distortion, parallax and occlusion. Since feature-based algorithms are based on the sparse feature points, the disturbance from above sources could misallocate the feature points and the parametric transformation. Various approaches have been proposed to deal with the error accumulation problem [2]. Most algorithms focus on improving the registration between two images. Harris-DoG [3] [4] and Harris-Laplacian [5] [6] detector are stable across affine change. Global adjustment, such as bundle adjustment [7], optimizes over all the homographies. Methods evenly distribute the error across the sequence as posterior processing are described in [8]. Uyttendaele [9] proposed a local adjustment method to determine the intensity value on overlay areas. [10] proposed a method for detecting and tracing incoherent local motion in dynamic scenes.

This work was supported by General Dynamics C4 Systems, and in part by the NSF under grant ECS-0002098.

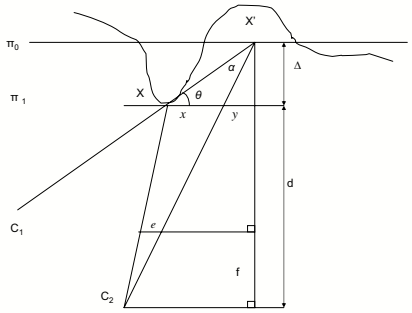


**Fig. 1**. Geometric modeling of parallax in 3-D scene.

Conventional image registration uses planar projective transformation, which is based on the assumption that the distance of the scene from the camera is much greater than the camera motion, such that the scene can be considered on a common plane surface approximately. However, in real cases when the distance of scene are comparable to the focal length, the depths vary tremendously, or the distances between camera centers are wide. The parallax effects caused by the three dimensional nature of the scene is no longer negligible. In this paper we propose an approach which allows optimization of global homography in each local region to compensate for the parallax.

The remainder of this paper is organized as follows. Section 2 briefly outlines the geometric model of parallax growing as viewing angles increases. In section 3 the coarse-to-fine image registration is described and section 4 presents the experimental results.

## 2. GEOMETRICAL MODELING

We consider a moving camera, taking images of a scene. The images taken at time $j$ is defined as $I_j$. The homography between two consecutive images is a $3 \times 3$ homogeneous matrix $H_{j+1}^{j}$ [11]. Let $\mathbf{x}'$ be a point in image $I_j$ and $\mathbf{x}$ be the corresponding point in image $I_{j+1}$, the spatial relationship between these two points is $\mathbf{x}' = H_{j+1}^{j}\mathbf{x}$.

The error inherent from 3-D scene is illustrated in Figure 1. $C_1$ and $C_2$ are two camera centers in the same scene but from varying angles, where $C_1$ is the reference view. The curve denotes the real surface of the scene. Plane $\pi_0$ is the common plane for the scene. $C_1$ reprojects a 3D point $X$ on $\pi_1$, onto the common surface plane $\pi_0$, to form a hypothetic 3D point $X'$. The error term $e$ reflects the residual of viewing $X$ from $C_2$:

$$\begin{aligned} e &= \frac{f}{d}x \\ &= \frac{f}{d}\frac{l\sin\alpha}{\sin(\theta+\alpha)} \\ &= \frac{f}{d}\frac{\frac{\Delta}{\sin\theta}\sin\alpha}{\sin\theta\cos\alpha+\cos\theta\sin\alpha} \\ &= \frac{f\Delta}{d(\frac{\sin^2\theta}{\tan\alpha}+\cos\theta\sin\theta)} \end{aligned} \quad (1)$$

Since $C_1$ is the reference view, angle $\theta$ is constant when camera center $C_2$ varies. The depth difference $\Delta$, focal length $f$ and scene depth $d$ are fixed terms as well. The constraint is that $\alpha$ is less than $\pi/2$, which is the usual case. According to Equation (1), $\alpha \uparrow \Rightarrow \quad \tan\alpha \uparrow \Rightarrow \quad e \uparrow$, which means the parallax increases as the viewing angle with respect to the reference view grows along the image sequence. And it is easy to see from Equation (1) that parallax is proportional to the range of depth difference, and the ratio between camera focal length and the distance from the scene to the camera.

## 3. APPROACH

### 3.1. Global Registration

We use a modified version of Lowe's SIFT descriptor [12] to retrieve affine-invariant interest points. Local extrema in DOG (Difference of Gaussian) filtered images are selected as interest points. The scale $\sigma$ at which DOG response has a local extremum is selected as well. The Harris corner condition is examined in both space and scale for each interest point. Points with edge response are eliminated. Edge response means the corner response function is less than a threshold. We make cornerness threshold adaptive to the spatial density of interest points in each region, where the regions are the output of a parallel process of segmentation. If the density of interest points in one region is above the global density of interest points, the threshold is increased to eliminate more points and vice versa. The density of interest points in region $E_i$ at scale $\sigma$ is defined as

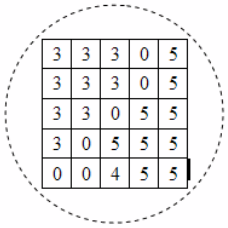


**Fig. 2**. A sample fifth order neighbourhood around a boundary pixel. Boundaries are assigned to one or more local regions.

$$D(E_i,\sigma) = \frac{\#\ interest\ points\ in\ E_i}{\#\ pixels\ in\ E_i} \quad (2)$$

Two interest points are defined as a potentially corresponding pair if the Mahalanobis distance between their local feature vectors is below a predetermined threshold [5].

$$d_M(\mathbf{m},\mathbf{m}') = \sqrt{(\mathbf{V_m}-\mathbf{V'_m})^T\Lambda^{-1}(\mathbf{V_m}-\mathbf{V'_m})} \quad (3)$$

where $\mathbf{V_m}$ and $\mathbf{V'_m}$ are local descriptors associated with point $\mathbf{m}$ and $\mathbf{m}'$ respectively, and $\Lambda$ is the covariance matrix formed from $\mathbf{V_m}$ and $\mathbf{V'_m}$. The Mahalanobis distance is the weighted form of the Euclidean distance, as a consequence it is rotation-invariant compared to the Euclidean distance in point matching.

Clustering all point-to-point correspondences forms a set of initial similarity transforms. The robust matching algorithm, RANdom SAmple Consensus, is used to obtain the initial transforms. RANSAC has the advantage of insensitive to outliers.

### 3.2. Segmentation in Scale Space

For the image segmentation, the image is segmented into homogeneous labeled regions in term of gray intensity using watershed segmentation algorithm. During the procedure of robust image matching, for a point pair $(\mathbf{x}_m, \mathbf{x}_n)$, the scale pair, $(\sigma_{j,m}, \sigma_{j+1,n})$, is recorded. $\sigma_{j,m}$ is the scale of interest point $\mathbf{x}_m$, and $\sigma_{j+1,n}$ is the scale of interest point $\mathbf{x}_n$. K-means clustering classifies the scale pairs, and the centroid of the biggest cluster, say $(\sigma_j, \sigma_{j+1})$, is chosen as the global scale pair between image $I_j$ and image $I_{j+1}$. The image at scale level $\sigma_{j+1}$ is selected from image pyramid $L(j+1, \sigma_{j+1})$

for segmentation. The regions are compared and aligned to $L(j, \sigma_j)$ in local adjustment. Here $L(j, \sigma) = G(j, \sigma) * I_j$.

Boundary pixels between regions are not assigned to any region in segmentation. But they are also taken care of in our experiments. The reason is that these pixels normally locate in high activity areas. Misalignments of these pixels are more visible than the pixels in flat areas. Therefore the local adjustment should include boundaries into consideration. The boundary pixel is assigned to the region(s) according to its fifth order neighborhood system:

$$\begin{cases} \mathbf{x} \in E^k_{j,\sigma}, & if N_k > N_m, m \neq k \\ \mathbf{x} \in E^k_{j,\sigma}, E^m_{j,\sigma}, & if N_k = N_m, m \neq k \end{cases} \quad (4)$$

where $N_k$ is the number of pixel belonging to region $k$ over the $5 \times 5$ neighouring window around $\mathbf{x}$, where $j$ is the index of image sequence and $E^k_{j,\sigma}$ is the region at $L(j, \sigma)$ labeled as $k$. Similarily $N_m$ indicates the number of pixels labeled as region $m$. As depicted in Figure 2 the boundary point in the center of the window is assigned to both region labeled as 3 and region 5. Therefore the regions may overlap over boundaries.

### 3.3. Local Adjustment

In the local adjustment step the global homography $H^j_{j+1}$ is applied to region $E^k_{j+1,\sigma_{j+1}}$ for evaluation. The reliability of an initial homograhpy is indicated by its residual error, $e^2_{j+1,k}$, which is denoted as follows:

$$e^2_{j+1,k} = \frac{1}{N_k} \sum_{\mathbf{x}} [L_{j,\sigma_j}(\mathbf{x}') - L_{j+1,\sigma_{j+1}}(\mathbf{x})]^2, \quad (5)$$
$$where \;\; \mathbf{x}' = H^j_{j+1}\mathbf{x}, \; \mathbf{x} \in E^k_{j+1,\sigma_{j+1}}$$

where $N_k$ is the number of pixels in region $E^k_{j+1,\sigma_{j+1}}$; $L_{j+1,\sigma_{j+1}}(\mathbf{x})$ is the gray intensity at point $\mathbf{x}$ in image $j+1$ at scale $\sigma_{j+1}$; $\mathbf{x}'$ is the location of the counterpart of $\mathbf{x}$ in image $L_{j,\sigma_j}$. If the residual is greater than a pre-defined threshold this region is considered incoherent with the global homography. Usually the registration residual is the consequence of inconsistent occlusions, parallax, or moving objects. The goal of local alignment is to find a homography such that,

$$H^j_{E^k_{j+1,\sigma_{j+1}}} = \arg\min e^2_{j+1,k}. \quad (6)$$

We use the Lervenberg-Marquardt method (LM) [13], which is an approximate second order optimization method for solving the minimum of a nonlinear function. This method uses the whole set of pixels in the overlapping area. It directly minimizes the discrepancy in the intensities between region pairs.

### 3.4. Iterative Refinement

After local adjustment for region $E_i$ reaching an optimal point, the local homography is applied to the overall image. If the global registration calculated by Equation (1) decreases, local homography replaces the global model; if the error increases, local homography is discarded and local adjustment for next region is performed. The above steps iterate until all the regions are processed. It is easy to see this is a linear regression process.

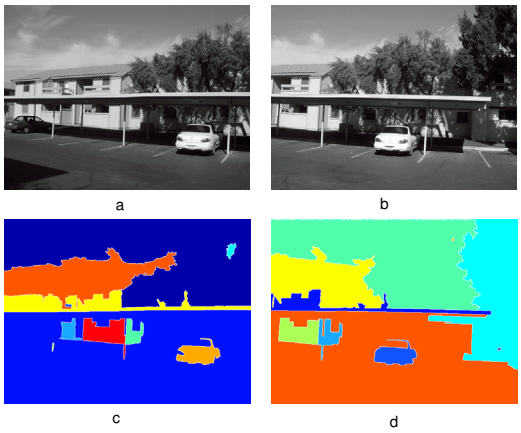


**Fig. 3**. a-b, two images with slight scale factor used for registration; c d, output of watershed segmentation. Regions of diffrent scene depths are seperated.

## 4. EXPERIMENTAL RESULTS

**Fig. 4**. Example mosaic result taking Figure 3.a and Figure 3.b for input

The proposed algorithm is applied to two images as shown

in Figure 3.a and Figure3.b. Note that Figure 3.b has a slight scale factor compared to reference image 3.a. Image matching between the two images are wide-baselined. The depths of the scene, range from close parking roof, to trees and buildings, until infinity into the sky. Figure 3.c and Figure 3.d shows the output of watershed segmentation. Regions with different depths are separated, and compensated respectively in local adjustment.

Figure 4 shows the mosaic result after global registration followed by local adjustment in scale space. The frame size is $640 \times 480$. The camera is uncalibrated nor providing any prior information. Registration error without local adjustment is 22.56, while Registration with local adjustment is 19.42. Part of the error comes from different illumination conditions in the input images.

## 5. CONCLUSIONS

Hierarchical image registration uses local information to refine the initial global homography in each region. This approach is useful if the scene is not precisely planar, which is the common case. The parallax is compensated to result in a higher quality mosaic image. In local region adjustment, since the regions are denser than sparse feature points the algorithm is more robust to Gaussian distributed noise and occlusions. In addition, feature points are usually selected from gradient maxima, which means the regions with rich texture contributes more to the global transformation. Our method balances the spatial distribution of feature points across regions, hence avoids ill-posed cases that those textured regions do not agree with the true global model. The global model is obtained first as the prior information for region adjustment. Otherwise the local registration may be stuck at some local optimal point.

## 6. REFERENCES


[1] Michal Irani, P. Anandan, and Steve Hsu, "Mosaic based representations of video sequences and their applications," in *Proc. Fifth Int. Conf. on Computer Vision*, 1995, pp. 605–611.

[2] Heung-Yeung Shum and Richard Szeliski, "Construction and refinement of panoramic mosaics with global and local alignment," IEEE Int'l Conf. Computer Vision, 1998, pp. 953–958.

[3] G. David Lowe, "Object recognition from local scale-invariant features," in *Proceedings of the International Conference on Computer Vision*, 1999, pp. 1150–1157.

[4] Y. Dufournaud, C. Schmid, and R. Horaud, "Image matching with scale adjustment," vol. 93, no. 2, pp. 175–194, February 2004.

[5] K. Mikolajczyk and C. Schmid, "Indexing based on scale invariant interest points," 2001, pp. I: 525–531.

[6] K. Mikolajczyk and C. Schmid, "A performance evaluation of local descriptors," in *Proceedings of Conference on Computer Vision and Pattern Recognition CVPR 2003*, USA, June 2003.

[7] H. Shum and R. Szeliski, "Construction and refinement of panoramic mosaics with global and local alignment," IEEE Int'l Conf. Computer Vision, 1998.

[8] M.G. Gonzalez, P. Holifield, and M. Varley, "Improved video mosaic construction by accumulated alignment error distribution," in *BMVC98*, 1998, pp. xx–yy.

[9] A. Eden M. Uyttendaele and R. Szeliski, "Eliminating ghosting and exposure artifacts in image mosaics," IEEE Computer Society Conference on Computer Vision and Pattern Recognition, December 2001, pp. 509–516.

[10] J. Davis, "Mosaics of scenes with moving objects," CVPR, 1998, pp. 97–100.

[11] Zisserman A. Hartley, R., *Multiple View Geometry*, Cambridge University Press, UK, 2004.

[12] G. David Lowe, "Distinctive image feature from scale-invariant keypoints," *Int'l Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.

[13] TEUKOLSKY S. A. VETTERLING S. A. PRESS, W. H. and B. P. FLANNERY, *Numerical Recipe In C*, Cambridge University Press, UK, 1992.