Machine Vision and Applications (2005) 16: 75–84
Digital Object Identifier (DOI) 10.1007/s00138-004-0162-5

Machine Vision and Applications

Layer-based video registration

Jiangjian Xiao, Mubarak Shah

Computer Vision Lab, School of Computer Science, University of Central Florida, Orlando, FL 32816, USA
(e-mail: {jxiao, shah}@cs.ucf.edu)

Received: 16 September 2004 / Accepted: 23 September 2004
Published online: 19 January 2005 –

Abstract. Registration of a mission video sequence with a reference image without any metadata (camera location, viewing angles, and reference DEMs) is still a challenging problem. This paper presents a layer-based approach to registering a video sequence to a reference image of a 3D scene containing multiple layers. First, the robust layers from a mission video sequence are extracted and a layer mosaic is generated for each layer, where the relative transformation parameters between consecutive frames are estimated. Then, we formulate the image-registration problem as a region-partitioning problem, where the overlapping regions between two images are partitioned into supporting and nonsupporting (or outlier) regions, and the corresponding motion parameters are also determined for the supporting regions. In this approach, we first estimate a set of sparse, robust correspondences between the first frame and reference image. Starting from corresponding seed patches, the aligned areas are expanded to the complete overlapping areas for each layer using a graph-cut algorithm with level set, where the first frame is registered to the reference image. Then, using the transformation parameters estimated from the mosaic, we initially align the remaining frames in the video to the reference image. Finally, using the same partitioning framework, the registration is further refined by adjusting the aligned areas and removing outliers. Several examples are demonstrated in the experiments to show that our approach is effective and robust.

Keywords: Video registration – Layer-based registration – Level set – Graph cuts – Adaptive region expansion

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, and computer vision [3,16,24]. Registration techniques can be classified based on the following two factors: the motion model between mission and reference images, and the method of alignment [16].

The motion model depends on the geometry of the imaged scene and dynamics of the sensor and object motion. Given two images of a planar scene, a single motion model (affine or projective) can be fitted using the existing registration approaches (Fig. 1a). For a scene containing multiple planes (or layers), it is difficult to obtain correct registration using only two images (mission and reference) due to the inconsistent motion model. Hence, the registration may overfit one layer or the layer boundaries may not be accurate [19]. However, given a video sequence, an accurate layer segmentation can be obtained by exploiting spatiotemporal information [1,6,8,22] (Fig. 1b), which makes it possible to perform the layer-based registration.

Alignment methods can be broadly categorized into three classes: intensity-based (or appearance) methods, feature-based methods, and hybrid methods. The intensity-based methods are based on the well-known optical flow constraint equation [5], which can be solved by minimizing the sum of squares of pixelwise 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 [11,17]. In the feature-based methods, the main steps include: finding robust features, establishing correspondences, fitting some transformation, and applying the transformation to warp the images [3,23]. These methods are relatively fast and more suitable for the registration of two dissimilar images with a large and complicated motion or transformation. Recently, several hybrid methods have been proposed to integrate the merits of intensity-based and feature-based methods [7,15]. In these methods, a set of features is extracted, then an iterative optimization procedure is applied to the supporting regions around these features to minimize some dissimilar measurements.

Currently, some registration problems, such as video mosaicing and registration of video acquired by an airborne sensor to a reference image in the presence of camera information [7,13,14,20], have been solved quite well. However, some problems in this area remain unresolved:

1. How do we obtain a reliable initial estimation of motion parameters if the camera information (e.g., location, viewing angles, and sensor model) is not available? In particular, if camera location and orientation are quite different, such as

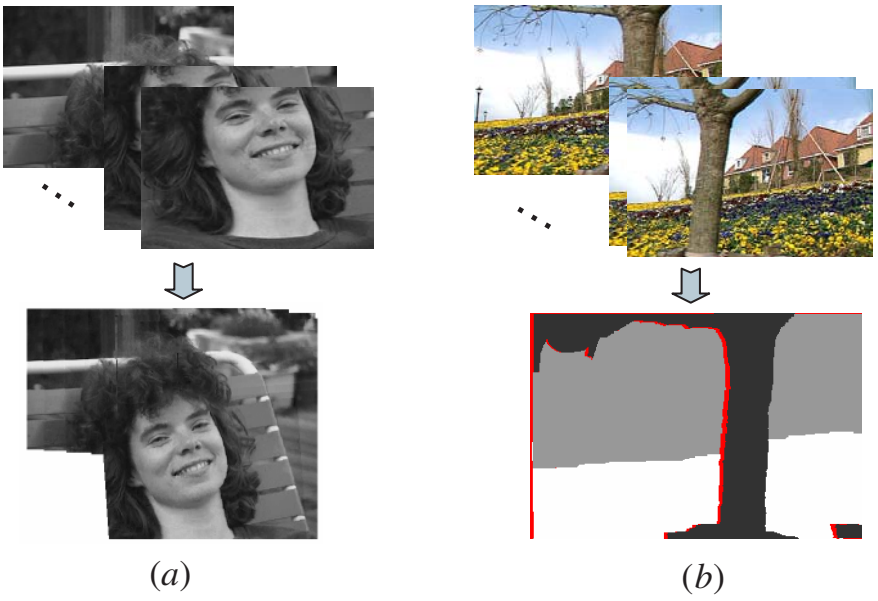


Fig. 1. Depending on the scene, a video sequence can be represented by one layer (**a**) or multiple layers (**b**). **a** This scene can be approximated by one plane due to the nonparallax camera motion and the generated mosaic. **b** "Flower garden" sequence, which can be represented by three layers: tree, garden, and background

with wide-baseline images, the initial estimation usually is quite difficult.
2. How do we deal with outlier regions when the images are taken at different times? These regions may look different due to appearing and disappearing objects, such as moving objects, shadows, and vegetation. Therefore, only a part of the image may be useful for the registration.
3. How do we handle complex motion models in a single 3D scene, e.g., multiple homographies as shown in Fig. 1b? Most existing approaches ignore these problems and attempt to align the whole image using a single motion model regardless of the number of layers.

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 metadata. In particular, given an image sequence of a mission or inspection video, we want to register it to a reference image, which may be taken at a different time, location, and orientation. The proposed approach first uses a motion layer extraction algorithm [22] to obtain an accurate layer segmentation of the mission video by exploiting spatiotemporal information. For each layer, a mosaic is generated and the relative transformation parameters between consecutive frames are estimated. Then, we formulate the image-registration problem into a partitioning framework, where the overlapping regions between two images are partitioned into supporting and nonsupporting regions for the registration. In this framework, a region expansion process is designed to adaptively propagate the alignment process from the high-confidence seed regions to the low-confidence areas and simultaneously remove outlier regions. In order to obtain such starting seed regions, we apply a wide baseline algorithm [21] to compute a set of reliable seed correspondences between the first mission frame and reference image. Then, starting from the seed regions, the initially aligned areas are expanded to the whole overlapping areas using a graph-cut algorithm integrated with the level set representation of the previous regions. Consequently, we achieve a robust layer alignment for each layer using the relative-motion parameters estimated by the layer mosaic, and the final multilayer video registration is obtained after back projection of layers.

In the remainder of this paper, we describe how to generate mosaics after extracting layers from the video in Sect. 2. Section 3 presents the region expansion algorithm for layer registration. In Sect. 4, we demonstrate the experimental results for single- and multiple-layer video registration to illustrate the robustness of our approach.

2 Layer mosaics

In a planar scene, only one layer is available, as shown in Fig. 2a. It is easy to generate a mosaic for this layer using an affine or projective motion model. However, if the scene contains multiple layers, the motion models can vary from a simple global motion model to multiple motion models, where pixel motions are mapped to several parameter clusters. Figure 2b shows one example of this case from a "door-wall" sequence, which contains two layers. It is impossible to obtain one mosaic using this mission video without severe misalignment or distortion. Fortunately, in the context of video registration, temporal information is available in the mission video sequence, from which the motion layers of the scene can effectively be extracted. In this paper, we use a multiframe graph-cut framework [22] to achieve an accurate layer segmentation of the mission video sequence. Figure 3 shows the video segmentation results obtained for the "door-wall" sequence. After the motion segmentation, we obtain precise supporting regions for each layer and the corresponding motion parameters between each consecutive frame, which can be used as the initial parameters for layer mosaicing.

2.1 Mosaic generation

Since the gap between consecutive frames of a video sequence is small, it is better to use an intensity-based registration

(a) (b)

Fig. 2. a Three frames from a mission video are shown on the *left* that are to be registered to a single layer in the reference image shown on the *right*. **b** Three frames from a mission video are shown on the *left* that are to be registered to two layers in the reference image shown on the *right*

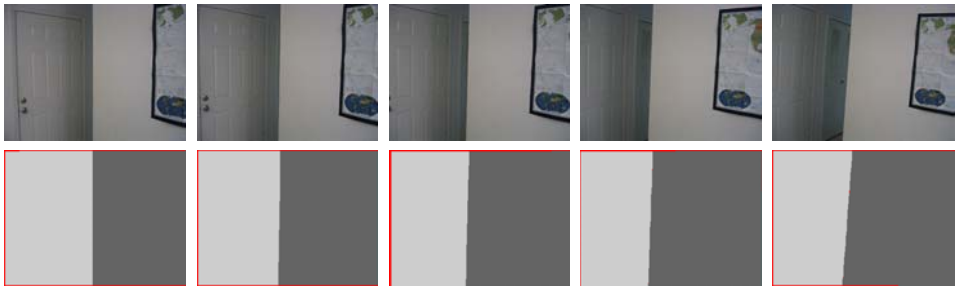


Fig. 3. Segmentation of "door-wall" sequence. *Top*: Several frames of the mission video sequence. *Bottom*: Segmentation results (occluded pixels in *red*)

method to minimize the image residue (or SSD), which can be written as

$$\epsilon = \sum_{\Omega} [(I_2(\mathbf{H}\mathbf{x}) - I_1(\mathbf{x})]^2 \,, \tag{1}$$

where I_1 and I_2 are two original images, Ω is the overlapping area between two consecutive frames, $\mathbf{x}$ represents homogenous pixel coordinates, and $\mathbf{H} \in \mathbb{R}^{3\times 3}$ is a homography matrix between two frames. Starting with $\mathbf{H} = \mathbf{I}$ (identity matrix), a nonlinear approach, such as Levenberg–Marquardt method, can be used to iteratively minimize the residue [17]. In this method, after computing image gradient $\bigtriangledown I$ from the two images, a gradient-descent direction is estimated that leads to a local minimum. However, due to changes in illumination and camera gain, the images may appear too bright or dark and the contrast of the images may be too weak to provide good gradient information for the minimization. Also, if the images do not have enough texture information, such as the door layer shown in Fig. 4a, the intensity map may be too flat. As a result, there is not much variation in the gradient map, and the registration process will be more sensitive to noise. Therefore, before generating the mosaic, we adjust the brightness of the images and enhance image contrast to obtain good gradient information for the minimization.

A simple way of enhancing image contrast is to normalize or equalize the image. Here we use a nonlinear equalization technique to distribute the histogram into the full-intensity range and enhance the gradient variation. Then, a bilateral filter [18] is applied to denoise the image, and the histogram is further smoothed and more uniformly distributed in the full range of intensity. Consequently, not only is the image contrast enhanced, but the gradient field also becomes continuous while the salient features are still retained (Fig. 4b).

Figure 5 compares the results of mosaic generation with and without image enhancement. For both "door" and "wall" layers, without applying the image enhancement, the image contrast and gradient information is not sufficient to get correct mosaics, as shown in Fig. 5a and c.

The transformation H_i between mission frame i with the mosaic becomes known after a mosaic is generated for each layer. Therefore, we have two choices when it comes to aligning the mission video to the reference image as shown in Fig. 6. In the first scheme, after aligning the layer mosaic to the reference image with transformation F, an initial transformation for a mission frame i to the reference image can be computed by $T_i = FH_i$. However, in this scheme, the error between frame f_i with frame f_1 will be accumulated with i increasing, which may not provide a good estimation between the layer mosaic and the reference image.

In this paper, we use an alternative solution, whereby each frame f_i is directly registered to the reference image based on the previous transformation of frame f_{i-1}. First, we align the first frame to the reference image and estimate its corresponding transformation T_1 by determining corresponding seed regions and using the region expansion approach, which will be described in the next section. Then the initial transformation for the second frame can be simply computed by $T_2 = T_1 H_1^{-1} H_2$, which can be further refined only using the region expansion process without computing seed correspondences. After estimating the precise transformation T_{i-1} of f_{i-1}, we iteratively compute the initial transformation for frame i by $T_i = T_{i-1} H_{i-1}^{-1} H_i$ using the previous frame f_{i-1}. As a result, we can avoid the accumulated error of the mosaic since the initial transformation is always computed employing the previous frame f_{i-1} instead of the first frame f_1. Hence, before registering the whole mission video sequence, we have to align frame 1 to the reference image and compute T_1. In the next section, we will present a novel solution for this critical frame registration.

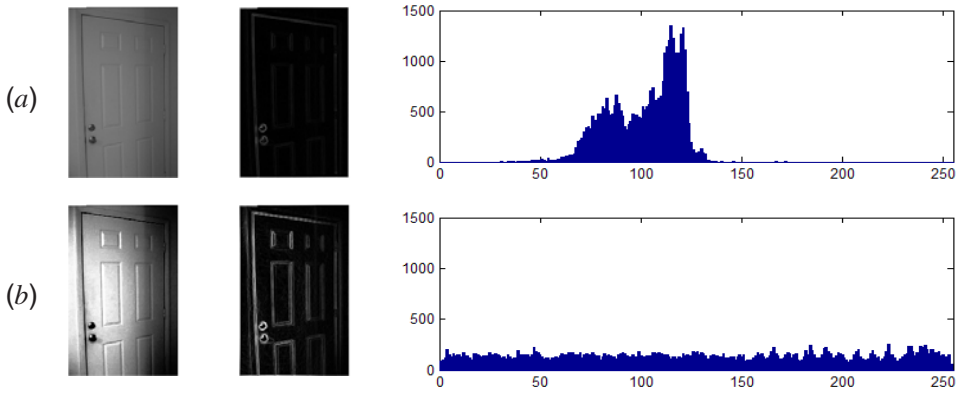


Fig. 4a,b. Image enhancement. **a** Original image of a door layer from sequence shown in Fig. 3 with gradient magnitude map and histogram. **b** Same as **a** after equalization and bilateral filtering. The histogram becomes more uniformly distributed, while the image contrast and gradient map are significantly enhanced

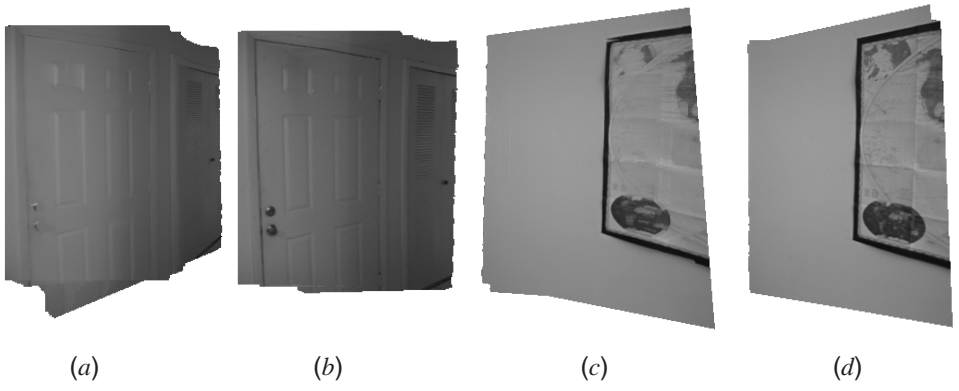


Fig. 5a–d. Image mosaics. **a,c** Mosaics of "door" and "wall" layers without image enhancement. **b,d** Mosaics after applying image enhancement. Compared to **b** and **d**, **a** and **c** contain some apparent distortions

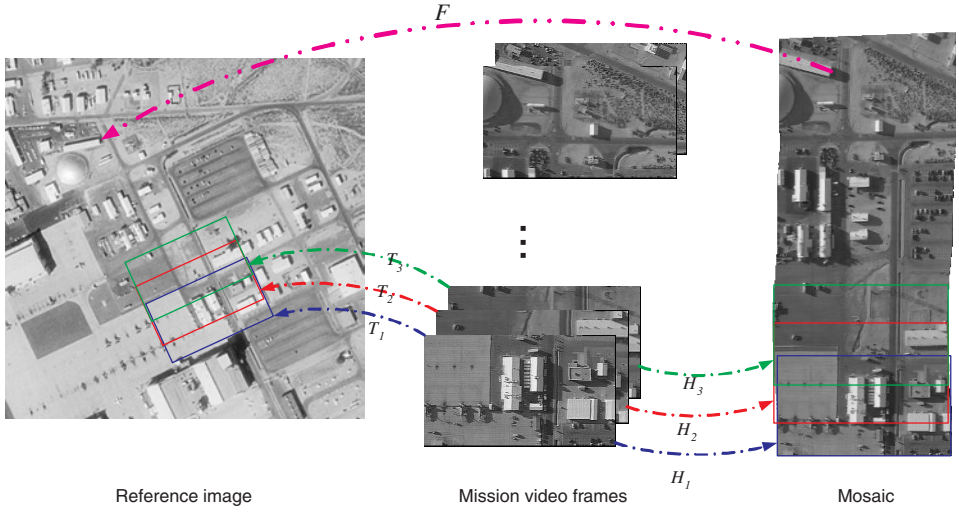


Fig. 6. The transformation among mission frames, reference image, and layer mosaic for one georegistration sequence. H_i is the transformation between mission frame i with mosaic. F is the transformation between mosaic and reference image. T_i is the final transformation between mission frame i and reference image

3 Layer registration by region expansion

The first issue that has to be tackled for layer registration algorithm development is the decision to use either sparse or dense image features for registration. Given two wide-baseline images without any metadata, it is difficult to perform alignment due to illumination variations and large motion between two images. Therefore, the use of sparse image features is ideal for the fast estimation of initial motion parameters. However, due to outliers and inaccuracy of these correspondences, the initial registration is usually not good enough. In this section, we propose a two-stage approach to integrating the merits of the sparse and dense image features. 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 same time, the outlier regions, such as appearing/disappearing objects that may harm the registration process, are detected and removed.

3.1 Determining correspondences

There are several methods for computing robust correspondences for wide-baseline images [4, 21]. Here we use Xiao and Shah's work [21] to determine a set of reliable corresponding corners. In this approach, a set of edge-corners is

identified in both images that provide robust and consistent matching primitives. Our feature detector first identifies Harris corners and then removes those corners that are not at the intersection of two edge lines. As a result, all the edge-corners are located at the junctions of multiple edges. Then the supporting regions of the corresponding corners should satisfy the following equation:

$$\mu I_2(\mathbf{A}\mathbf{x}+\mathbf{d})+\delta = I_1(\mathbf{x}), \tag{2}$$

where I_1 and I_2 are two original images, $\mathbf{x} \in \mathbb{R}^2$ represents pixel coordinates, $\mathbf{A} \in \mathbb{R}^{2\times 2}$ is an affine matrix, $\mathbf{d} \in \mathbb{R}^2$ is a translation vector, μ depends on the reflection angle of the light source, and δ depends on the camera gain. The illumination coefficients, μ and δ, are used to compensate for 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. We can compute the best match by minimizing the following residue:

$$\epsilon = \sum_{\Omega}[(\mu I_2(\mathbf{A}\mathbf{x}+\mathbf{d})+\delta) - I_1(\mathbf{x})]^2, \tag{3}$$

where Ω is the image patch and the patch size is 40×40. Instead of initializing the minimization with $\mathbf{A} = \mathbf{I}$ (identity matrix), $\mathbf{d} = [0\ 0]^{\mathrm{T}}$, $\mu = 1$, and $\delta = 0$, we first search the best rotation and scaling component implied in this affine transformation; we then apply Newton–Raphson iteration from $\mathbf{A} = \mathbf{R} \times \mathbf{S}$ (best rotation and scaling) to obtain stable results [21]. Simultaneously, an optimal affine transformation $\mathbf{A}$ and $\mathbf{d}$, and illumination coefficients μ and δ are obtained, which not only compensates for the geometric deformation but also efficiently adjusts the brightness and contrast of I_2.

Figure 7 shows the detailed matching process for a small patch. After computing the best affine transformation and illumination coefficients between the two patches in mission and reference images, the patch in I_2 (Fig. 7d) is warped as shown in Fig. 7e, which has a similar appearance to the patch in I_1 (Fig. 7c), and the residue is minimized.

3.2 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 small square region with its corresponding affine transformation. This implies that the corresponding small square regions in the mission and reference images are aligned by this initial affine transformation. For a general minimization case, we use a vector, Θ, to denote all the parameters used to minimize the errors between two images. The image dissimilarity function can be rewritten as:

$$\epsilon = \sum_{\Omega}[I_2(\mathbf{x},\Theta) - I_1(\mathbf{x})]^2, \tag{4}$$

where Θ includes two parts: the first part deals with illumination coefficients μ and δ, and the second part deals with motion parameters (i.e., affine or projective transformation parameters). Using linear [21] or nonlinear [17] optimization algorithms, ϵ can be iteratively minimized for this small patch (e.g., 41×41) and the corresponding Θ can be estimated. Nevertheless, this minimization process may create two problems. First, the estimated parameters obtained by using the small patch may overfit the pixels inside the region and may not correctly represent the global transformation of a larger region. Second, this process ignores the appearing/disappearing objects between two images, such as the moving objects, occlusion areas, and shadows.

To overcome the problems described above, we expand the region boundary to obtain more supporting pixels that are consistent with the motion parameters and also to identify the outlier pixels. Then we iteratively refine the motion parameters using these supporting pixels. Therefore, this registration

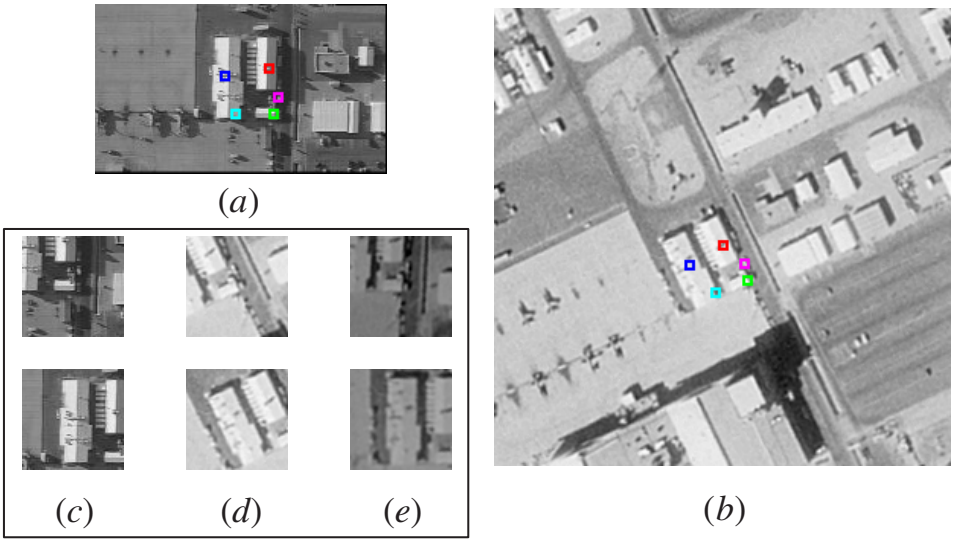


Fig. 7a–e. Determining correspondences between the mission frame and reference image. **a** Mission image. **b** Small part of reference image. Several correspondences are computed by the wide-baseline matching algorithm, each pair of correspondences is marked by *squares* with the same color. **c–e** Matching process of *green* (*top row*) and *blue* (*bottom row*) corners. **c** A patch from **a**. **d** Corresponding patch from **b**. **e** Warped patch **d** obtained after applying the best affine transformation, where patch **e** is similar to patch **c**. NB: Compared to the original patches **c** and **d**, the illumination effect is partially compensated between **c** and **e** by estimating μ and δ

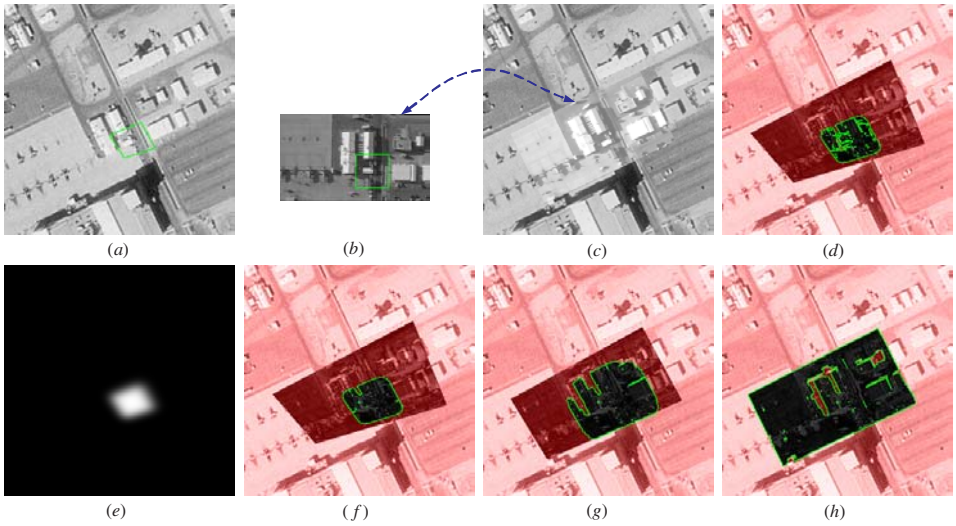


Fig. 8a–h. Region expansion process. **a,b** Initial corresponding patch contours in the reference and mission images, respectively. **c** Final registration result, where the intensities of the embedded mission image are adjusted by illumination coefficients μ and δ. **d** Simple expansion and partitioning started from the initial contour shown in **a**. **e** Level set representation of initial contour **a**. **f–h** Intermediate results using graph-cut method with the level set representation, which can guarantee that the expansion gradually evolves from the center to a boundary. NB: The *green boxes* in **a** and **b** are the initial seed regions. **f–h** Difference images between the warped **b** and **a** and the *green* contours in **f–h** are supporting region boundaries obtained after using a bipartitioning algorithm. The nonsupporting pixels are masked by *red*

problem is essentially converted into a partitioning problem that can be stated as follows: Determine the optimal supporting regions and their corresponding motion parameters for image registration.

One straightforward solution is to apply a threshold on ϵ between the original and warped windows to detect additional supporting pixels in a neighboring area of the previous region. However, the expanded region is sensitive to noise and may not be compact and smooth. Figure 8d shows one result obtained by using this simple scheme. Thus, we propose a novel alternative approach to gradually expanding the seed region by identifying the supporting pixels using a bipartitioning graph-cut method integrated with a level set representation. First, a smoothness energy term between neighboring pixels is introduced that maintains piecewise smoothness of the partitions [2,9]. 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 recast into the graph-cut framework. In this framework [2], we seek the labeling function f that partitions the pixels in region Ω into two groups: the first group represents the supporting regions, labeled $f = 0$; the other represents the outlier regions, labeled $f = 1$. This partitioning can be achieved by minimizing the following energy function:

$$E = \sum_{(p,q)\in\mathcal{N}} V(p,q) + \sum_{p\in\Omega} D_p(f_p)\,, \tag{5}$$

where the first term is a piecewise smoothness term, the second term is a data penalty term, $\mathcal{N}$ is a 4-neighbor system, and f_p is the label of a pixel p. $D_p(f_p)$ can be approximated by a Heaviside function:

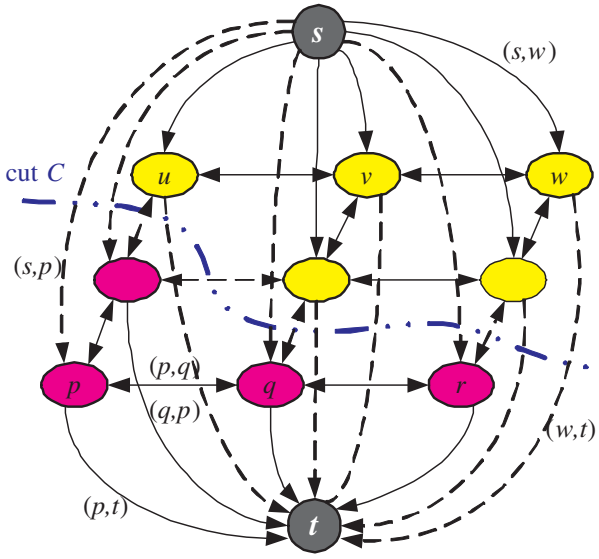
$$D_p(f_p) = \begin{cases} \tan^{-1}(d_p - \tau) + \pi/2 \text{ if } f_p = 0, \\ \pi/2 - \tan^{-1}(d_p - \tau) \text{ if } f_p = 1, \end{cases} \tag{6}$$

where $d_p = [I_2(\mathbf{x}_p, \Theta) - I_1(\mathbf{x}_p)]^2$, $\mathbf{x}_p$ is the pixel coordinates of p and τ is an empirical threshold. $V(p,q)$ is designed to more likely maintain the same label for p and q if they have similar intensities, such as:

$$V(p,q) = \begin{cases} 3\lambda \text{ if } \max(|I_1(p) - I_1(q)|, |I_2(p) - I_2(q)|) < 8, \\ \lambda \quad \text{otherwise.} \end{cases}$$

To minimize the energy function, a weighted graph $\mathcal{G} = \langle \mathcal{V}, \mathcal{E} \rangle$ is constructed as shown in Fig. 9, 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 using the table shown in Fig. 9, we can compute a minimum cut $\mathcal{C}$ using a standard graph-cut algorithm and partition the original region into the supporting and outlier regions. However, using this process we cannot expand the region from the initial seed patch to the exterior to obtain more supporting pixels. Hence, we must use the contour of the previous seed region prior to computing the level set representation for this region [10,12], which allows the region contour to evolve along the normal direction. After enforcing the level set regulation on the sink-side weight of graph $\mathcal{G}$, we can effectively control the graph-cut algorithm to gradually expand the seed region.

Figure 8 shows a detailed expansion process starting from one initial seed region. Figures 8a and b show the initial contours of the corresponding seed regions. Based on the initial contour of the original seed region Ω^0 (Fig. 8b), 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. 8e) 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 weight of the sink-side t-link for each pixel, such that the weights of the pixels inside the region are almost



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

Fig. 9. Example of a weighted graph $\mathcal{G}$ for an image. Nodes p, q, r, $\cdots$, w are the pixels in the image. s is the source node and t is the sink node. The edges connected to the 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 nonsupporting pixels (sink). The *dotted links* are crossed by $\mathcal{C}$. Table 1 shows the weights

unchanged 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 exclude the outside pixels and label them as the nonsupporting pixels for this region. This way, the new expanded supporting region Ω^1 can be computed as shown in Fig. 8f. After several iterations as shown in (Figs. 8f–h), the region's boundary gradually propagates from the center to the exterior until it reaches the overlapping boundary of two images, and the alignment is stable. Figure 8h shows the final region Ω^5 after five iterations, and Fig. 8c shows the final registration results using the projective transformation computed by this approach.

If several initial seed regions share the same motion transformation for some layer, we expand the multiple regions simultaneously to speed up the registration process. We show one such example in Sect. 4. Figure 8 shows that our approach can obtain the piecewise smooth region expansion, which is insensitive to noise. The outlier regions due to shadows are also detected and removed. At the same time, the transformation T_1 for the key frame is estimated. After applying the initial transformation $T_i = T_{i-1}H_{i-1}^{-1}H_i$ to frame i, we initialize the alignment of frame i to the reference image. Then, employing the region expansion approach to the ith frame, we remove outliers and refine the alignment to compute the transformation T_i for this frame. The final video registration results are shown in Fig. 10.

4 Experiments

We performed several experiments on different real data sets, where the metadata information was not available. In all of the experiments, we applied the wide-baseline matching algorithm to estimate sparse correspondences, which can provide an approximated initial alignment between mission and reference images. For a single layer registration, after determining the sparse correspondences, we expanded these seed regions simultaneously to speed up the alignment process. The initial homography between two images could be computed in two ways: select the most robust affine transformation of the seed regions using the RANSAC technique or estimate a homography voted on by all of these correspondences.

In Fig. 11, we show an example of the multiseed expansion process. Since a number of correspondences are determined, it is easy to estimate a robust initial homography using all the correspondences. Then, starting with the initial homography, we expand all the initial seed regions simultaneously until the overlapping areas between the mission and reference images are covered. Our graph-cut algorithm also detects and removes the outlier regions, most of which are due to vegetation or shadows. Figures 11h,i compare the zoomed results before and after applying the region expansion process.

Figure 12 shows another set of results for georegistration using single seed region expansion where only three correspondences are determined due to the small size of the mission frame. Since we cannot obtain a good initial projective transformations from these few correspondences, we use RANSAC to determine the robust affine transformation of the seed regions, which is shown in blue in Figs. 12a,b. Then, starting from one seed region (blue), we perform the adaptive region expansion alignment and obtain the registration results as shown in Figs. 12c,d.

Figure 13 shows the final registration results for the "doorwall" sequence. After obtaining the layers for each frame, we align the different layers to the reference image separately using the adaptive region expansion approach. The final registration results of the first frame are shown in Figs. 13d,e. Compared to the direct registration, our approach has two advantages. First, to align the corresponding layers, we employ different sets of motion parameters to correctly represent the mapping of the pixels in these layers. Second, the layer segmentation also provides accurate supporting regions for each layer, which prevent the region expansion process across the layer boundaries. Therefore, for each layer registration, our approach can effectively avoid the pixels from the other layers and achieve more accurate aligned regions for each layer.

In all of our experiments, after determining correspondences, the computational time for a single layer registration

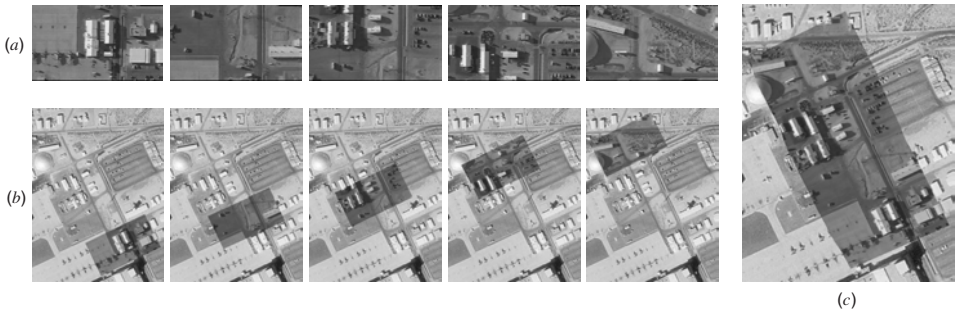


Fig. 10a–c. Video registration results. **a** Mission video frames. **b** Registration results for several frames, where the mission images are superimposed on the reference image. **c** Full registration of all mission video frames

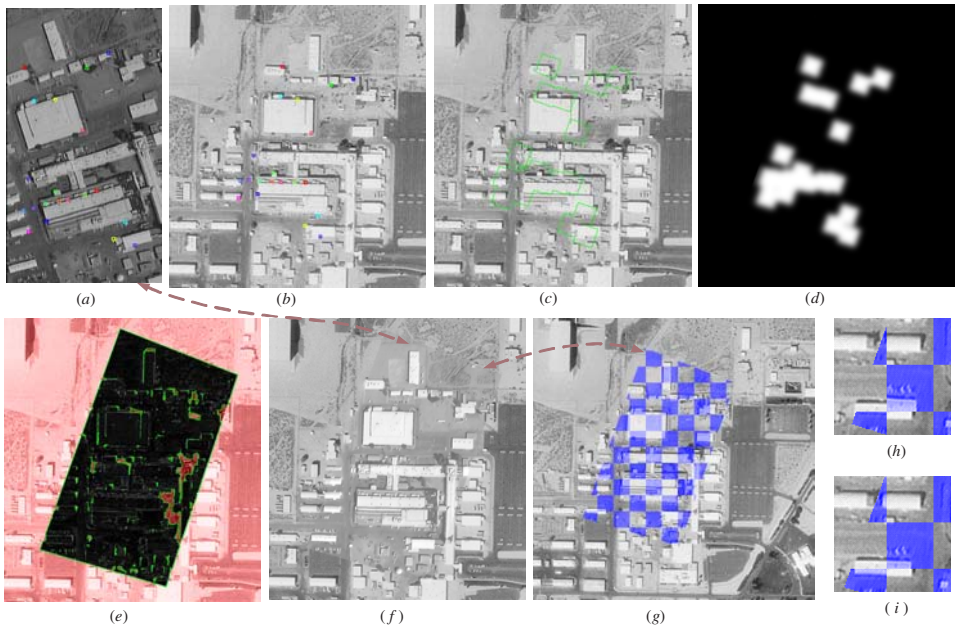


Fig. 11a–i. Registration using multiseed region expansion. **a** Mission image. **b** Small part of a reference image. The correspondences are marked in **a** and **b** by the same colors. **c** Initial seed regions and **d** corresponding level set representation. **e** Final region contour after expansion, where the nonsupporting regions are indicated by *red*. **f** Registration results. **g** Checkered display after alignment. **h,i** Zoomed alignment results before and after applying the region expansion alignment

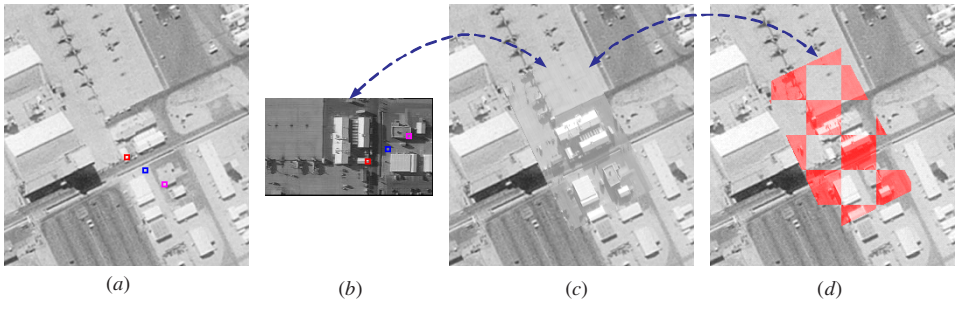


Fig. 12a–d. Registration using one seed region expansion. **a** Small part of a reference image. **b** Mission image. **c** Registration results. **d** Checkered display after alignment

is less than 10 s per frame. NB: All of the results are available at our Web site.[1]

5 Conclusions

In this paper, we presented a layer-based framework for video registration without any metadata. After extracting layers from the video sequence, layer mosaics are built. Then, an adaptive region expansion algorithm is used to efficiently propagate the alignment process from the high-confidence areas to the low-confidence areas. In addition, the outlier regions are also

[1] `http://www.cs.ucf.edu/~vision/projects/layer_registration/`

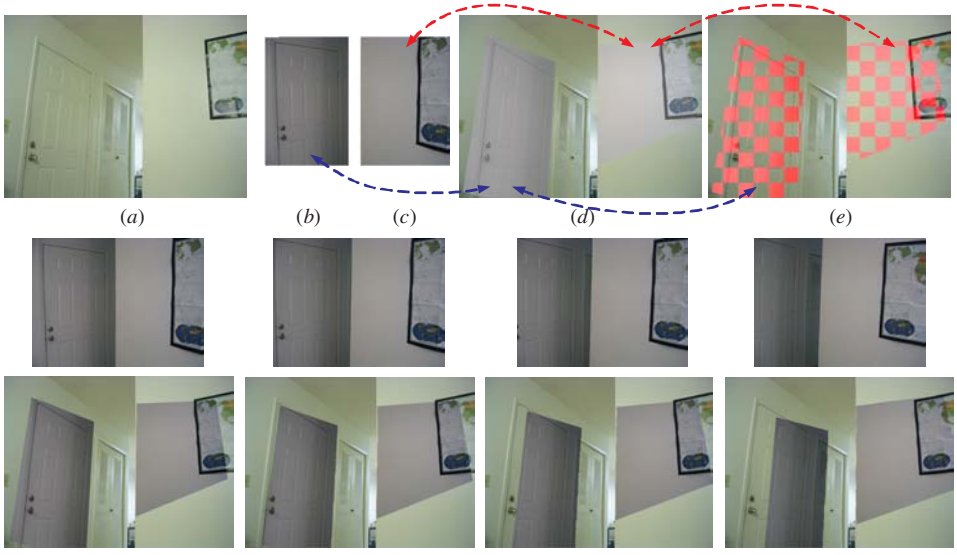


Fig. 13a–e. Multiple-layer registration. **a** Reference image. **b,c** Layers of door and wall in frame 1. **d** Registration results of frame 1. **e** Checkered display after alignment. *Middle*: Some mission video frames. *Bottom*: Corresponding video registration for these frames, where the mission images are split into two parts during the registration

identified and removed. Based on an initial transformation from the mosaic, each video frame is aligned to the reference image and finally combined together to achieve multilayer video registration.

In the future, we would like to investigate different dissimilarity measures to handle other kinds of registration problems, such as multisensor image registration and 2D image to 3D model registration.

References

1. Ayer S, Sawhney H (1995) Layered representation of motion video using robust maximum-likelihood estimation of mixture models and MDL encoding. In: International conference on computer vision
2. Boykov Y, Veksler O, Zabih R (2001) Fast approximate energy minimization via graph cuts. IEEE Trans Pattern Anal Mach Intell 23(11):1222–1239
3. Brown L (1992) A survey of image registration techniques. ACM Comput Surv 24(4):325–376
4. Ferrari V, Tuytellars T, Van Gool L (2003) Wide-baseline multiple-view correspondences. In: IEEE conference on computer vision and pattern recognition
5. Horn B, Schunck B (1981) Determining optical flow. Artif Intell 17:185–203
6. Ke Q, Kanade T (2002) A robust subspace approach to layer extraction. In: IEEE workshop on motion and video computing
7. Keller Y, Averbuch A (2003) Implicit similarity: a new approach to multi-sensor image registration. In: IEEE conference on computer vision and pattern recognition
8. Khan S, Shah M (2001) Object based segmentation of video using color, motion and spatial information. In: IEEE conference on computer vision and pattern recognition
9. Kolmogorov V, Zabih R (2002) What energy functions can be minimized via graph cuts? In: European conference on computer vision
10. Osher S, Fedkiw R (2003) Level set methods and dynamic implicit surfaces. Springer, Berlin Heidelberg New York
11. Sawhney H, Hsu S, Kumar R (1998) Robust video mosaicing through topology inference and local to global alignment. In: European conference on computer vision
12. Sethian J (1999) Level set methods and fast marching methods. Cambridge University Press, Cambridge, UK
13. Sheikh Y, Shah M (2004) Aligning 'dissimilar' images directly. In: Asian conference on computer vision
14. Sheikh Y, Khan S, Shah M, Cannata R (2003) Geodetic alignment of aerial video frames. In: Video registration, video computing series. Kluwer, Dordrecht
15. Shen D, Davatzitos C (2002) HAMMER: Hierarchical attribute matching mechanism for elastic registration. IEEE Trans Med Imag 21:1421–1439
16. Shah M, Kumar R (eds) (2003) Video registration. Kluwer, Dordrecht
17. Szeliski R (1996) Video mosaics for virtual environments. IEEE Comput Graph Appl 16:22–30
18. Tomasi C, Manduchi r (1998) Bilateral filtering for gray and color images. In: International conference on computer vision
19. Wills J, Agarwal S, Belongie S (2003) What went where. In: IEEE conference on computer vision and pattern recognition
20. Wildes R, Hirvonen D, Hsu S, Kumar R, Lehman W, Matei B, Zhao W (2001) Video georegistration: algorithm and quantitative evaluation. In: International conference on computer vision
21. Xiao J, Shah M (2003) Two-frame wide baseline matching. In: International conference on computer vision
22. Xiao J, Shah M (2004) Motion layer extraction in the presence of occlusion using graph cut. In: IEEE conference on computer vision and pattern recognition
23. Zheng Q, Chellappa R (1993) A computational vision approach to image registration. IEEE Trans Image Process 2(3):311–326
24. Zitova B, Flusser J (2003) Image registration methods: a survey. Image Vis Comput 21:977–1000

Jiangjian Xiao received the M.S. and Ph.D. degrees in computer science, both from University of Central Florida (UCF), Orlando, USA, in 2001 and 2004 respectively. He also received the B.S. degree in material engineering and M.S. degree in automatic control both from Beijing University of Aeronautics and Astronautics, Beijing, P.R. China, in 1994 and 1997 respectively. Currently he is a research assistant of computer vision lab at UCF. His research interests include wide baseline matching, multi-view synthesis, motion segmentation, video registration, video synthesis, and image-based rendering.

Mubarak Shah, a professor of Computer Science, and the founding director of the Computer Vision Laboratory at University of Central Florida (UCF), is a researcher in computer vision. He is a co-author of two books "Video Registration", (2003) and "Motion-Based Recognition", (1997), both by Kluwer Academic Publishers. He has supervised several Ph.D., M.S., and B.S. students to completion, and is currently directing twenty Ph.D. and several B.S. students. He has published close to one hundred fifty papers in leading journals and conferences on topics including activity and gesture recognition, violence detection, event ontology, object tracking (fixed camera, moving camera, multiple overlapping and non-overlapping cameras), video segmentation, story and scene segmentation, view morphing, ATR, wide-baseline matching, and video registration. Dr. Shah is a fellow of IEEE, was an IEEE Distinguished Visitor speaker for 1997–2000, and is often invited to present seminars, tutorials and invited talks all over the world. He received the Harris Corporation Engineering Achievement Award in 1999, the TOKTEN awards from UNDP in 1995, 1997, and 2000; Teaching Incentive Program award in 1995 and 2003, Research Incentive Award in 2003, and IEEE Outstanding Engineering Educator Award in 1997. He is an editor of international book series on "Video Computing"; editor in chief of Machine Vision and Applications journal, and an associate editor of Pattern Recognition journal. He was an associate editor of the IEEE Transactions on PAMI, and a guest editor of the special issue of International Journal of Computer Vision on Video Computing.