Department of Computer Science and Engineering, Texas A&M University, Technical Report tamu-cs-tr-2011-12-1

An Iterative Image Registration Technique Using a Scale-Space Model

Joseph Lee
CSE Department
Texas A&M University
jslee@cse.tamu.edu

S. Susan Young
U.S. Army Research Laboratory
Adelphi, MD
shiqiong.susan.young@us.army.mil

Ricardo Gutierrez-Osuna
CSE Department
Texas A&M University
rgutier@cse.tamu.edu

Abstract

Registration between two images is a key problem in computer vision. Current methods tend to separate the scale estimation process from translation and rotation estimation. This is due to the fact that the scale parameter is inherently related to the image resolution. In this paper, we present an area-based image registration technique that can simultaneously estimate translation, rotation, and scale parameters and take into account differences in resolution between two images. We first develop a scale-space model that relates the entire reference image pixels to a single observed image pixel with a scale parameter. This model is then easily generalized to include x-y translation and rotation parameters. By embedding this scale-space model into a non-linear least squares method, we can iteratively estimate the four registration parameters (x-y shift, rotation, and scale) in a unied manner. We test the validity of the proposed method on both simulated and real image data.

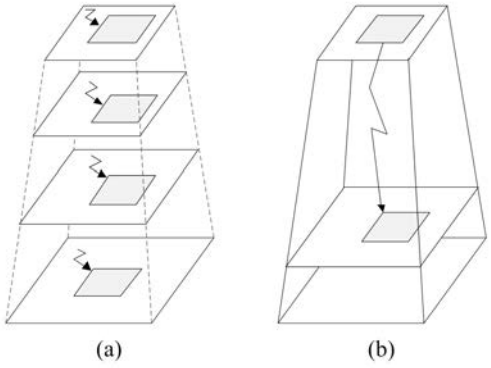

Figure 1. Comparison between the pyramid and scale-space approach for image matching. In the conventional approach (a) translation and rotation parameters are estimated at each level; scale estimation is performed independently. In the proposed approach (b) a scale-space model is embedded in the estimation process; this allows simultaneous estimation of translation, rotation, and scale parameters.

1. Introduction

Image registration is a crucial step in a variety of computer vision tasks, from image stitching to super-resolution and object recognition. In general, registration aims to align two images (the *reference* image and the *observed* image) by estimating translation, rotation, and scale. However, because of the inherent relationship between scale and image resolution [4], current methods tend to separate scale estimation from translation and rotation estimation.

To find a match between two images, conventional methods first construct an image pyramid by downscaling the reference image [1]. Then, an observed image is matched at each level by searching the translation and rotation parameters. The level at which the best match is found determines the scale factor between the reference and observed image. The disadvantage of this approach is that (1) the scale can be estimated at only fixed levels and (2) the scale is estimated independently from other registration parameters. In this paper, we present an iterative image registration technique that can estimate translation, rotation, and scale in a unified manner. This is achieved by embedding a *scale-space model* into a non-linear least squares framework. The approach is illustrated in Figure 1.

Image registration can be broadly categorized into area-based (or pixel-based) and feature-based methods [10]. *Area-based* methods work by directly matching pixel intensities of the two images as a whole. In contrast, *feature-based* methods extract higher-level structures from the two images and find the corresponding features to perform registration. Thus, feature-based methods are useful when corresponding features can be reliably detected [10]. However, if the reference image and the observed image are both in low-resolution, corresponding feature points may not be

detected accurately. For this case, our area-based method can also be used to register low-resolution images with sub-pixel accuracy.

This paper is organized as follows. Section 2 reviews previous image registration methods that incorporate scale estimation. Section 3 describes the proposed model for simultaneous estimation of scale, translation and rotation parameters; the method relates a reference image to an observed image with a single scale parameter and incorporates translation and rotation parameters. Section 4 describes how registration parameters are estimated using non-linear least squares. Experimental results on simulated and real data are shown in Section 5. We conclude this study and provide future work in the last section.

2. Related work

Different image representations have been used for scale estimation. A classic method for scale estimation is to use an *image pyramid* structure [1]. This method first builds a pyramid of reference images of decreasing scale, and then matches the observed image at each level. The level with the best match is taken as the scaling factor between the observed and reference image. Because the scale change at each level is discrete, the actual scale may not be precisely determined.

An improvement over pyramid representations is the *scale-space* representation [5]. Scale-space representations use a continuous scale parameter to express the possible scales of an image but, as a result, the search through the possible scale levels becomes prohibitive. Feature-based image registration methods such as [6, 4] use the scale-space concept in their approach. However, these methods still build an explicit pyramid structure that is discrete. Our method is able to efficiently search the scale parameter using the continuous scale-space model and avoid building a predetermined pyramid structure for the reference image.

Besides the pyramid-like representations, other image representations have also been used to address the scale estimation problem. In [8], scale estimation for object distance measurement was performed using a wavelet transform. Correlation techniques such as in [9, 2] use polar representations which enable rotation and scale invariance matching. Although, the polar representation can unify rotation and scale estimation, the translation component needs to be estimated separately. Our method estimates translation, rotation, and scale simultaneously in the Cartesian coordinates.

Although Lucas and Kanade [7] do not use the above image structures, their method provides a unified framework to simultaneously estimate translation, rotation, and scale parameters. In their method, the objective function is set as

$$J = \sum_{\mathbf{x}} [I_r(A\mathbf{x} + \mathbf{t}) - I_o(\mathbf{x})]^2$$

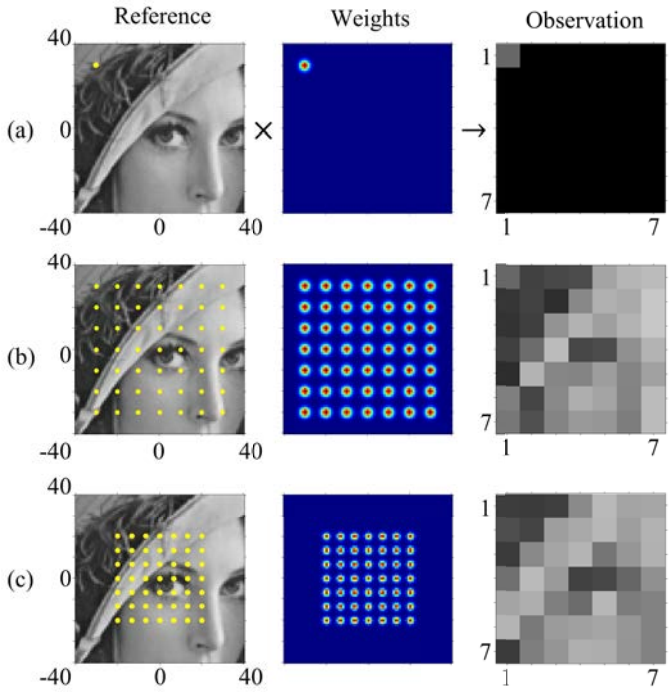

Figure 2. The relationship between a reference image and an observed image. (a) Each observation pixel is the weighted sum of reference pixels covered by the Gaussian window. (b) The observation image. (c) As the size of the Gaussian window G decreases, the number of effective reference pixels in $I_r(x, y)$ that are covered decreases.

where I_r and I_o are the reference and observed image, respectively. The $A\mathbf{x} + \mathbf{t}$ expresses the affine transformation of the two-dimensional coordinates of $\mathbf{x}$. However, this objective function does not consider the fact that when the scale changes, the image intensity for the corresponding pixels at x changes due to the difference in resolution. Our method is related to [7] in that we also use non-linear least squares to simultaneously estimate translation, rotation, and scale parameters. However, we embed the scale-space model into the objective function to handle images with different resolution. As in [7], we assume that a rough registration between the reference and observed image is provided.

3. Image scaling model

Consider the problem of registering two images $I_o(x, y)$ and $I_r(x, y)$, defined by their pixel intensities at index (x, y). We want to find the registration parameters that minimize the difference between I_r and I_o. Depending on the scale factor between these two images, I_r and I_o may be in different resolution. If we set I_r in a higher-resolution grid, we can define a transformation from $I_r(x, y)$ to $I_o(x, y)$ using a single scale parameter. This transformation can be further generalized to incorporate translation and rotation

parameters.

3.1. Image scale parameter

The relationship between the reference image and the observed image can be defined by

$$I_o(i,j) = \sum_{x,y} G(x,y;i,j,s) I_r(x,y), \tag{1}$$

where (x, y) represent the column and row indices of reference image I_r, and (i, j) are the column and row indices of image I_o. The scale parameter s determines the width of a two-dimensional Gaussian kernel G

$$G(x,y;x_0,y_0,\sigma) = \frac{1}{2\pi\sigma^2} e^{-\frac{(x-x_0)^2+(y-y_0)^2}{2\sigma^2}}, \tag{2}$$

where σ is the standard deviation and (x_0, y_0) is the mean vector. The model in equation (1) represents a single observation pixel as a weighted sum of all the pixels in the reference image. This is illustrated in Figure 2.

For convenience, we assume that the pixel width and height of the reference image $I_r(x, y)$ are set to unit length. Thus, if the observed image is scaled by a factor s relative to the reference image, an observed pixel will cover a width of $1/s$ in the reference image. Since most of the mass in a Gaussian density is contained within $\pm 3\sigma$, equating the observation pixel and the Gaussian window yields $6\sigma = 1/s$. Therefore, the standard deviation σ can be expressed in terms of the scaling factor s as:

$$\sigma = \frac{1}{6s}. \tag{3}$$

Assuming a rectangular image, the mean location of the $G(x, y; i, j, s)$ can be expressed in terms of the observed image coordinates (i, j) and the observed image pixel width $1/s$ as

$$\begin{bmatrix} x_0 \\ y_0 \end{bmatrix} = \begin{bmatrix} x_1 + (i-1)\frac{1}{s} + \frac{1}{2s} \\ y_1 + (j-1)\frac{1}{s} + \frac{1}{2s} \end{bmatrix} \tag{4}$$

where (x_1, y_1) are the left-top coordinates of the observed image[1]. Substituting expressions (3) and (4) into the kernel equation (2) yields

$$G(x,y;i,j,s) = \frac{1}{2\pi\left(\frac{1}{6s}\right)^2} e^{-\frac{\left(x-x_1-\frac{2i-1}{2s}\right)^2+\left(y-y_1-\frac{2j-1}{2s}\right)^2}{2\left(\frac{1}{6s}\right)^2}}. \tag{5}$$

Although the model equation (5) is continuous, the reference image is discrete. For an observed image pixel $I_o(x, y)$ to be a valid intensity, the Gaussian weights in equation (1) need to sum up to 1. However, as shown in Figure 3(a), if the number of reference pixels that is covered by the Gaussian kernel is too small, the weights will not add up to 1. Therefore, the reference image should be sufficiently upsampled as shown in Figure 3(b) so that σ is not too small.

[1] If an $m \times n$ image is centered at the origin $(0, 0)$, the left-top coordinates are $x_1 = -m/2$ and $y_1 = n/2$.

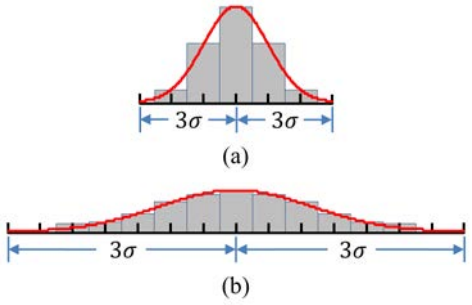

Figure 3. A one-dimensional view of the continuous Gaussian kernel on a discrete reference image. Only a few reference pixels are covered by the Gaussian kernel in (a) whereas enough reference pixels are covered by the Gaussian kernel that the sum of weights becomes close to 1.

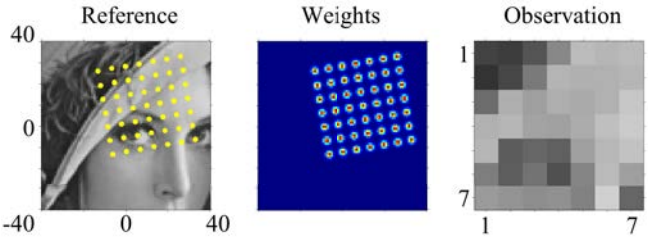

Figure 4. The relationship between the reference image and observation image. The observed image is a scaled, rotated, and translated version of the reference image.

3.2. Generalization to translation and rotation

As shown in Figure 4, the mean location (x_0, y_0) of the Gaussian window can be affinely transformed by

$$\begin{bmatrix} x_0' \\ y_0' \end{bmatrix} = \begin{bmatrix} \cos\theta & \sin\theta \\ \sin\theta & \cos\theta \end{bmatrix} \begin{bmatrix} x_0 \\ y_0 \end{bmatrix} + \begin{bmatrix} t_x \\ t_y \end{bmatrix}$$

where θ is the rotation parameter and t_x and t_y are the horizontal and vertical translation. Therefore, we can generalize equation (5) as

$$G(x,y;i,j,t_x,t_y,\theta,s) = \frac{1}{2\pi\left(\frac{1}{6s}\right)^2} e^{-\frac{(x-x_0')^2+(y-y_0')^2}{2\left(\frac{1}{6s}\right)^2}}.$$

Finally, the model equation (1) for the reference image and the observed image becomes

$$I_o(i,j) = \sum_{x,y} G(x,y;i,j,t_x,t_y,\theta,s) I_r(x,y). \tag{6}$$

$$
\begin{aligned}
I_o(1,1) &= G(1,1;1,1,\mathbf{x})I_r(1,1) + G(1,2;1,1,\mathbf{x})I_r(1,2) + \cdots + G(q,p;1,1,\mathbf{x})I_r(q,p) \\
I_o(1,2) &= G(1,1;1,2,\mathbf{x})I_r(1,1) + G(1,2;1,2,\mathbf{x})I_r(1,2) + \cdots + G(q,p;1,2,\mathbf{x})I_r(q,p) \\
I_o(1,3) &= G(1,1;1,3,\mathbf{x})I_r(1,1) + G(1,2;1,3,\mathbf{x})I_r(1,2) + \cdots + G(q,p;1,3,\mathbf{x})I_r(q,p) \\
&\vdots \\
I_o(n,m) &= G(1,1;n,m,\mathbf{x})I_r(1,1) + G(1,2;n,m,\mathbf{x})I_r(1,2) + \cdots + G(q,p;n,m,\mathbf{x})I_r(q,p)
\end{aligned} \tag{7}
$$

which, for a $p \times q$ reference image and an $m \times n$ observed image, yields the system of equations (7), where the column vector $\mathbf{x}$ denotes

$$\mathbf{x} = [t_x, t_y, \theta, s]^T.$$

4. Image registration algorithm

Given the model equation (6), image registration parameters (t_x, t_y, θ, s) are estimated through an iterative non-linear least squares algorithm[2]. Namely, we seek to find an estimate $\hat{\mathbf{x}}$ that minimizes the objective function

$$J = \frac{1}{2}[\tilde{\mathbf{y}} - \mathbf{g}(\hat{\mathbf{x}})]^T[\tilde{\mathbf{y}} - \mathbf{g}(\hat{\mathbf{x}})]$$

where $\tilde{\mathbf{y}}$ is the observed image in raster scan order and, through equation (6), $\mathbf{g}(\hat{\mathbf{x}})$ is a transformed image of the reference image with registration parameters $\hat{\mathbf{x}}$.

An initial value $\mathbf{x}_c$ for $\hat{\mathbf{x}}$ is required to start the estimation process. We assume that a good initial estimate is provided. For a given estimate $\mathbf{x}_c$, its goodness is computed by the error term

$$\Delta\mathbf{y}_c = \tilde{\mathbf{y}} - \mathbf{g}(\mathbf{x}_c)$$

and the Jacobian matrix, which expresses the linear change of the predicted image at current state $\mathbf{x}_c$, is computed as

$$
H = \begin{bmatrix}
\left.\frac{\partial g(1,1)}{\partial x_1}\right|_{\mathbf{x}_c} & \cdots & \left.\frac{\partial g(1,1)}{\partial x_4}\right|_{\mathbf{x}_c} \\
\left.\frac{\partial g(1,2)}{\partial x_1}\right|_{\mathbf{x}_c} & \cdots & \left.\frac{\partial g(1,2)}{\partial x_4}\right|_{\mathbf{x}_c} \\
\vdots & \ddots & \vdots \\
\left.\frac{\partial g(n,m)}{\partial x_1}\right|_{\mathbf{x}_c} & \cdots & \left.\frac{\partial g(n,m)}{\partial x_4}\right|_{\mathbf{x}_c}
\end{bmatrix},
$$

where $[x_1, x_2, x_3, x_4]^T = [t_x, t_y, \theta, s]^T$. Once $\Delta\mathbf{y}_c$ and H have been computed, the correction term, which gives the minimum error by the weighted least squares solution [3], can be expressed as:

$$\Delta\mathbf{x} = (H^T W H)^{-1} H^T W \Delta\mathbf{y}_c$$

[2]Note that although the weights $G(x, y)$ are linear in equation (7), the registration parameters t_x, t_y, θ, and s are non-linear terms.

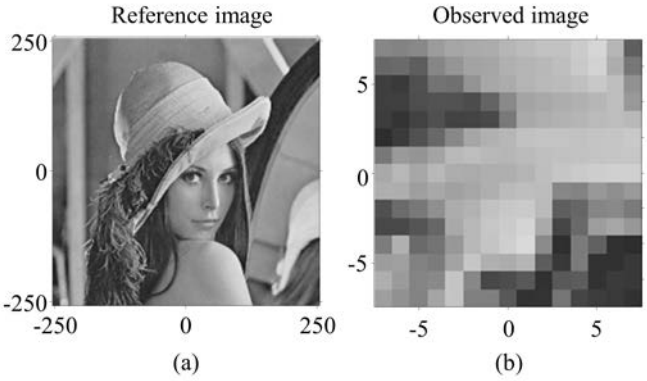

Figure 5. Reference (a) and observed image patch (b). The observed patch was obtained by down-sampling the 512×512 reference image by a factor of 0.1 followed by cropping. The observed image is 15×15.

where the weight matrix W allows us to emphasize certain observed pixels. As an example, if the region of interest we seek to register does not occupy the whole image, the weight matrix can be used as a mask. For the experiment in this study, we use an identity matrix which puts equal weight to each observed pixel.

With the correction term $\Delta\mathbf{x}$, the current state estimate $\mathbf{x}_c$ is iteratively updated by

$$\mathbf{x}_c = \mathbf{x}_c + \Delta\mathbf{x}.$$

The iterative process continues until a stopping condition is satisfied or after a fixed number of iteration is reached [3]. As an example, if the predicted residual at each iteration i is defined by

$$J_i = \Delta\mathbf{y}_c^T W \Delta\mathbf{y}_c,$$

the stopping criteria is given by

$$\frac{|J_i - J_{i-1}|}{J_i} < \frac{\varepsilon}{\|W\|},$$

where ε is a small value that determines the tolerance.

5. Experimental results

We will demonstrate the proposed technique on two case studies that require simultaneous estimation of translation,

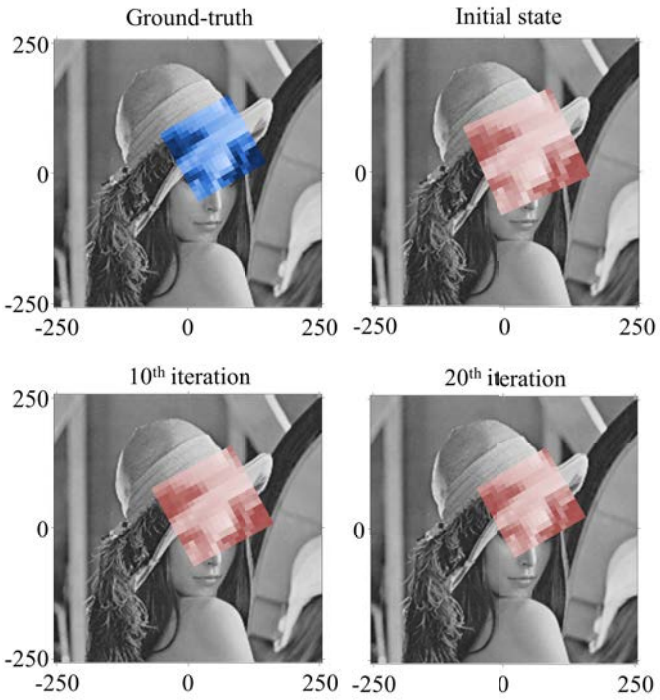

Figure 6. The ground-truth position of the observed image (blue) and its updated positions at the 10^{th} and 20^{th} iterations (red). In this case, the initial position starts with a 20% error (by multiplying 0.8 to the ground-truth values) for each registration parameter.

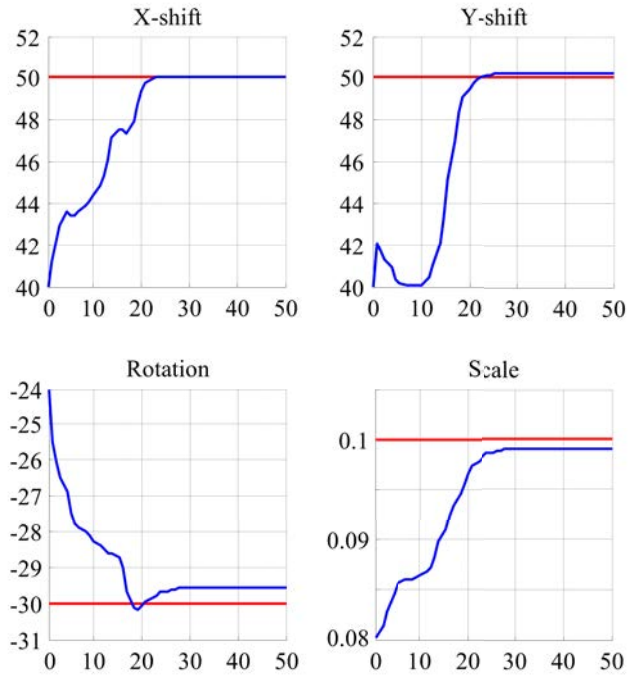

Figure 7. Estimated values for translation (x-shift, y-shift), rotation, and scale in each iteration (blue). The ground-truth values were $t_x = 50$, $t_y = 50$, $\theta = -30$, $s = 0.1$ (red). The initial estimate starts with a 20% error for each registration parameter.

rotation, and scale parameters between two images with different resolution. For the first case study, we generate an observed image by scaling, rotating, and translating the reference image. This allows us to compare registration results against ground truth. For the second study, we use two aerial images captured at different times and use one for the observed image and the other for the reference image.

5.1. Simulated problem

We will illustrate the performance of the method on two image registration scenarios: (1) when the observed image has lower resolution than the reference image, and (2) when both images are in low-resolution. Figure 5 shows the reference and observed images. To provide an initial estimate, the observed image is manually positioned on the reference image.[3] Then the observed image was iteratively registered on the reference image that was upsampled by a factor of 2 using bicubic interpolation. The upsampling was applied so that the sum of the reference pixel weights covered by the Gaussian kernel approaches 1 (Figure 3). Registration results are shown in Figures 6 and 7; given a rough initial registration, the proposed method finds the correct registration parameters.

For a quantitative evaluation of our method, we measured the number of registration successes in 300 trials for eight different configurations (Figure 8). Thus, there were $300 \times 8 = 2,400$ trials in total. Figure 5(a) was set as the reference image, and for each trial, an observed image was cropped from the reference image at random locations and scaled by a factor of 0.1. Registration was performed after displacing the observed image from its true position by adding random translation, rotation, and scale. For each configuration, the amount of displacement was increased to make the registration increasingly difficult. The displacement added for each configuration was within 0, ± 5, ± 10, ± 15, ± 20, ± 25, ± 30, and ± 35 for x-y translation and rotation. The scale factor was randomly initialized between 0.5 and 1.0 for all configurations. Figure 8 shows registration performance as a function of level of difficulty. In the first configuration, out of 300 trials, 80(26.67%) failed to converge to the true position while 198(66%) failed for the last configuration.

Figure 9 shows the two low-resolution images for the second scenario. In this case, both images had the same size but were generated with slightly different translation, rotation, and scaling. The reference and observed images were generated by downsampling the 512×512 image in

[3]The initial estimates can be obtained from other image processing steps by using an automatic detection system. Our method can be used to perform finer registration.

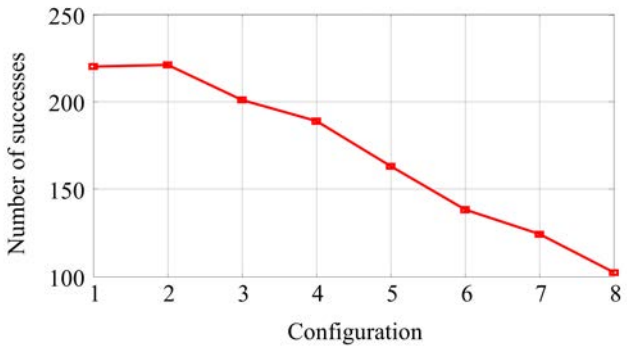

Figure 8. The number of successful registrations for eight different configurations. We defined success as x-y translation error less than 3 pixels, rotation error less than 1 degree, and scale error less than 0.01. The performance decreases as the registration problem becomes more difficult.

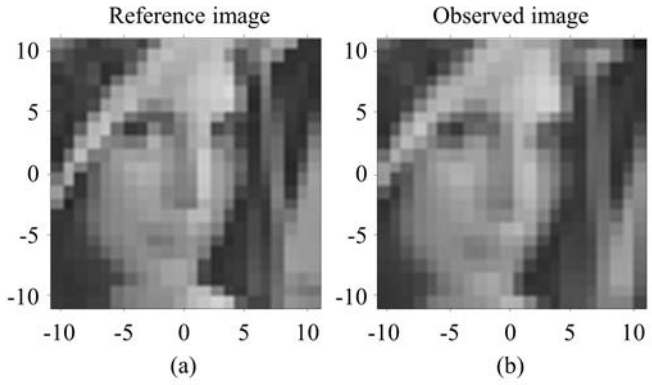

Figure 9. Reference (a) and observed image (b). Both images were obtained by down-sampling the 512×512 original image by a factor of 0.1 and 0.15, respectively, followed by cropping. Both images are 22×22.

Figure 5(a). The observed image was manually positioned on the reference image for initialization. Then the observed image was iteratively registered on the reference image that was upsampled by a factor of 50 using bicubic interpolation. We used a larger scaling factor for upsampling compared to the previous scenario since the reference image was in lower resolution. The registration process is shown in Figures 10 and 11; with a rough initial estimate, the proposed method is able to find the correct registration parameters. The results show that our method can simultaneously estimate the sub-pixel level changes in translation, rotation, and scale between the two images.

5.2. Real problem

Finally, we applied the method to register two aerial images taken at different instances. The two images are shown in Figure 12. As before, we manually registered the observed image to the reference image to give a rough initial estimate.[4] Then, we applied the image registration method to update the estimates. Two estimation runs are shown: in the first run (Figure 13) the initial scale factor is larger than the true scale; in the second run (Figure 14), the initial scale factor is smaller than the true scale and the observation image was rotated differently. During the estimation process, the x-y translation, rotation, and scale parameters of the observed image were updated simultaneously. The results show that the proposed technique can match the aerial images with different initializations.

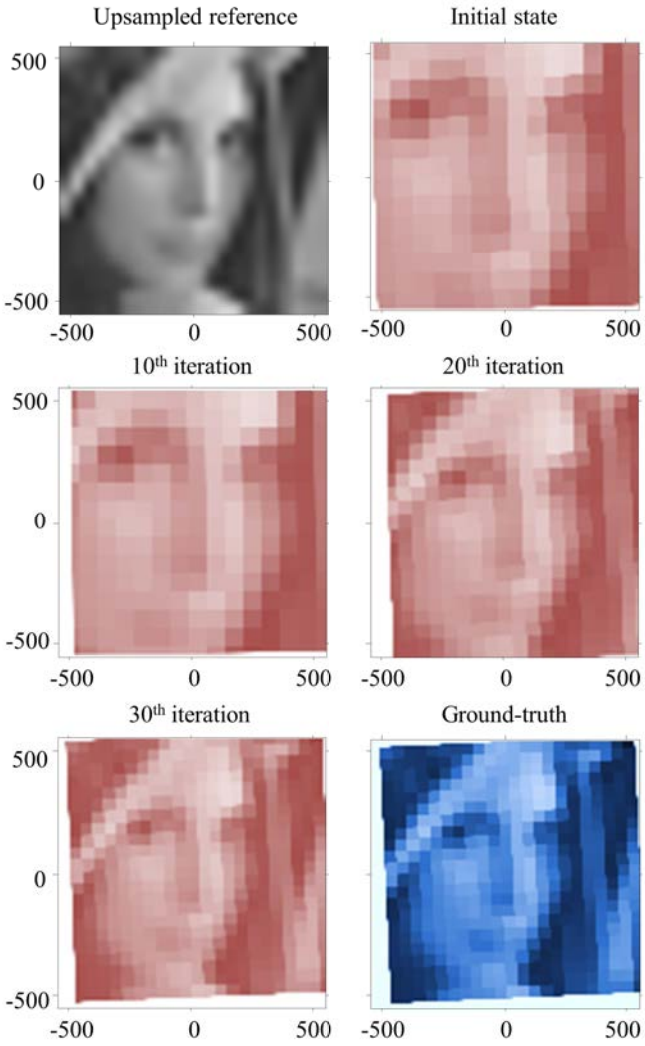

Figure 10. Ground-truth position (blue) and updated estimates of the observed image at the 10^{th}, 20^{th}, and 30^{th} iterations (red). Ground-truth values were $t_x = 0.5$, $t_y = 0.5$, $\theta = -2.5$, $s = 1.05$. The initial position starts with a 40% error for each registration parameter.

[4] For this particular domain, initial registration estimates for the two images can be obtained from GPS on an airborne sensor.

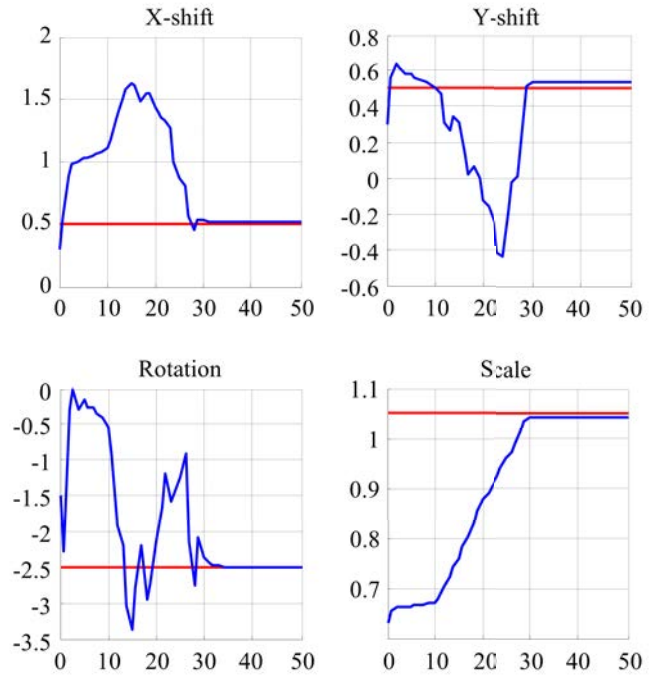

Figure 11. Estimated values for x-shift, y-shift, rotation, and scale in each iteration (blue). Ground-truth values were $t_x = 0.5$, $t_y = 0.5$, $\theta = -2.5$, $s = 1.05$ (red). The initial estimate starts with a 40% error for each registration parameter.

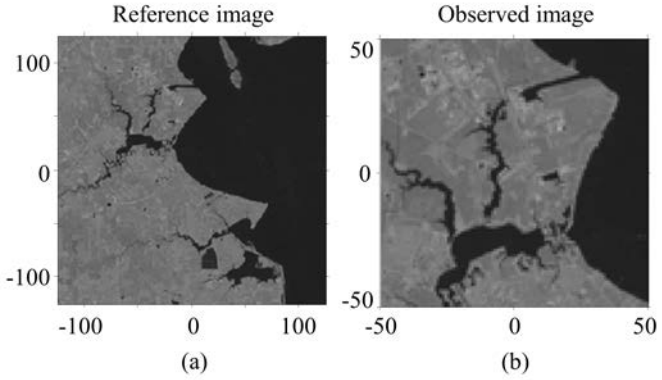

Figure 12. The reference (a) and observed image (b). The observed image is a downscaled and cropped version of the original image. The downscaling was applied to change the scale factor between the two original images. The example images are Landsat 7 images from the U.S. Geological Survey.

6. Conclusions

In this paper, we presented an image registration technique that can estimate translation, rotation, and scale in a unified manner and also take into account resolution differences caused by scaling. By assuming images are rectangular, we defined a scale-space model that relates the reference image and the observed image with a single scale parameter. This scale-space model is easily generalized to handle translation and rotation. The proposed model is embedded into a non-linear least squares method for simultaneous translation, rotation, and scale estimation.

Although the proposed area-based image registration technique can be used as a general image registration tool, it can be especially useful for low-resolution image registration, where salient feature points are difficult to extract accurately. Such situations occur in super-resolution problems where sub-pixel accuracy alignment is required. Future research will involve studies on the performance metrics of the proposed registration method to improve convergence speed and accuracy.

References

[1] E. H. Adelson, C. H. Anderson, J. R. Bergen, P. J. Burt, and J. M. Ogden. 1984, Pyramid methods in image processing. *RCA Engineer*, 29(6):33–41, 1984. 1, 2

[2] P. Bone, R. Young, and C. Chatwin. Position-, rotation-, scale-, and orientation-invariant multiple object recognition from cluttered scenes. *Optical Engineering*, 45(7):077203, 2006. 2

[3] J. L. Crassidis and J. L. Junkins. *Optimal Estimation of Dynamic Systems (Chapman & Hall/Crc Applied Mathematics & Nonlinear Science)*. Chapman & Hall/CRC, Apr. 2004. 4

[4] Y. Dufournaud, C. Schmid, and R. Horaud. Matching images with different resolutions. In *Computer Vision and Pattern Recognition, 2000. Proceedings. IEEE Conference on*, volume 1, pages 612 –618 vol.1, 2000. 1, 2

[5] T. Lindeberg. *Scale-Space*. John Wiley & Sons, Inc., 2007. 2

[6] D. Lowe. Object recognition from local scale-invariant features. In *Computer Vision, 1999. The Proceedings of the Seventh IEEE International Conference on*, volume 2, pages 1150 –1157 vol.2, 1999. 2

[7] B. Lucas and T. Kanade. An iterative image registration technique with an application to stereo vision (darpa). In *Proceedings of the 1981 DARPA Image Understanding Workshop*, pages 121–130, April 1981. 2

[8] R. Rao and S. Lee. A video processing approach for distance estimation. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 3, 2006. 2

[9] S. B. Reddy and B. N. Chatterji. An FFT-based technique for translation, rotation, and scale-invariant image registration. *IEEE Transactions on Image Processing*, 5(8):1266–1271, Aug. 1996. 2

[10] B. Zitová and J. Flusser. Image registration methods: a survey. *Image and Vision Computing*, 21:977–1000, 2003. 1

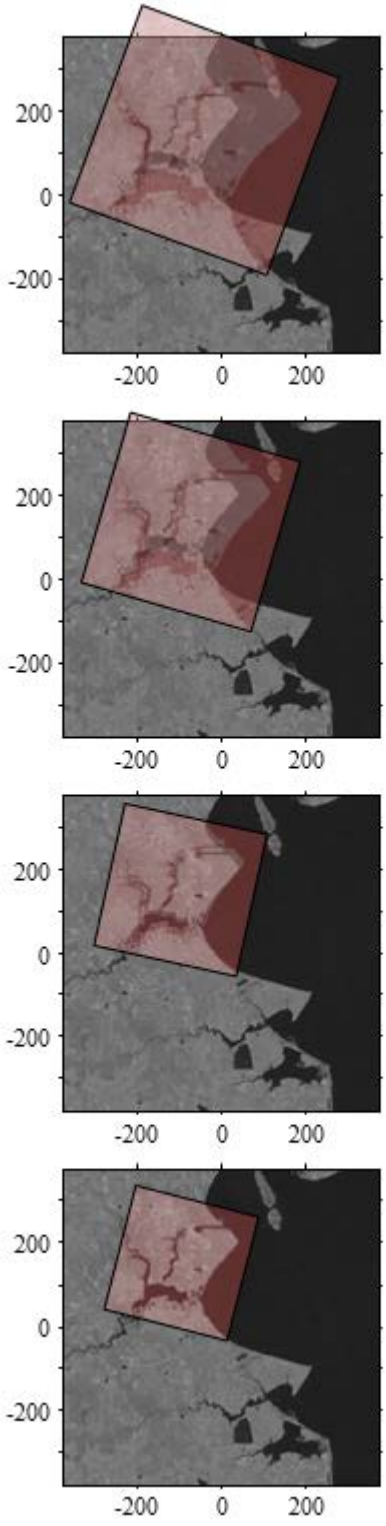

Figure 13. The updated estimates of the observed image position (red). The initial position starts from $t_x = -40$, $t_y = 130$, $\theta = 20$, $s = 1.7$. The estimates at certain iterations are shown. The estimation eventually converges and the observed image matches the reference image.

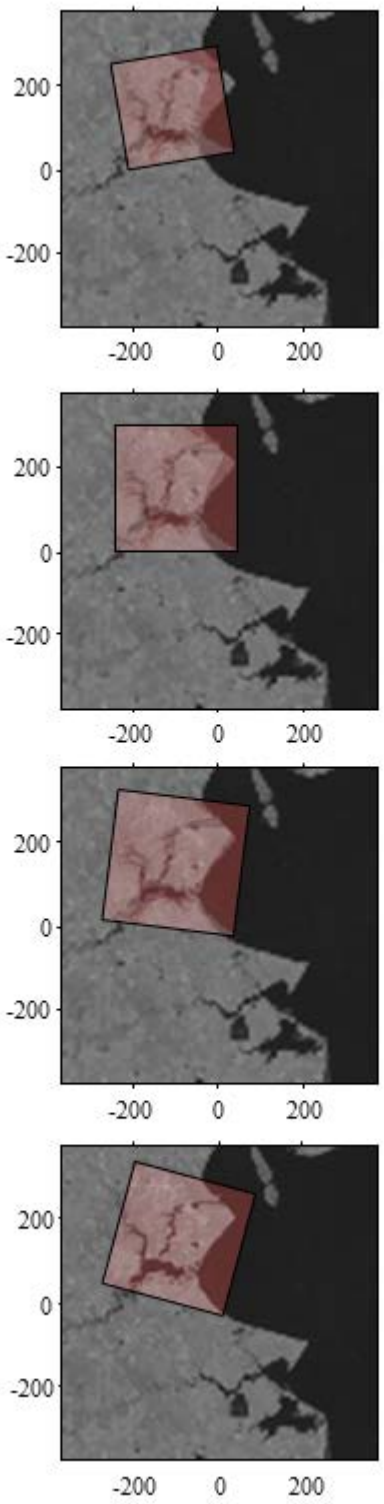

Figure 14. The updated estimates of the observed image position (red). The initial position starts from $t_x = -110$, $t_y = 150$, $\theta = -10$, $s = 0.8$. The estimates at certain iterations are shown. The estimation eventually converges and the observed image matches the reference image.