Adaptive Region-Based Video Registration

Jiangjian Xiao Yunjun Zhang Mubarak Shah

Computer Vision Lab, School of Computer Science
University of Central Florida, Orlando, Florida 32816, USA
{*jxiao, yjzhang, shah*}*@cs.ucf.edu*

Abstract

Video registration without meta data (camera location, viewing angles, and reference DEMs) is still a challenging problem. With the aim of handling this kind of problem, this paper presents an adaptive region expansion approach to propagate the alignment process from high confidence areas (reliable salient features) to low confidence areas and to simultaneously remove outlier regions. Hence, we re-cast the image registration problem as a partitioning problem such that we determine the optimal supporting regions and their corresponding motion parameters for the registration. First, we determine sparse robust correspondences between mission and reference images by using our wide baseline algorithm. Next, starting from the seed regions, the aligned areas are expanded to the whole overlapping areas using the graph cut algorithm, which is controlled by the level set representation of the previous region shape. Consequently, a robust video registration is achieved if the scene can be represented by one homography. Furthermore, we extend this approach to multi-homography video registration problem for 3D scenes, which cannot be directly solved by the current alignment methods. Using our motion layer extraction algorithm, the mission video first is segmented into several layers, then each layer is respectively aligned to the reference image by employing the region expansion algorithm. Several examples are demonstrated in the experiments to show that our approach is effective and robust.

1 Introduction

Image registration and alignment have been studied for a long time in different areas, including photogrammetry, remote sensing, image processing, computer graphics, medical imaging, computer vision, etc [11, 17, 2]. Registration techniques can be classified based on the following two factors: Model of transformation (motion) between two images (mission and reference), and method of alignment used [11].

The motion model is dependent on the geometry of the imaged scene and dynamics of the sensor and object motion. Given two images of one planar scene, a single motion model can be effectively fitted using the existing registration approaches. If the scene has multiple planes, it is difficult to achieve accurate motion layer clustering, segmentation, and alignment using two wide baseline images [13]. However, given a video sequence an accurate layer segmentation can be obtained by exploiting spatiotemporal information [16, 4], which makes it possible to perform the layer-based registration.

Alignment methods can be broadly categorized into three classes: intensity-based methods, feature-based methods, and hybrid methods. The intensity-based methods are based on the well-known optical flow constraint equation $f_x u + f_y v + f_t = 0$ [8], which can be solved by minimizing the sum of squares of pixel-wise differences (SSD). Generally, these methods are more useful for frame-to-frame registration of video frames with a simple camera motion, where the pixel motion is small and the image intensities are similar [12, 9]. In feature-based methods, the main steps are finding robust features, establishing correspondences, fitting some transformation, and applying the transformation to warp the images [2]. These methods are relatively fast and more suitable for registration of two dissimilar images with large and complicated motion or transformation. Recently, several hybrid methods have been proposed to integrate the merits of intensity-based and feature-based methods [5]. In these methods, first a set of features are extracted, then an iterative optimization procedure is applied on the supporting regions around these features to obtain the robust registration.

Currently, some registration problems, such as video mosaicing, and registration of video acquired by airborne sensor with the reference image in presence of camera information [5, 14], have been solved quite well. However, there are still some unsolved open problems in this area. First, how to obtain a reliable initial estimation for the motion parameters if the camera information (e.g. location, viewing angles, and sensor model) is not available (typically for the wide baseline images). Second, how to deal with the outliers when the images are taken at different times, which may look different due to moving objects, shadow changing, and vegetation growing, etc. Third, how to handle complex motion models

in a single 3D scene with multiple homographies.

With the aim of addressing the above limitations of the current methods, we propose a novel framework to perform video registration of a 3D scene (which can be approximated by multiple planes) without any knowledge of the meta data. The proposed approach is as follows. First, we reformulate the image registration problem as a partitioning problem, which can effectively determine the optimal supporting regions and their corresponding motion parameters for the registration. In single layer video registration, we design a region expansion process to adaptively propagate the alignment process from high confidence seed regions to the low confidence areas and simultaneously remove outlier regions. In order to obtain the starting seed regions, we apply our wide baseline algorithm [15] to compute a set of reliable seed correspondences between the mission and reference images. Then, starting from the seed regions, the initially aligned area is expanded to the whole overlapping areas using a graph cut algorithm, which is regulated by the level set of the previous region shape. Consequently, a robust video registration is achieved for this layer. If the scene contains multiple planar layers, we first employ the layer extraction algorithm [16] to determine the number of layers and their descriptions which can provide the accurate support region for each layer. Then the region expansion algorithm is applied to each layer respectively and the final multi-layer registration is achieved.

In the remainder of this paper, we first describe the algorithm for single layer video registration, then we extend the framework for multi-layer video registration of complex scene. After that, we further illustrate our approach for video registration using multiple seed region expansion.

2 Single Layer Registration

In this section, we propose a two-stage approach to integrate the merits of the sparse and dense image features for single layer video registration. In the first stage, we determine a set of sparse correspondences between the mission and reference images. Then, starting from the initial seed correspondences, the aligned regions are gradually expanded to cover the whole overlapping areas between both images. At the meantime, the outlier regions are detected and removed, such as the appearing/disappearing objects which may harm the registration process.

2.1 Determining correspondences

There are several methods to compute robust correspondences for wide baseline images [3, 15]. Here we use our previous work [15] to determine a set of reliable corresponding corners. First, we identify a salient features, edge-corners, in both wide baseline images, which provide robust and consistent matching primitives. Then, the supporting regions of the corresponding corners should satisfy the following equation:

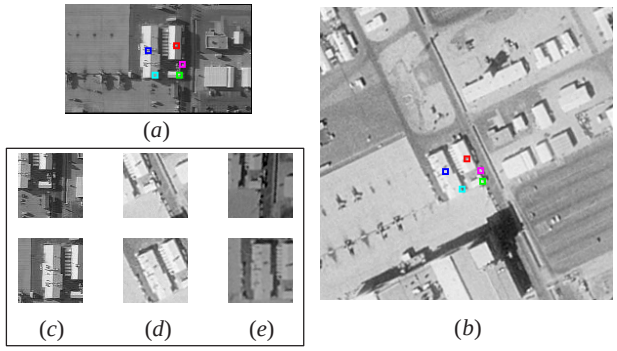


Figure 1: Determining correspondences between wide baseline images. (a) mission image. (b) a small part of reference image. Several correspondences are computed by the wide baseline matching algorithm. (c)-(e) illustrate the detail of matching process of green (the top row) and blue (the bottom row) corners. (c) patch windows from (a). (d) the corresponding patch windows from (b). (e) warping patch (d) by the best affine transformation, where patch (e) is similar to patch (c).

$$\mu I_2(\mathbf{A}\mathbf{x}+\mathbf{d})+\delta = I_1(\mathbf{x}), \tag{1}$$

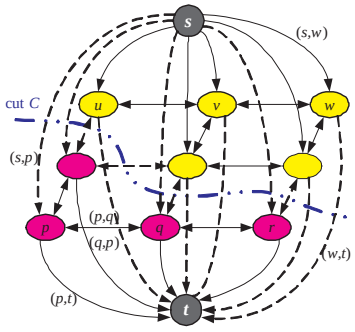
where I_1 and I_2 are two original images, $\mathbf{x} \in \mathbb{R}^2$ are the pixel coordinates, $\mathbf{A} \in \mathbb{R}^{2\times 2}$ is an affine matrix and $\mathbf{d} \in \mathbb{R}^2$ is a translation vector, μ depends on the reflection angle of the light source (corresponding to image contrast), and δ depends on the camera gain (corresponding to image brightness). The illumination coefficients, μ and δ, are used to compensate the illumination change between wide baseline images, while the affine model, $\mathbf{A}$ and $\mathbf{d}$, is used to compensate the geometric distortion between two image patches (or corners). We can compute the best match by minimizing the residual (or SSD)

$$\epsilon = \sum_{\Omega}[(\mu I_2(\mathbf{A}\mathbf{x}+\mathbf{d})+\delta) - I_1(\mathbf{x})]^2, \tag{2}$$

where Ω is the image patch. Instead of minimizing the function starting from $\mathbf{A} = \mathbf{I}$ (identity matrix), $\mathbf{d} = [0\ 0]^{\mathrm{T}}$, $\mu = 1$, and $\delta = 0$, we first search the best rotation and scaling component implied on the this affine transformation, then we apply Newton-Raphson iteration from $\mathbf{A} = \mathbf{R} \times \mathbf{S}$ (the best rotation and scaling) to speed up the minimization procedure and obtain stable results. Fig. 1 shows the detailed matching process for a small patch.

2.2 Adaptive Region Expansion Process

Once the seed correspondences are estimated between the mission and reference images, our purpose is to perform registration for each plane (layer) in the scene. For each pair of correspondences, we consider a small patch centered around each seed region, which can be approximated as a planar patch (or an initial layer) in the scene. Therefore, we get a number of initial layers, and each layer is supported by a



Edge	Weight
(s,p)	$D_p(0)$
(p,t)	$D_p(1)$
(p,q) (q,p)	V(p,q)

Figure 2: An example of a graph $\mathcal{G}$ for a image. Nodes p, q, r, $\cdots$, w are the pixels in the image. s is source node and t is sink node. The edges connected to source or sink are called t-links, such as (s,p) and (p,t). The edges connected to two neighboring pixel nodes are called n-links, which have two directions, such as (p,q) and (q,p). After computing minimum cut $\mathcal{C}$, the nodes are partitioned into supporting pixels (source) and non-supporting pixels (sink). The dotted links are crossed by $\mathcal{C}$. Table shows the weights.

small square region with its corresponding affine transformation. For a general minimization case, we use a vector Θ to denote all the parameters which can be used to minimize the errors between two images. The image dissimilarity function can be rewritten as:

$$\epsilon = \sum_{\Omega} [I_2(\mathbf{x}, \Theta) - I_1(\mathbf{x})]^2. \tag{3}$$

where Θ includes two parts: one is the illuminational coefficients μ and δ, the other is motion parameters (i.e. affine or homography parameters). Using linear [15] or nonlinear [12] optimization algorithms, ϵ can be iteratively minimized on this small patch (e.g., 41×41) and the corresponding Θ is estimated. Nevertheless, this minimization process may produce two issues. First, the estimated parameters Θ using the small patch may over-fit the pixels inside the region, and may not correctly represent the global transformation of a larger region. Second, this process ignored the appearing/disappearing between two images, such as the moving objects, occlusion area, shadow changing, etc.

To overcome the above problems, we propose a novel approach to gradually expand the seed region by identifying the supporting pixels using a bi-partitioning graph cut method integrated with level sets. First, a smoothness energy term between neighboring pixels is introduced, which can effectively maintain the partitions piecewisely smooth [6, 1]. Then, using the level set representation of the previous region, the contour of the seed region is gradually evolved by propagating the region's front along its normal direction. Our registration problem can be easily formulated into the graph cut framework. In this framework [1], we seek the labeling function f which partitions the pixels in region Ω into two groups: one is supporting region which has label $f = 0$, the other is outlier region which has label $f = 1$. This partition can be achieved by minimizing the energy function

$$E = \sum_{(p,q)\in\mathcal{N}} V(p,q) + \sum_{p\in\Omega} D_p(f_p), \tag{4}$$

where the first one is a piecewise smoothness term, the second one is a data penalty term, $\mathcal{N}$ is a 4-neighbor system, f_p is the label of a pixel p, $D_p(f_p)$ is data penalty function, and $V(p,q)$ is smoothness penalty function which is is designed to more likely maintain the same label for p and q if they have similar intensities [16].

To minimize the energy function, a weighted graph $\mathcal{G} = \langle \mathcal{V}, \mathcal{E} \rangle$ is constructed as shown in Fig. 4, where $\mathcal{V}$ is a node set (image pixels) and $\mathcal{E}$ is a link set that connects the nodes. After assigning weights for the links following the table in Fig. 4, we compute a minimum cut $\mathcal{C}$ using standard graph cut algorithm and partition the original region into the supporting region and outlier region. However, this process cannot expand the region from the initial seed patch to the exterior to obtain more supporting pixels. Hence, we use the contour of the previous seed region as a prior to compute the level set regulation of this region [10, 7], which allows the region contour to evolve along the normal direction.

Fig. 3 shows a detailed expansion process starting from one initial seed region. Fig. 3.a and b show the initial contours of the corresponding seed regions. Based on the initial contour of the original seed region Ω^0 (Fig. 3.b), we construct a mask β of this region, which has a value in [0, 1], where the interior pixels of the region are marked by 1 and the others are marked by 0. Then, a level set ϕ (Fig. 3.e) can be simply computed by convolving the region mask with a Gaussian kernel as: $\phi = G * \beta$, where the value of ϕ falls down along the contour normal direction until $\phi_p = 0$. Then, we warp the second image using the corresponding homography, and construct a graph $\mathcal{G}$ for the pixel with $\phi_p > 0$. After that, we apply the level set ϕ to change the weights of the sink side t-links for each pixels, such that the weights of the pixels inside the region are almost not changed while the weight (p,t) will decrease when the pixel p is away from the boundary. As a result, the minimum cut $\mathcal{C}$ is most likely to cut the outside pixels, and label them as the non-supporting pixels for this region. This way, the new expanded supporting region Ω^1 can be computed as Fig. 3.f. After several iterations (Fig. 3.f-h), the region's boundary gradually propagate from the center to exterior until stopping at the overlapping boundary of the two images, and the alignment will be stable. Fig. 3.h shows the final region Ω^5 after five iterations and Fig. 3.c shows the final registration results using the projective transformation computed by this approach.

3 Multiple Layer Registration

In a single plane scene, only one layer is available. After expanding and merging the seed regions between mission and reference images, one unique merged layer registration will

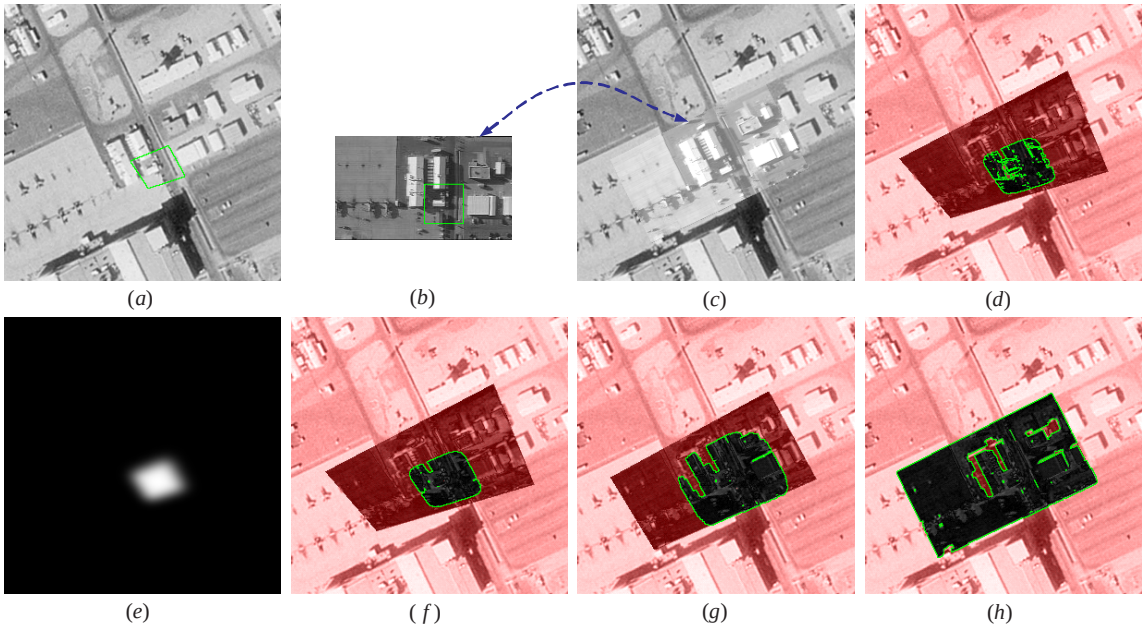


Figure 3: Region expansion process. $(a-b)$ initial corresponding patch contours on the reference and mission images respectively. (c) the final registration result, where the intensities of the embedded mission image are adjusted by illumination coefficients μ and δ. (d) the simple expansion and partitioning started from contour (a). (e) the level set representation of the initial contour (a). $(f-h)$ are intermediate results using graph-cut method with the level set regulation, which can guarantee the expansion is gradually evolved from the center to boundary. *Note:* The green boxes in (a) and (b) are the initial seed regions. $(f-h)$ are difference maps between the warped (b) and (a), and the green contours in $(f-h)$ are supporting region boundaries obtained after using bi-partitioning algorithm. The unsupporting pixels are masked by red.

Figure 4: Registration with multiple layers. (a) One frame of mission video. (b) Reference image. Instead of registration using one simple motion model, the alignment of this scene requires two motion models, which correspond to the layers of the door and wall respectively.

be achieved. However, if the scene includes multiple layers, the motion models can vary from a simple global parametric map to multiple parametric mappings, where the pixel's motion is mapped to several parameter clusters. Fig. 4 shows one example of this case from a "door-wall" sequence, which contains two layers. It is very difficult to directly align the whole mission image to the reference image without any layer information.

Theoretically, using two images (one from the mission video, the other from the reference image), the layer clustering and hence the segmentation can be computed by minimizing a proper energy function. Nevertheless, similar to the wide baseline matching problem, there are several very difficult to be resolved for the segmentation, such as illumination variation, shadows, and moving objects, etc. Hence, the results usually are not accurate enough, typically on the layer boundary [13]. Fortunately, in context of video registration, the temporal information is available in the mission video, where the motion layers can effectively be extracted [16, 4]. For each layer, the motion parameters are different and only represent the mapping of the pixels in this layers. In this paper, we use our multi-frame graph cut framework [16] to achieve an accurate segmentation for mission video sequence.

Fig. 5.a and b show video segmentation results for the "door-wall" sequence. In this framework, we can not only obtain the accurate layer segmentations which correspond to the wall and door respectively (Fig. 5.d), but also can explicitly identify the occluded pixels between overlapping layers as the red pixels shown in Fig. 5.b. During the motion segmentation, the supporting region and its motion parameters are computed. After given a reference image (Fig. 5.c) which was taken from a fairly different location and orientation with different camera white-balance, we can apply the extracted layer information to perform layer based registration. Using the segmented layers (Fig. 5.d-e), we align the different supporting regions to the reference image separately using the adaptive region expansion approach. The final registration

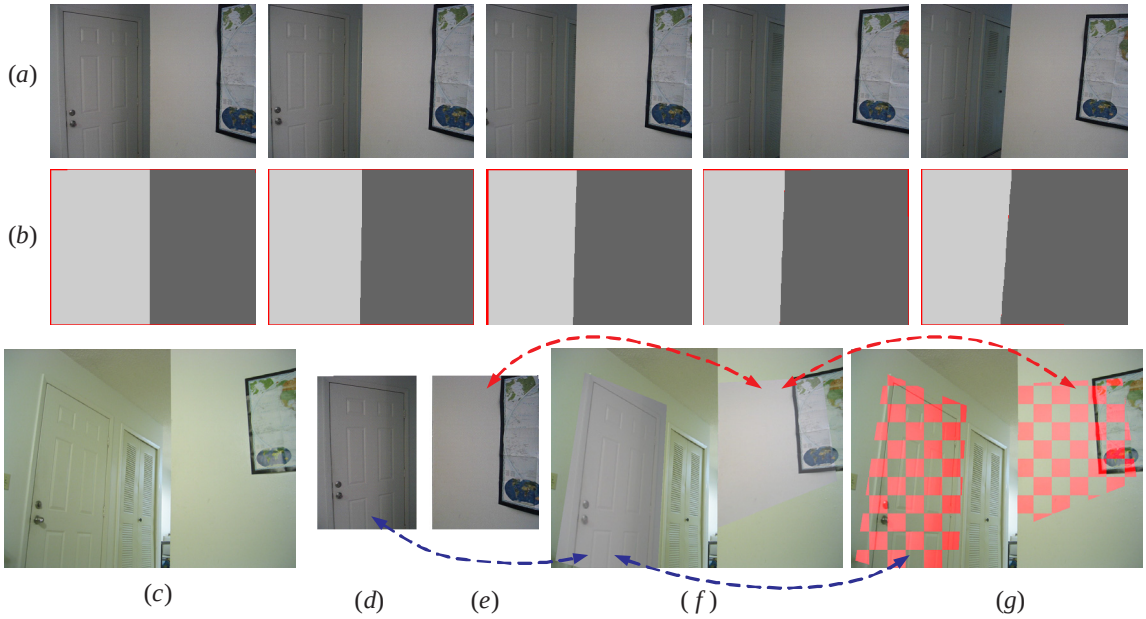


Figure 5: Multiple layers registration. (a) several frames from the mission image sequence. (b) corresponding segmentation results of (a). (c) reference image. $(d-e)$ are the layers of door and wall in frame 1. (f) Registration results of frame 1. (h) Checkered display after alignment.

results of the first frame are shown in Fig. 5.f and 5.h.

In direct registration method, only one set of motion parameters is available, which can not represent multiple layers very well. The registration process will have to fit all of the pixels in the images, which may over-fit for one true layer and mis-align for the other true layers or cause some undesired distortions. Compared to the direct registration, our approach has two advantages: First, we employs different sets of motion parameters to align the corresponding layers, which correctly represents the mapping of the pixels in this layers. Second, the layer segmentation also provide an accurate supporting region for each layer, which prevents the later region expansion process over the layer boundary. Therefore, for each layer registration, our approach can effectively avoid to include the pixels from the other layers, typically for low-textured area (such as the white door or wall), and achieve the accurate alignment region for each layer.

4 Multi-seed region expansion

If the scene only includes one layer, several initial seed regions will share the same motion transformation. Starting from the multiple initial seed regions simultaneously, our region expansion algorithm can be speeded up as shown in our experiments later. From our results (Fig. 3), it is obvious that our approach can obtain the piecewise region expansion, which is insensitive to noise. The outlier regions are also detected and removed, such as shadows of the buildings. Therefore, we can effectively improve the accuracy of the alignment by using this approach.

For single layer registration, after determining sparse correspondences between mission and reference images, we can also expand these seed regions simultaneously to speed up the alignment process. The initial homography can be computed by two ways: One is to select the most robust affine transformation of the seed regions using RANSAC technique; the other is to estimate an initial homography voted by all of these correspondences.

In Fig. 6, we show the multi-seed expansion process. Since a number of correspondences are determined, it is easily to estimate a robust initial homography using all of correspondences. Then, starting from the initial homography, we expand all of the initial seed regions simultaneously until the overlapping areas between the mission and reference images are covered. Our graph cut algorithm also detects and remove the outlier regions, most of which are due to growing vegetation or shadow changing. Fig. 6.h and 6.i compare the zoomed results before and after applying region expansion process. *Note: All of our results are available at our web site [18].*

5 Conclusions

In this paper, we successfully formulated the video registration problem in the partitioning framework, where the optimal supporting regions of the different layers and their corresponding motion parameters are determined. For each layer, the adaptive region expansion algorithm can efficiently propagate the alignment process from high confidence areas (reliable salient features) to low confidence areas. In addition, the outlier regions are also identified and removed. We further extend this approach to multi-layer video registration

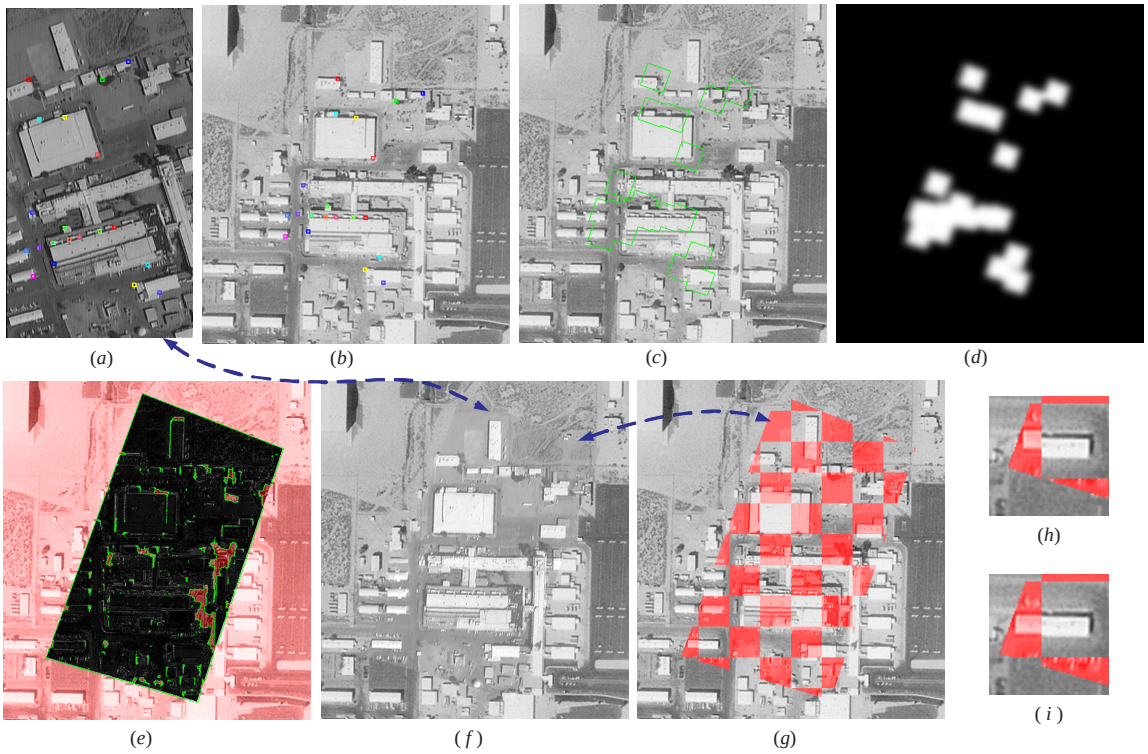


Figure 6: Registration using multi-seed region expansion. (a) mission image. (b) small part of a reference image. The correspondences are marked on the (a) and (b) by the same colors. (c) and (d) are the initial seed region and its level set representation. (e) final region contour after expansion. The un-supporting regions are shown by red. (f) registration results. (g) checkered display after alignment. (h) and (i) the zoomed alignment results before and after applying region expansion alignment.

of a 3D scene, which cannot be directly solved by the current alignment methods. After segmenting the mission video into several layers using our layer extraction algorithm, we can align each layer to the reference image separately for the video sequence.

In the future, we would like to investigate different dissimilarity measurements on this partitioning framework to handle other kinds of registration problems, such as multi-sensor image registration, image to 3D registration, etc.

References

[1] Y. Boykov, O. Veksler, R. Zabih, "Fast Approximate Energy Minimization via Graph Cuts", IEEE PAMI, 2001.

[2] L. Brown, "A Survey of Image Registration Techniques", ACM Computing Surveys, 1992.

[3] V. Ferrari, T. Tuytellars, L. Van Gool, "Wide-Baseline Multiple-View Correspondences", CVPR, 2003.

[4] Q. Ke, T. Kanade, "A Robust Subspace Approach to Layer Extraction", IEEE Workshop on Motion and Video Computing, 2002.

[5] Y. Keller, A. Averbuch, "Implicit similarity: a new approach to multi-sensor image registration", CVPR, 2003.

[6] V. Kolmogorov and R. Zabih, "What Energy Functions can be Minimized via Graph Cuts?", ECCV, 2002.

[7] S. Osher, R. Fedkiw, "Level Set Methods and Dynamic Implicit Surfaces". The Springer-Verlag Press, 2003.

[8] B. Horn and B. Schunck, "Determining optical flow", Artificial Intelligence, 1981.

[9] H. Sawhney, S. Hsu, R. Kumar, "Robust Video Mosaicing through Topology Inference and Local to Global Aligment", ECCV, 1998.

[10] J. Sethian. "Level Set Methods and Fast Marching Methods". Cambridge University Press, 1999.

[11] Mubarak Shah and Rakesh Kumar (editors),"Video Registration", Kluwer Academic Publishers, 2003.

[12] R. Szeliski, "Video Mosaics for Virtual Environments", IEEE Computer Graphics and Applications, 16(2), pp. 22-30, 1996.

[13] J. Wills, S. Agarwal, S. Belongie, "What Went Where", CVPR, 2003.

[14] R. Wildes, D. Hirvonen, S. Hsu, R. Kumar, W. Lehman, B. Matei, W. Zhao, "Video georegistration: Algorithm and quantitative evaluation", ICCV, 2001.

[15] J. Xiao, and M. Shah, "Two-Frame Wide Baseline Matching", ICCV, 2003.

[16] J. Xiao, and M. Shah, "Motion Layer Extraction in the Presence of Occlusion using Graph Cut", CVPR, 2004.

[17] B. Zitova, J. Flusser, "Image registration methods: a survey", Image and Vision Computing, 2003.

[18] http://www.cs.ucf.edu/∼vision/projects/layer_registration/.