Depth Based Image Registration via Geometric Segmentation

Bing Han, Chris Paulson, and Dapeng Wu
Department of Electrical and Computer Engineering
University of Florida Gainesville, FL 32611
Correspondence author: Prof. Dapeng Wu, wu@ece.ufl.edu, http://www.wu.ece.ufl.edu

Abstract

Image registration is a fundamental task in computer vision and it significantly contributes to high-level computer vision and benefits numerous practical applications. Although there are already a lot of image registration techniques existing in literature, there is still a significant amount of research to be conducted because there are a lot of issues that need to be solved such as the parallax problem. The traditional image registration algorithms suffer from the parallax problem due to their underling assumption that the scene can be regarded approximately planar which is not satisfied when large depth variations exist in the images with high-rise objects. To address the parallax problem, we present a new strategy for 2D image registration by leveraging the depth information from 3D image reconstruction. The novel idea is to recover the depth in the image region with high-rise objects to build accurate transform function for image registration. We segment the 3D space in several separate regions and use surface fitting algorithms to estimate the 3D dense depth map. In order to segment the space geometrically, we propose a non-linear deterministic annealing algorithm for space partitioning. From the experimental results, the new method is able to mitigate the parallax problem and achieve robust image registration results. Our algorithm is attractive to numerous practical applications.

Index Terms

3D reconstruction, image registration, depth estimation, parallax problem, geometric segmentation

I. Introduction

Image registration is a fundamental task in image processing and computer vision which matches two or more images taken at different times and different viewpoints, by geometrically aligning reference and sensed images. There has been a broad range of techniques developed over the years in literature. A comprehensive survey of image registration methods was published in 1992 by Brown [1], including many classic methods still in use. Due to the rapid development of image acquisition devices, more image registration techniques emerged afterwards and were covered in another survey published in 2003 [2].

Different applications due to distinct image acquisition require different image registration techniques. In general, manners of the image acquisition can be divided into three main groups:

- *Different viewpoints (multiview analysis)*. Images of the same scene are acquired from different viewpoints. The aim is to gain a larger 2D view or a 3D representation of the scanned scene.
- *Different times*. Images of the same scene are acquired at different times, often on regular basis, and possibly under different conditions. The aim is to find and evaluate changes in the scene which appeared between the consecutive image acquisitions.
- *Different sensors*. Images of the same scene are acquired by different sensors. The aim is to integrate the information obtained from different source streams to gain more complex and detailed scene representation.

The prevailing image registration methods, such as Davis and Keck's algorithm [3], [4], assume all the feature points are coplanar and build a homography transform matrix to do registration. The advantage is that they have low computational cost and can handle planar scenes conveniently; however, with the assumption that the scenes are approximately planar, they are inappropriate in the registration applications

when the images have large depth variation due to the high-rise objects, known as the parallax problem. Parallax is an apparent displacement of difference of orientation of an object viewed along two different lines of sight, and is measured by the angle or semi-angle of inclination between those two lines. Nearby objects have a larger parallax than further objects when observed from different positions. Therefore, as the viewpoint moves side to side, the objects in the distance appear to move slower than the objects close to camera.

In this paper, we propose a depth based image registration algorithm by leveraging the depth information. Our method can mitigate the parallax problem caused by high-rise scenes in the images by building accurate transform function between corresponding feature points in multiple images. Given an image sequence, we first select a number of feature points and then match the features in all images. Then we estimate the depth of each feature point from feature correspondences. With the depth information, we can project the image in 3D instead of using a homography transform. Further more, fast and robust image registration algorithm can be achieved by combining the traditional image registration algorithms and depth based image registration method proposed in this paper. The idea is that we first compute the 3D structure of a sparse feature points set and then divide the scene geometrically into several approximately planar regions. For each region, we can perform a depth based image registration. Accordingly, robust image registration is achieved.

The remainder of this paper is organized as follows. We present the system scheme for 2D image registration in Section II. Section III reviews the 3D reconstruction algorithm we used in our new method. In Section IV, we describe how to use 3D depth information for 2D image registration and propose a non-linear deterministic annealing algorithm for space partitioning. Section V presents the experimental results and we compare our algorithm with Davis and Keck's algorithm on the same test video sequence. We conclude this paper in Section VI.

II. The scheme of the new 2D image registration system

Due to the diversity of images to be registered and various types of degradations, it is impossible to design a universal method applicable to all registration tasks. Every method should take into account not only the assumed type of geometric deformation between the images but also the radiometric deformations and noise corruption, required registration accuracy and application-dependent data characteristics. Nevertheless, the majority of the registration methods consists of the following four steps: feature detection, feature matching, transform model estimation, image resampling and transformation. Although they may differ in some specific part, most 3D reconstruction approaches are generally based on the same pipeline. The pipeline is given in Fig. 1.

A widely used feature detection method is corner detection. Kitchen and Rosenfeld [5] proposed to exploit the second-order partial derivatives of the image function for corner detection. Dreschler and Nagel [6] searched for the local extrema of the Gaussian curvature. However, corner detectors based on the second-order derivatives of the image function are sensitive to noise. Thus Forstner [7] developed a more robust, although time consuming, corner detector, which is based on the first-order derivatives only. The reputable Harris detector [8] also uses first-order derivatives for corner detection. Feature matching includes area-based matching and feature-based matching. Classical area-based method is cross-correlation (CC) [9] exploit for matching image intensities directly. For feature-based matching, Goshtasby [10] described the registration based on the graph matching algorithm. Clustering technique, presented by Stockman et al. [11], tries to match points connected by abstract edges or line segments. After the feature correspondence has been established the mapping function is constructed. The mapping function should transform the sensed image to overlay it over the reference image. Finally interpolation methods such as nearest neighbor function, bilinear, and bicubic functions are applied to the output of the registered images.

In our new image registration system, we use a 3D model instead of 2D motion model used in existing works. The system scheme is slightly different from the previous one. We give the new scheme in Fig. 2.

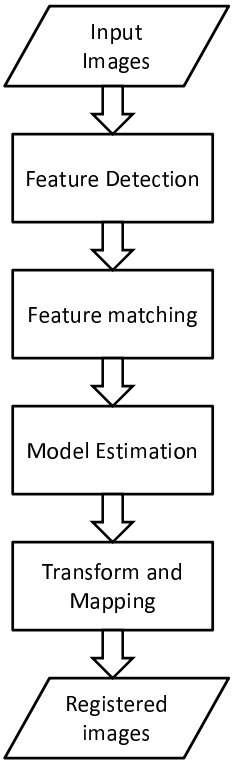

Fig. 1. The pipeline for 2D image registration system.

In the new system scheme, we first apply 3D reconstruction to the input images and recover the 3D geometric structure of the scene in the images. The 3D model is more accurate compared to the 2D motion models estimated in the previous works. Then we segment the 3D Euclidean space geometrically into several separate regions. Each region could be modeled by a linear plane. With the segmentation, we can estimate the 3D depth for every pixel in each region and recover the dense structure of the scene. The 3D dense structure enables the pixel by pixel mapping of the input images. We describe the 3D reconstruction algorithm in Section III. In Section IV, we present the geometric segmentation and depth based mapping in 3D, and also propose a non-linear deterministic annealing algorithm for space partitioning.

III. 3D RECONSTRUCTION FROM VIDEO SEQUENCES

Here, we simply review the 3D reconstruction algorithm described in Ma et. al's book [12]. When developing a stereo vision algorithm for registration, the requirements for accuracy vary from those of standard stereo algorithms used for 3D reconstruction. For example, a multi-pixel disparity error in an area of low texture, such as a white wall, will result in significantly less intensity error in the registered image than the same disparity error in a highly textured area. In particular, edges and straight lines in the scene need to be rendered correctly.

The 3D reconstruction algorithm is implemented using the following steps. First, geometric features are detected automatically in each individual images. Secondly, feature correspondence is established across all the images. Then the camera motion is retrieved and the camera is calibrated. Finally the Euclidean structure of the scene is recovered.

A. Feature selection

The first step in 3D reconstruction is to select candidate features in all images for tracking across different views. Ma et al. [12] use point feature in reconstruction which is measured by Harris' criterion,

$$C(\mathbf{x}) = \det(G) + k \times \text{trace}^2(G) \tag{1}$$

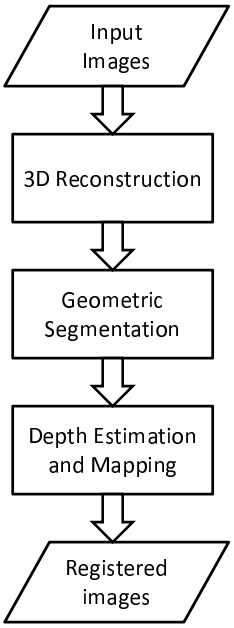

Fig. 2. The new image registration system scheme.

where $\mathbf{x} = [x, y]^T$ is a candidate feature, $C(\mathbf{x})$ is the quality of the feature, k is a pre-chosen constant parameter and G is a 2×2 matrix that depends on $\mathbf{x}$, given by

$$G = \begin{bmatrix} \sum_{W(\mathbf{x})} I_x^2 & \sum_{W(\mathbf{x})} I_x I_y \\ \sum_{W(\mathbf{x})} I_x I_y & \sum_{W(\mathbf{x})} I_y^2 \end{bmatrix} \tag{2}$$

where $W(\mathbf{x})$ is a rectangular window centered at $\mathbf{x}$ and I_x and I_y are the gradients along the x and y directions which can be obtained by convolving the image I with the derivatives of a pair of Gaussian filters. The size of the window can be decided by the designer, for example 7×7. If $C(\mathbf{x})$ exceeds a certain threshold, then the point $\mathbf{x}$ is selected as a candidate point feature.

B. Feature correspondence

Once the candidate point features are selected, the next step is to match them across all the images. In this subsection, we use a simple feature tracking algorithm based on a translational model.

We use the sum of squared differences (SSD) [13] as the measurement of the similarity of two point features. Then the correspondence problem becomes looking for the displacement $\mathbf{d}$ that satisfies the following optimization problem:

$$\min_{\mathbf{d}} \doteq \sum_{\mathbf{x} \in W(\mathbf{x})} \left[I_2(\mathbf{x} + \mathbf{d}) - I_1(\mathbf{x}) \right]^2 \tag{3}$$

where $\mathbf{d}$ is the displacement of a point feature of coordinates $\mathbf{x}$ between two consecutive frames I_1 and I_2. Lucas and Kanade also give the close form solution of 3

$$\mathbf{d} = -G^{-1}\mathbf{b} \tag{4}$$

where

$$\mathbf{b} \doteq \begin{bmatrix} \sum_{W_{(\mathbf{x})}} I_x I_t \\ \sum_{W_{(\mathbf{x})}} I_y I_t \end{bmatrix} \tag{5}$$

G is the same matrix we used to compute the quality of the candidate point feature in Eq. 1, and $I_t \doteq I_2 - I_1$.

TABLE I
EIGHT-POINT ALGORITHM

Given a set of initial point feature correspondences expressed in pixel coordinates $(\mathbf{x'}^j_1, \mathbf{x'}^j_2)$ for $j = 1, 2, ..., n$:
• **A first approximation of the fundamental matrix:** Construct the matrix $\chi \in \mathbb{R}^{n\times 9}$ from the transformed correspondences $\tilde{\mathbf{x}}^j_1 \doteq [\tilde{x}^j_1, \tilde{y}^j_1, 1]^T$ and $\tilde{\mathbf{x}}^j_2 \doteq [\tilde{x}^j_2, \tilde{y}^j_2, 1]^T$, where the jth row of χ is given by $[\tilde{x}^j_1\tilde{x}^j_2, \tilde{x}^j_1\tilde{y}^j_2, \tilde{x}^j_1, \tilde{y}^j_1\tilde{x}^j_2, \tilde{y}^j_1\tilde{y}^j_2, \tilde{y}^j_1, \tilde{x}^j_2, \tilde{y}^j_2, 1]^T \in \mathbb{R}^9$. Find the vector $F^s \in \mathbb{R}^9$ of unit length such that $||\chi F^s||$ is minimized as follows: Compute the singular value decomposition (SVD) of $\chi = U\Sigma V^T$ and define F^s to be the ninth column of V. Unstack the nine elements of F^s into a square 3×3 matrix $\tilde{F}$.
• **Imposing the rank-2 constraint:** Compute the SVD of the matrix F recovered from data to be $\tilde{F} = U_F diag\{\sigma_1, \sigma_2, \sigma_3\}V^T_F$. Impose the rank-2 constraint by letting $\sigma_3 = 0$ and reset the fundamental matrix to be $F = U_F diag\{\sigma_1, \sigma_2, 0\}V^T_F$.

C. Estimation of camera motion parameters

In this subsection, we recover the projective structure of the scene from the established feature correspondence. We will follow the notation used in Ma et al.'s book [12]. For the detail of the proof of this algorithm, please refer to the reference.

The reconstruction algorithm is based on a perspective projection model with a pinhole camera. Suppose we have a generic point $p \in \mathbb{E}^3$ with coordinates $\mathbf{X} = [X, Y, Z, 1]^T$ relative to a world coordinate frame. Given two frames of one scene which is related by a motion $g = (R, T)$, the two image projection point $\mathbf{x}_1$ and $\mathbf{x}_2$ are related as follows:

$$\lambda_1\mathbf{x}'_1 = \Pi_1\mathbf{X}_p, \quad \lambda_2\mathbf{x}'_2 = \Pi_2\mathbf{X}_p \tag{6}$$

where $\mathbf{x}' = [x, y, 1]^T$ is measured in pixels, λ_1 and λ_2 are the depth scale of $\mathbf{x}_1$ and $\mathbf{x}_2$, $\Pi_1 = [K, 0]$ and $\Pi_2 = [KR, KT]$ are the camera projection matrices and K is the camera calibration matrix. In order to estimate λ_1, λ_2, Π_1 and Π_2, we need to introduce the epipolar constraint. From Eq. 6, we have

$$\mathbf{x'}^T_2 K^{-T}\hat{T}RK^{-1}\mathbf{x}'_1 = 0 \tag{7}$$

The fundamental matrix is defined as:

$$F \doteq K^{-T}\hat{T}RK^{-1} \tag{8}$$

With the above model, we could estimate the fundamental matrix F via the Eight-point algorithm[12]. Then we could decompose the fundamental matrix to recover the projection matrices Π_1 and Π_2 and the 3D structure. We only give the solution here by canonical decomposition:

$$\Pi_1 p = [I, 0], \Pi_2 p = [(\widehat{T'})^T F, T'], \lambda_1\mathbf{x}'_1 = \mathbf{X}_p, \lambda_2\mathbf{x}'_2 = (\widehat{T'})^T F\mathbf{X}_p + T' \tag{9}$$

D. Depth estimation

The Euclidean structure $\mathbf{X}_e$ is related to the projective reconstruction $\mathbf{X}_p$ by a linear transform $H \in \mathbb{R}^{4\times 4}$,

$$\Pi_{ip} \sim \Pi_{ie}H^{-1}, \mathbf{X}_p \sim H\mathbf{X}_e, i = 1, 2, ..., m \tag{10}$$

where $\sim$ means equality up to a scale factor and

$$H = \begin{bmatrix} K_1 & 0 \\ -\nu^T K_1 & 1 \end{bmatrix} \in \mathbb{R}^{4\times 4} \tag{11}$$

With the assumption that K is constant, we could estimate the unknowns K and ν with a gradient decent optimization algorithm. In order to obtain a unique solution, we also assume that the scene is generic and the camera motion is rich enough.

Fig. 3 shows the first frame and the 88th frame of the test video sequence 'oldhousing'. In our experiment, we will register all the frames in the video sequence to the first frame. Fig. 4 show the selected feature points on the first frame which are used for camera pose estimation. Fig. 5 show the estimated depth map of the selected feature points and the camera pose.

(a) The 1st frame in the ‘oldhousing’ video sequence (b) The 88th frame in the ‘oldhousing’ video sequence

Fig. 3. Original frames used for image registration

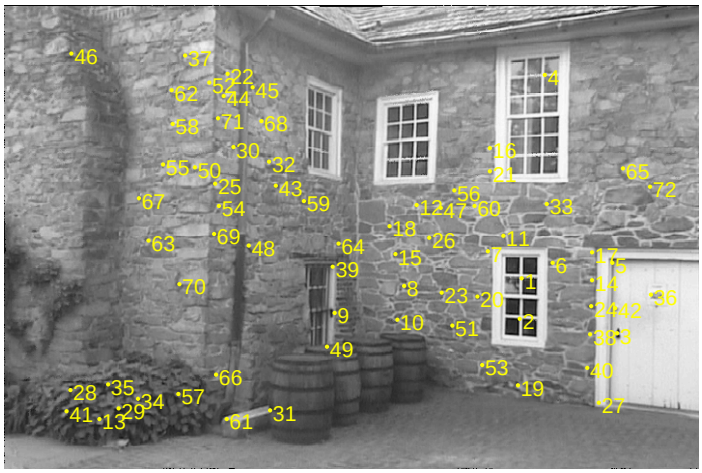

Fig. 4. The feature points selected for depth estimation on the 1st frame.

IV. Image Registration with Depth Information

Once we obtain the 3D structure of the feature points, the motion, and calibration of the camera, we can start to register the rest of the pixels in the images with the estimated depth information. The traditional image registration algorithms, such as the algorithm proposed by Davis and Keck [3], [4], try to register the two images by computing the homography matrix H between corresponding feature points. The limit of this algorithm is that they assume all the points in the physical world are coplanar or approximately coplanar, which is not true with high-rise scenes. In order to mitigate this problem, we propose a novel algorithm which first segment the image geometrically and then perform the registration to each region with depth estimation.

A. Geometrical segmentation

In order to perform the geometrical segmentation, the most intuitive method is to obtain the dense surface model of the scene and then segment the surface into several regions based on the depth of the

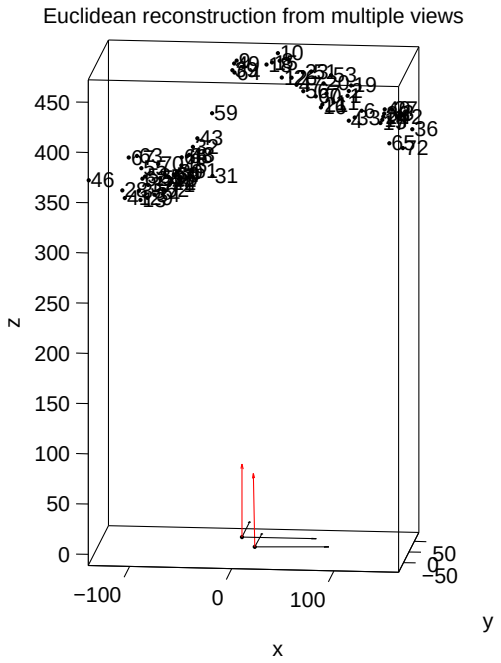

Fig. 5. The estimated depth map and camera pose for the selected feature points of the 1st and 88th frames.

points. However, we need to know the correspondence for almost all the pixels to compute the dense surface model, which means we need to know all the pixel correspondence before the registration. In order to avoid this dilemma, we will not use the traditional 3D reconstruction algorithm to estimate the dense surface model. Instead, we directly segment the scene into several regions by clustering the sparse 3D points set that we obtained in Section III. With the assumption that each segment region of the scene is approximately coplanar in the physical world, we could easily estimate the plane model and project the 3D plane onto the image frames. Comparing the assumption that the whole scene is coplanar in the physical world used in the traditional image registration algorithms, this assumption is valid in most circumstances.

There are a lot of algorithms for data clustering. The most famous hard-clustering algorithm is k-means [14]. The k-means algorithm assigns each data point to the cluster whose centroid is nearest. Here, we use the distance to a 3D plane in the physical world as the measurement. For each cluster, we could choose the plane that has the smallest sum of distance of all the data points in the cluster. However, the descent based learning methods suffer from a serious limitation. The non-global optima of the cost surface may easily resulting in poor local minima to the above methods. Techniques adding penalty terms to the cost function further increases the complexity of the cost surface and worsen the local minimum problem.

In this paper, we propose a non-linear deterministic annealing approach to solve the 3D geometrical fitting problem. We follow the deterministic annealing approach [15] and use the geometrical structure for clustering. Deterministic Annealing introduce the entropy constraint to explore a large portion of the cost surface using randomness, while still performing optimization using local information, which is similar to fuzzy c-means algorithm. Eventually, the amount of imposed randomness is lowered so that upon termination DA optimizes over the original cost function and yields a solution to the original problem.

To solve the space partitioning problem, we do not use prototype to calculate the difference. The reason is that the prototype in space partitioning is generally not sufficient to represent a plane in 3D space. Instead, we estimate the linear plane model and calculate the fitting error as the Euclidean distance between the data and the plane. The traditional local optimization algorithm will likely stuck at a local optima. In order to avoid local optima, we use local geometric structure from neighboring data points and embedded the data vectors to a higher dimension as follows.

The input data is given as a 3D point, $\mathbf{x}_i = [x_i, y_i, z_i]^T$. With the assumption that nearest data points are on the same plane, we could estimate the local plane model, $\mathbf{L}_i = [a_i, b_i, c_i]^T$ of data point $\mathbf{x}_i$ and its K nearest neighbor points.

$$\mathbf{L} = \begin{bmatrix} a(\mathfrak{X}) \\ b(\mathfrak{X}) \\ c(\mathfrak{X}) \end{bmatrix} \tag{12}$$

$$\mathbf{f} = \begin{bmatrix} \mathbf{x} \\ \mathbf{L} \end{bmatrix} \tag{13}$$

Then we revise the distortion function as follows,

$$D(\mathbf{f}_i, g_{\theta_j}) = D_1(I_1\mathbf{f}_i, g_{\theta_j}) + D_2(I_2\mathbf{f}_i, g_{\theta_j}) \tag{14}$$

$$I_1 = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \end{bmatrix} \tag{15}$$

$$I_2 = \begin{bmatrix} 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \tag{16}$$

where $D_1 = d_{i,j}$ calculate the fitting error between the data point and the estimated plane, and D_2 calculate the difference between the local estimated plane model and the cluster scale estimated plane model. D_2 is defined as follows:

$$D_2(I_2\mathbf{f}_i, g_{\theta_j}) = \frac{I_2\mathbf{f}_i^T \times g_{\theta_j}}{|I_2\mathbf{f}_i| \times |g_{\theta_j}|} \tag{17}$$

After the mapping, we apply deterministic annealing algorithm to partition the data into several clusters as follows.

$$\min_{g_{\theta_j}} F = D - TH \tag{18}$$

where $g_{\theta_j} = [a_j, b_j, c_j]$ is the geometrical surface model parameter to be estimated, D is the sum of square of geometrical fitting error and H is the entropy constraint. We define D and H as follows:

$$D = \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{K} p(\mathbf{x}_i, g_{\theta_j}) d(\mathbf{x}_i, g_{\theta_j}) = \sum_{i=1}^{N} p(\mathbf{x}_i) \sum_{j=1}^{K} p(g_{\theta_j}|\mathbf{x}_i) d(\mathbf{x}_i, g_{\theta_j}) \tag{19}$$

$$H(\mathbf{X}, g_\theta) = \sum_{i=1}^{N}\sum_{j=1}^{K} p(\mathbf{x}_i, g_{\theta_j}) \log p(\mathbf{x}_i, g_{\theta_j}) \tag{20}$$

To perform optimization we need to further analyze its terms. We can rewrite equation (20) by applying the chain rule of entropy as

$$H(\mathbf{X}, g_\theta) = H(\mathbf{X}) + H(g_\theta|\mathbf{X}) \tag{21}$$

Notice that the first term $H(\mathbf{X})$ is the entropy of the source and is therefore constant with respect to the cluster g_{θ_j} and association probabilities $p(g_{\theta_j}|\mathbf{x}_i)$. Thus we can just focus on the conditional entropy

$$H(g_\theta|\mathbf{X}) = \sum_{i=1}^{N} p(\mathbf{x}_i) \sum_{j=1}^{K} p(g_{\theta_j}|\mathbf{x}_i) \log p(g_{\theta_j}|\mathbf{x}_i) \tag{22}$$

The minimization of F with respect to association probabilities $p(g_{\theta_j}|\mathbf{x}_i)$ gives rise to the Gibbs distribution

$$p(g_{\theta_j}|\mathbf{x}_i) = \frac{\exp(-\frac{d(\mathbf{x}_i, g_{\theta_j})}{T})}{Z_x} \tag{23}$$

where the normalization is

$$Z_x = \sum_{j=1}^{K} \exp(-\frac{d(\mathbf{x}_i, g_{\theta_j})}{T}) \tag{24}$$

The corresponding minimum of F is obtained by plugging equation (23) back into equation (18)

$$F^* = \min_{p(g_{\theta_j}|\mathbf{x}_i)} F = -T \sum_{i=1}^{N} p(\mathbf{x}_i) \log Z_x \tag{25}$$

To minimize the Lagrangian with respect to the cluster model g_{θ_j}, its gradients are set to zero yielding the condition

$$\nabla_{g_{\theta_j}} F = \frac{1}{N} \sum_{i=1}^{N} p(g_{\theta_j}|\mathbf{x}_i) \nabla_{g_{\theta_j}} d(\mathbf{x}_i, g_{\theta_j}) = 0 \tag{26}$$

Non-linear deterministic annealing method (NDA) introduces the entropy constraint to explore a large portion of the cost surface using randomness, while still performing optimization using local information, which is similar to fuzzy c-means algorithm. Eventually, the amount of imposed randomness is lowered so that upon termination NDA optimizes over the original cost function and yields a solution to the original problem.

However, there is no close form solution, therefore we use a gradient descent algorithm to solve this problem. I present our algorithm in Figure. 6.

B. Depth estimation

Here, we only consider two images. Suppose for the first image, we have the 3D point set $\mathbf{X}_e^j, j = 1, 2, ..., n$ which could be divided into three clusters,$\mathbf{X}_{e1}$, $\mathbf{X}_{e2}$, $\mathbf{X}_{e3}$. For each cluster, there are at least three non-collinear points. Then we could have the plane model for this cluster. Let's take the example of $\mathbf{X}_{e1}$, suppose there are m points in the cluster and we have the plane model as follows:

$$\mathbf{A} \cdot p = 1. \tag{29}$$

where $\mathbf{A} = [\mathbf{X}_{e1}^i], i = 1, ..., m$ and $p = [a, b, c]^T$ is the plane parameter.

Given an arbitrary point $\mathbf{x}^i = [x^i, y^i]^T$ measured in pixels in the first cluster, we could estimate it's depth scale λ^i by solving the following equation.

$$\lambda^i \mathbf{x'}^i = H_1^{-1} \Pi_1 \mathbf{X}_e^i. \tag{30}$$

where $\mathbf{x'}^i = [x^i, y^i, 1]^T$, H_1^{-1} and Π_1 are estimated in Section III. In Eq. 30, only λ^i is unknown and with the constraint on $\mathbf{X}_e^i$ with Eq. 29, we could easily get the value of λ^i.

Then, with $\Pi_1 = [I, 0]$, we could have $X_p^i = [\lambda_1^i x^i, \lambda_1^i y^i, \lambda_1^i, 1]$. from Eq. 6, we can get the relation between two image projection point $\mathbf{x}_1^i$ and $\mathbf{x}_2^i$ as follows:

$$\widehat{\mathbf{x_2^i}'} = \Pi_2 \mathbf{X}_p^i. \tag{31}$$

where $\widehat{\mathbf{x_2^i}'} = [\lambda_2^i x_2^i, \lambda_2^i y_2^i, \lambda_2^i]$. We could then get the position of the corresponding point $\mathbf{x}_2^i = [x_2^i, y_2^i]$ in the second image.

1) Algorithm 6 **NDA based geometrical segmentation algorithm**
2) **Set Limit**
3) K_{max}: maximum number of clusters
4) T_{init}: starting temperature
5) T_{min}: minimum temperature
6) δ: perturbation vector
7) α: cooling rate (must be < 1)
8) I_{max}: maximum iteration number
9) th: Iteration threshold
10) sth: Surface distance threshold
11) **Initialization**
12) $T = T_{init}, K = 2, \Lambda_1 = (X^T X)^{-1} X^T \vec{1}, \Lambda_2 = \Lambda_1, [p(\mathbf{\Lambda}_1|\mathbf{x}_i), p(\mathbf{\Lambda}_2|\mathbf{x}_i)] = [\frac{1}{2}, \frac{1}{2}], \forall i.$
13) **Perturb**
14) $\Lambda_j = \Lambda_j + \delta, \forall j.$
15) $L_{old} = D - TH.$
16) **Loop until convergence,** $i = 0\ \forall j$
17) For all $\mathbf{x}_i$ in the training data, compute the association probabilities

$$p(\mathbf{\Lambda}_j|\mathbf{x}_i) = \frac{\exp(-\frac{d(\mathbf{x}_i, \mathbf{\Lambda}_j)}{T})}{\sum_{j=1}^{K} \exp(-\frac{d(\mathbf{x}_i, \mathbf{\Lambda}_j)}{T})} \tag{27}$$

18) update the surface model

$$\mathbf{\Lambda}_j \longleftarrow \mathbf{\Lambda}_j + \alpha \nabla_{\mathbf{\Lambda}_j} F. \tag{28}$$

19) i = i+1;
20) if ($i > I_{max}$ or $\nabla_{\mathbf{\Lambda}_j} F < th$) End **Loop**
21) **Model Size Determination**
22) if($d(\mathbf{\Lambda}_j, \mathbf{\Lambda}_{j+1}) < sth$)
23) replace $\mathbf{\Lambda}_j, \mathbf{\Lambda}_{j+1}$ by a single plane
24) $K =$number of planes after merging
25) **Cooling Step**
26) $T = \alpha T.$
27) if ($T < T_{min}$)
28) perform last iteration for $T = 0$ and STOP
29) **Duplication**
30) Replace each plane by two planes at the same location, $K = 2K.$
31) **Goto Step 10**

Fig. 6. NDA based geometrical segmentation algorithm

V. Experimental Results

The data includes a sequence of 88 images captured from one camera. We first select 72 feature points in the first image and then find the corresponding feature points in the rest of the images. The depth estimates of these points are calculated by the algorithm introduced in Section III.

In our experiment, we regard the first image's local coordinate system as world coordinate system so the first image can be viewed as a reference image. Then the rest of the images are registered to the reference image. We also applied the algorithm proposed by Davis and Keck [3] to register the input images for comparison purpose.

Fig. 3 is the 1st frame and the 88th frame in the test image sequence. Fig. 7 is the registration result using our algorithm and Fig. 8 is the output of the algorithm proposed by Davis and Keck [3]. Fig. 9 shows the difference image between the registered image and the first image using our algorithm and Fig. 10 shows the difference image from the algorithm of Davis and Keck [3]. We can see that our result can mitigate the parallax problem since the roof and wall corners are registered correctly; on the contrary, the registered image by the algorithm of Davis and Keck [3] has a lot of artifacts caused by the parallax problem. We also show some registration results using our algorithm in Fig. 11$\sim$ Fig. 12.

In order to further compare our algorithm to the algorithm proposed by Davis and Keck, we compute the root of mean squared errors (RMSE) of the registration results from both algorithms. Fig. 13 shows that the registration error of our algorithm is less than 50% than that of the algorithm proposed by Davis

Fig. 7. Our algorithm test result, in which the 88th frame is registered to the 1st frame.

Fig. 8. The test result under the algorithm of Davis and Keck, in which the 88th frame is registered to the 1st frame.

and Keck.

The result shows that our image registration algorithm can mitigate the parallax problem because most of the scene is registered without vibration, as opposed to registration results under the algorithm of Davis and Keck in which the high-rise scene in the sensed images significantly moved after registration to the reference images. The reason is that the algorithm of Davis and Keck assumes all the points in the images are coplanar. While this assumption is satisfied when the distance between the camera and the interested scene is so large that the small depth variation can be neglected, it fails in the case of high-rise scene. Therefore, depth information should be used to accomplish the registration for this specific high-rise region of the images.

Finally, we would like to point out that the algorithm proposed by Davis and Keck [3] assumes a planar registration. Their scheme was designed for use with high-altitude aerial imagery where planar transformations are fairly good approximations. Furthermore, their scheme uses RANSAC to remove poor matching points during the computation. This can help to deal with some depth discontinuities that

Fig. 9. The difference image between the registered 88th image (using our algorithm) and the 1st image.

Fig. 10. The difference image between the registered 88th image (using the algorithm of Davis and Keck) and the 1st image.

may be present in the high-altitude aerial images. In our experiments, the test images contain salient 3D scenes; these images are out of the domain for the algorithm of Davis and Keck. This is the reason why the algorithm of Davis and Keck does not perform well.

VI. Conclusion

In this paper, we propose a new 2D image registration method by leveraging depth information. While traditional image registration algorithms fail to register high-rise scene accurately because the points cannot

Fig. 11. The 37th frame in the 'oldhousing' video sequence.

Fig. 12. Our algorithm test result, in which the 37th frame is registered to the 1st frame.

be assumed to be simply planar, our image registration algorithm can mitigate the parallax problem.

Our future works include:

- Develop a robust 3D model based on the state-of-the-art depth estimate algorithm [16][17] given a video sequence. The reliability of the depth estimates is crucial to depth-based registration algorithm; therefore, the highly robust 3D reconstruction technique is required to implement our algorithm. Up to now, most recent depth recovery algorithms reported in the literature claim to recover consistent depth from some challenging video sequences [16][17]. We can apply or modify this state-of-the-art

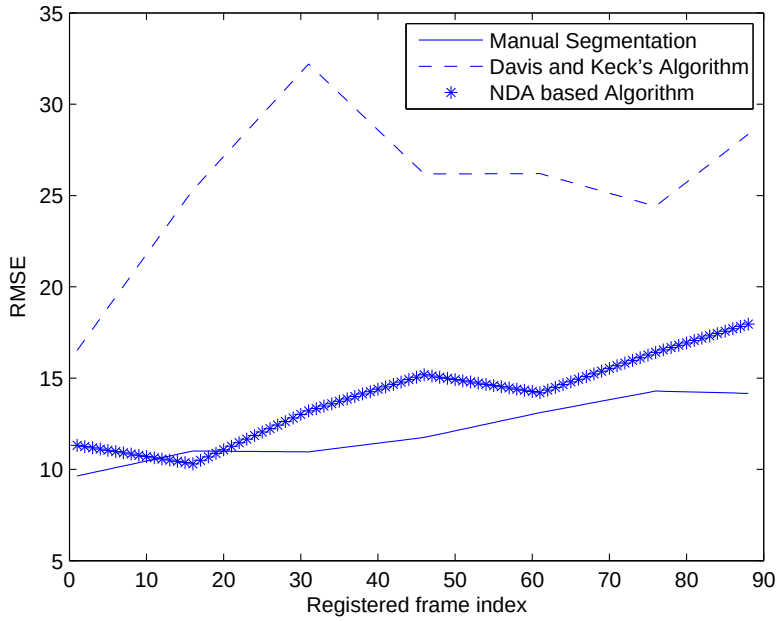

Fig. 13. Our algorithm test result comparing to that under the algorithm of Davis and Keck, in which all the 88 frames are registered to the 1st frame.

depth map recovery method to develop depth-based image registration algorithm.

- Combine depth-based image registration method with traditional algorithms. In other words, we can use depth information to register high-rise region while applying traditional registration algorithm for other planar region of the image. The purpose is to tradeoff between the accuracy of the registration and the high computational cost introduced by 3D reconstruction. The combined algorithm thus can enjoy both the high efficiency of the traditional algorithm and the high robustness of the depth-based registration method.
- We would use our depth-based image registration algorithm in practical applications to further verify the performance of our algorithm compared to the traditional ones.

DISCLAIMERS

The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of AFRL or the U.S. Government.

ACKNOWLEDGEMENT

This material is based on research sponsored by AFRL under agreement number FA8650-06-1-1027. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright notation thereon. The authors would like to thank Dr. James W. Davis and Mark Keck from the Ohio State University for permission to use their registration algorithm and code.

REFERENCES

[1] L. Brown, "A survey of image registration techniques," *ACM computing surveys (CSUR)*, vol. 24, no. 4, pp. 325–376, 1992.
[2] B. Zitova and J. Flusser, "Image registration methods: a survey," *Image and vision computing*, vol. 21, no. 11, pp. 977–1000, 2003.
[3] J. Davis and M. Keck, "OSU Registration Algorithm," *Internal Report, Ohio State University, USA*.

[4] O. Mendoza, G. Arnold, and P. Stiller, "Further exploration of the object-image metric with image registration in mind," in *Proceedings of the SPIE, Symposium on Multisensor, Multisource Information Fusion: Architectures, Algorithms, and Applications* (B. V. Dasarathy, ed.), vol. 6974, pp. 697405–697405–12, April 2008.
[5] L. Kitchen and A. Rosenfeld, "Gray-level corner detection," 1980.
[6] L. Dreschler and H. Nagel, *Volumetric model and 3D-trajectory of a moving car derived from monocular TV-frame sequences of a street scene*. Univ., Fachbereich Informatik, 1981.
[7] W. Forstner and E. Gulch, "A fast operator for detection and precise location of distinct points, corners and centres of circular features," in *Proc. ISPRS Intercommission Conference on Fast Processing of Photogrammetric Data*, pp. 281–305, 1987.
[8] J. Noble, "Finding corners," *Image and Vision Computing*, vol. 6, no. 2, pp. 121–128, 1988.
[9] W. Pratt *et al.*, "Digital image processing," *New York*, pp. 429–32, 1978.
[10] A. Goshtasby and G. Stockman, "Point pattern matching using convex hull edges.," *IEEE TRANS. SYST. MAN CYBER.*, vol. 15, no. 5, pp. 631–636, 1985.
[11] G. Stockman, S. Kopstein, and S. Benett, "Matching images to models for registration and object detection via clustering," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 4, pp. 229–241, 1982.
[12] Y. Ma, S. Soatto, J. Kosecka, Y. Ma, S. Soatta, J. Kosecka, and S. Sastry, *An invitation to 3-D vision*. Springer, 2004.
[13] B. Lucas and T. Kanade, "An iterative image registration technique with an application to stereo vision," in *International joint conference on artificial intelligence*, vol. 3, p. 3, 1981.
[14] S. Lloyd, "Least squares quantization in PCM," *IEEE Transactions on Information Theory*, vol. 28, no. 2, pp. 129–137, 1982.
[15] K. Rose, "Deterministic annealing for clustering, compression, classification, regression, and related optimization problems," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2210–2239, 1998.
[16] G. Zhang, J. Jia, T. Wong, and H. Bao, "Recovering consistent video depth maps via bundle optimization," in *IEEE Conference on Computer Vision and Pattern Recognition, 2008. CVPR 2008*, pp. 1–8, 2008.
[17] G. Zhang, J. Jia, T. Wong, and H. Bao, "Consistent Depth Maps Recovery from a Video Sequence," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 974–988, 2009.