Hierarchical Multi-sensor Image Registration Using Evolutionary Computation

Ju Han and Bir Bhanu

Center for Research in Intelligent Systems
University of California, Riverside, CA 92521, USA
{jhan,bhanu}@cris.ucr.edu

ABSTRACT

Image registration between multi-sensor imagery is a challenging problem due to the difficulties associated with finding a correspondence between pixels from images taken by the two sensors. However, the moving people in a static scene provide cues to address this problem. In this paper, we propose a hierarchical approach to automatically find the correspondence between the preliminary human silhouettes extracted from synchronous color and infrared (IR) image sequences for image registration using evolutionary computation. The proposed approach reduces the overall computational load without decreasing the final estimation accuracy. Experimental results show that the proposed approach achieves good results for image registration between color and IR imagery.

Categories: I. Computing Methodologies
I.4 Image Processing and Computer Vision
I.4.8 Scene Analysis
Subject Descriptor: Sensor fusion
General Terms: Algorithms
Keywords: Sensor Fusion, Automatic Image Registration, Color Image Sequence, IR Image Sequence, Genetic Algorithm.

1. INTRODUCTION

Images from different kind of sensors generally have different pixel characteristics due to the phenomenological differences between the image formation processes of the sensors. In recent years, sensor fusion approaches have been employed to improve the performance of object detection and recognition, especially in the field of automated target recognition [1, 2] and remote sensing [3, 4, 5, 6].

Image registration is essential for precise comparison or fusion of images from multiple sensors at the pixel level. Sometimes, manual image registration is employed in many sensor fusion approaches [1, 2]. This needs lots of human interaction which is not desirable in processing large collections of image data taken under different field-of-views of the sensors.

Several approaches have been proposed for automatic registration between Synthetic Aperture Radar (SAR) and optical images. Li et al. [3] proposed an elastic contour matching scheme based on the active contour model for multi-sensor image registration between SAR (microwave) and SPOT (visible and near IR) images. Inglada and Adragna [4] proposed an approach for automatic image registration between SAR and SPOT images. They first extracted edges in both images, and then used a genetic algorithm to estimate the geometric transformation which minimized the matching error between corresponding edges. Similarly, Ali and Clausi [5] automatically registered SAR and visible band remote sensing images using an edge-based pattern matching method. In order to locate reliable control points between SAR and SPOT images, Dare and Dowman [6] proposed an automatic image registration approach based on multiple feature extraction and matching methods, rather than just relying on one method of feature extraction. Zheng and Chellappa [7] proposed an automatic image registration approach to estimate 2-D translation, rotation and scale of two partially overlapping images obtained from the same sensor. They extract features from each image using a Gabor wavelet decomposition and a local scale interaction method to detect local curvature discontinuities. Hierarchical feature matching is performed to obtain the estimate of translation, rotation and scale. Li and Zhou [8, 9] extended this single sensor image registration approach to the work of automatic color/IR and SAR/IR image registration. Their approach is based on the assumption that some strong contours are presented in both the color and IR images. Consistent checking is required to remove inconsistent features between images from different sensors.

Due to the difficulty in finding the correspondence between images with different physical characteristics, image registration between imagery from different sensors is still a challenging problem. In our task of human silhouette extraction, objects in color and IR images appear different due to different phenomenology of color and IR video sensors. Also, there are differences in the field-of-view and resolution of the sensors. Therefore, it is generally difficult to precisely determine the corresponding points between color and IR images. However, human motion provides enough cues for automatic image registration between synchronized color and thermal images in our human silhouette extraction application. Compared with the correspondence of individual points, the preliminary extracted body silhouette regions provide a more reliable correspondence between color and thermal image pairs. In this paper, we propose a automatic image registration method to perform a match of the transformed color silhouette to the thermal silhouette.

We use Genetic Algorithm (GA) to solve the optimization problem in silhouette matching. However, the accurate subpixel corresponding point search requires longer bit length for each coordinate

GECCO'05, June 25–29, 2005, Washington, DC, USA.

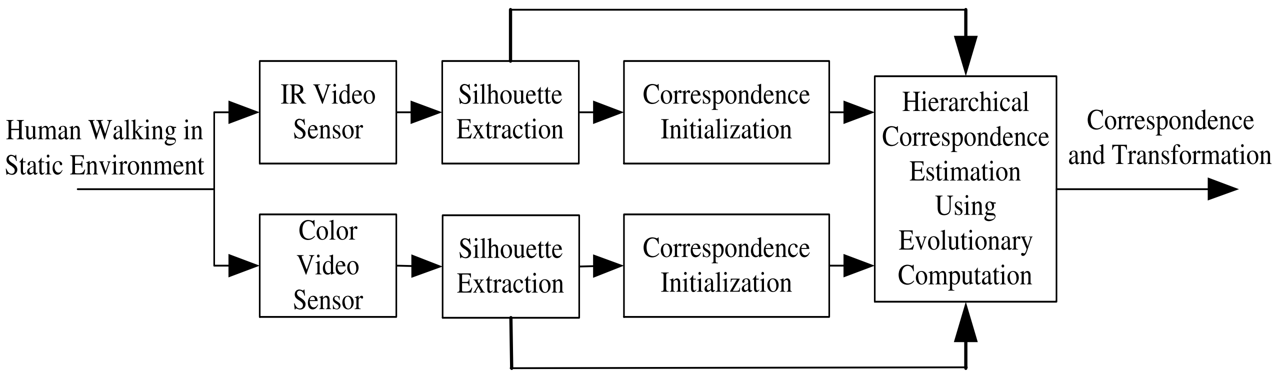


Figure 1: Proposed hierarchical genetic algorithm based multi-modal image registration approach.

value of each point. As a result, the population size of GA needs to be large to reduce the probability of falling into a local maxima. Due to the costly fitness function, the large population size is not desirable. In this paper, we propose a Hierarchical Genetic Algorithm (HGA) based search scheme to estimate the model parameters within a series of windows with adaptively reduced size at different levels. The proposed approach reduces the overall computational load of the GA without decreasing the final estimation accuracy.

2. TECHNICAL APPROACH

In this paper, we propose a Genetic Algorithm (GA) based hierarchical correspondence search approach for automatic image registration between synchronous color and IR image sequences as shown in Figure 1. Input of the system are videos possibly with humans, recorded simultaneously by both color and IR video sensors. Next, a background subtraction method is applied to both color and IR images to extract silhouettes from the background. Silhouette centroids are then computed from the color and IR silhouettes to form the initial correspondence between color and IR images. A hierarchical Genetic Algorithm (HGA) based scheme is employed to estimate the exact correspondence so that the silhouettes form the synchronous color and IR images are well matched. The transformation so obtained from this correspondence is used for the registration of images from color and IR video sensors. Mandava et al. [10] proposed an adaptive GA based approach for medical image registration with manually selected region-of-interest. As compared to their approach, our approach employs the similar concept of hierarchical search space scaling in evolutionary computation. However, the two approaches are different in strategies, implementation and applications.

2.1 Image Transformation Model

We use one color video sensor and one IR video sensor for sensor fusion. We locate the color and IR video sensors as close as possible without interference, and adjust their parameters so that the field-of-views of both sensors contain the desired scene where human motion occurs. Such a geometric transformation can be represented by a 3-D rotation transformation and a 3-D translation. We transform points in the color image plane into points in the IR image plane because the IR images have a higher resolution and a smaller field-of-view. The 2-D point (X, Y) in the color image is transformed into the 2-D point (X', Y') in the IR image as follows:

$$\begin{pmatrix} x' \\ y' \\ z' \end{pmatrix} = \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} + \begin{pmatrix} \Delta x \\ \Delta y \\ \Delta z \end{pmatrix}, \quad (1)$$

$$\begin{pmatrix} X' \\ Y' \end{pmatrix} = \begin{pmatrix} f'x'/z' \\ f'y'/z' \end{pmatrix} \quad \text{and} \quad \begin{pmatrix} X \\ Y \end{pmatrix} = \begin{pmatrix} fx/z \\ fy/z \end{pmatrix}, \quad (2)$$

where (x, y, z) and (x', y', z') are the 3-D location of the points in color and IR video sensor coordinates, respectively; (Δx, Δy, Δz) is the 3-D displacement vector of two sensors in the world coordinate system; f and f' are focal length of two video sensors.

According to the degree of elasticity of the transformations, they can be rigid, affine, projective, or curved [11]. In our example set up the geometric transformation is more complex than the rigid, affine, and projective models since the surfaces of objects encountered in the scene are not flat. However, the geometric transformation for planar objects can be strictly represented by a projective transformation [12]. Assuming that there is a large distance between the video sensor and the walking people, the visible human surface from the video sensor view can be approximated as planar. In this case, the projective transformation model is appropriate for image registration for our application. Furthermore, assuming that the person walks along the frontoparallel direction with respect to the image plane and video sensor axes are parallel to the ground, the visible human surface during the walking sequence approximately parallels to the image planes of both color and IR video sensors. Under this assumption, the geometric transformation can be further simplified as the rigid model. A rigid transformation can be decomposed into 2-D translation, rotation, and reflection. In the rigid transformation, the distance between any two points in the color image plane is preserved when these two points are mapped into the IR image plane. The 2-D point (X, Y) in the color image plane is transformed into the 2-D point (X', Y') in the IR image plane as follows:

$$\begin{pmatrix} X' \\ Y' \end{pmatrix} = \begin{pmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{pmatrix} \begin{pmatrix} X \\ Y \end{pmatrix} + \begin{pmatrix} \Delta X \\ \Delta Y \end{pmatrix}, \quad (3)$$

where θ is the rotation angle, and $(\Delta X, \Delta Y)^T$ is the translation vector [11]. Rigid transformations are used when shapes in the input image are unchanged, but the image is distorted by some combination of translation, rotation, and scaling. Straight lines remain straight, and parallel lines are still parallel under the assumption as mentioned above. A minimum correspondence of two pairs of points is required in the rigid transformation.

Compared to the complex projective model, the rigid model reduces the probability of estimated transformation overfitting the distorted human silhouettes from color and IR video sensors. Although the frontoparallel assumption is generally unavailable in real surveillance situations, we only need one pair of sequences satisfying this assumption to register color/IR imagery. The registration result will be used for human silhouette extraction from other synchronous color/IR sequence pairs under the same sensor setup.

2.2 Silhouette Extraction and Correspondence Initialization

Assuming that both the color and IR video sensors are static and human is the only moving object in the scene, human silhouette can be extracted by a simple background subtraction method. To model the color background, we choose some frames from the given color image sequence that only contains background, and compute the mean and standard deviation values for each pixel in each color channel. Assuming that the background has a Gaussian distribution at each pixel, a pixel at (X, Y) in the input color image is classified as part of moving objects if

$$|r(X,Y) - \mu_r(X,Y)| > \beta\sigma_r(X,Y) \quad (4)$$

$$|g(X,Y) - \mu_g(X,Y)| > \beta\sigma_g(X,Y) \quad (5)$$

$$\text{or} \quad |b(X,Y) - \mu_b(X,Y)| > \beta\sigma_b(X,Y) \quad (6)$$

where r, g and b represent pixel color values of the input image for red, green and blue channels, respectively; μ_r, μ_g and μ_b represent mean values of the background pixel color; σ_r, σ_g and σ_b represent standard deviation values of the background pixel color; β is the arbitrarily selected threshold.

Similarly, a pixel at (X, Y) in the input IR image is classified as part of moving objects if

$$|t(X,Y) - \mu_t(X,Y)| > \beta\sigma_t(X,Y), \quad (7)$$

where t represents the pixel IR value in the input IR image; μ_t represents the mean value of the background pixel temperature; σ_t represents the standard deviation value of the background pixel temperature; β is the arbitrary selected threshold. The threshold β is chosen to have the same value for both color and IR images so that the extracted body silhouettes can be compared at the same level ($\beta = 15$ in our experiments).

After silhouettes are extracted from each color image and its synchronous IR image, the centroid of the silhouette region is computed as the initial correspondence between each pair of color and IR images.

2.3 Automatic Image Registration Using Evolutionary Computation

In manual image registration, a set of corresponding points are manually selected from the two images to compute the parameters of the transformation model, and the registration performance is generally evaluated by manually comparing the registered image pairs. The same step is repeated several times until the registration performance is satisfied. If the background changes, the entire procedure needs to be repeated again. This makes manual image registration inapplicable when data are recorded at different locations with changing time or with different sensor setup. The automatic image registration is desirable under this situation.

2.3.1 Parameter Estimation by Silhouette Matching

Due to the phenomenological differences of objects in color and IR images, it is difficult to automatically find accurate correspondence between color and IR images. Figure 2 shows different object appearances in color and IR images due to the phenomenological difference between the image formation process of color and IR video sensors. However, human motion provides enough cues for automatic image registration between synchronized color and IR images in our human silhouette extraction application. Compared with the correspondence of individual points, the preliminary extracted body silhouette regions provide more reliable correspondence between color and IR image pairs. Therefore, we propose a method to perform a match of the transformed color silhouette to the IR silhouette. That is, we estimate the set of model parameters $\mathbf{p}$ to maximize

$$\text{Similarity}(\mathbf{p}; I_i; C_i) = \prod_{i=1}^{N} \frac{\text{Num}(T_{C_i;\mathbf{p}} \cap I_i))}{\text{Num}(T_{C_i;\mathbf{p}} \cup I_i)} \quad (8)$$

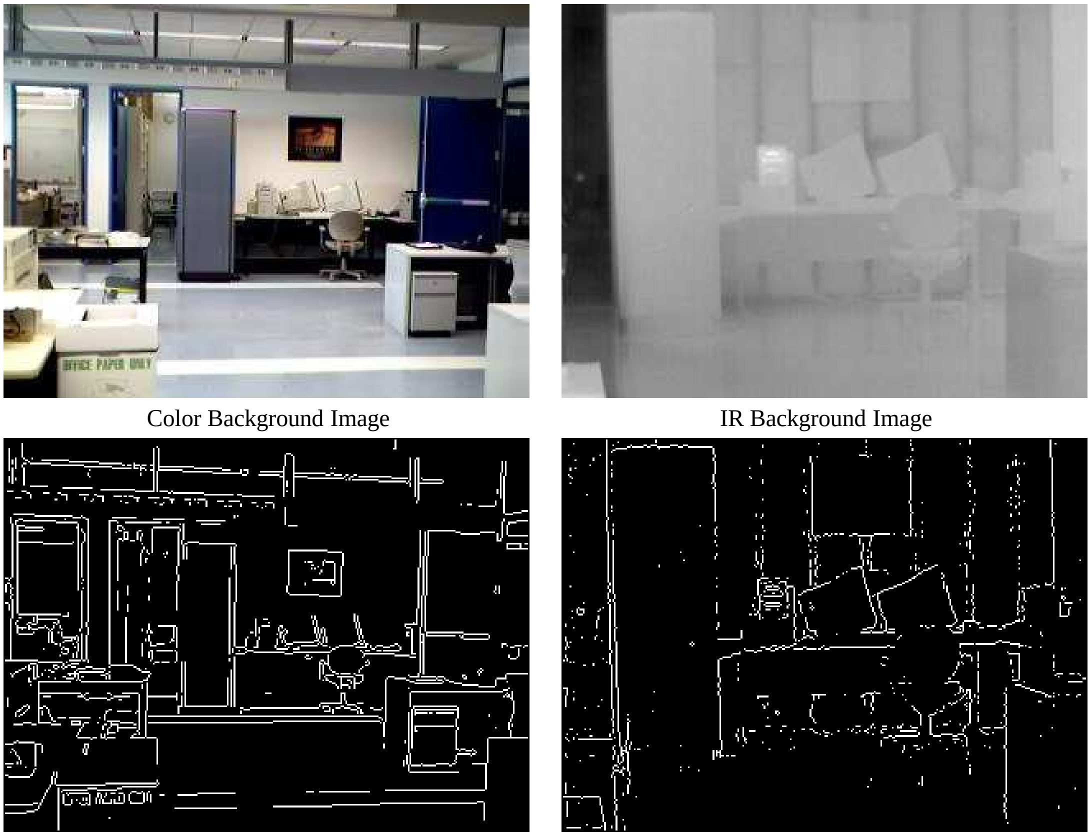
where I is the silhouette binary image obtained from the IR image, C is the silhouette binary image obtained from color image, $T_{C;\mathbf{p}}$ is the transformed binary image of C by rigid transformation with parameter set $\mathbf{p}$, N is the number of color and IR image pairs, and $\text{Num}(X)$ is the number of silhouette pixels in a silhouette image X. We use the product of similarity of image pairs instead of the sum to reduce the possibility of falling into local maxima on specific frame(s), i.e., to increase the possibility of the global maximum on all images pairs.

In the rigid transformation model, the parameters are the elements of the 2-D linear transformation matrix in Equation (3). However, the ranges of these parameters are difficult to be determined. In the rigid transformation model, a maximum correspondence of four points (two pairs) is required. If we fix two points in the IR image as the reference points, the 2-D coordinates of their corresponding points in the synchronous color image can be used to determine the rigid transformation model. Because the locations of the corresponding points should exist in limited local areas, the ranges of new parameters can be determined.

For each pair of color and IR images, we have obtained one pairs of initial correspondence, i.e., centroid of the preliminary extracted silhouettes. Under the assumption of planar object surface (i.e., human walks along the same direction in the scene so that the human body surface from the video sensor view in each frame lies on the same plane over the whole sequence), we can choose initial correspondence from two image pairs in the given color and IR image sequences. In this way, we have two pairs of initial correspondence: two points in the IR image are reference points and two points in the color images are initial model parameters whose exact values need to be estimated. If all the two pairs of points are chosen from a small area, the resulting registration performance may be unsatisfied in other areas. To avoid this problem, these points should be located as far away as possible in the images. This is also the reason that we do not choose all the corresponding points from the silhouettes of one color and IR image pair.

2.3.2 Genetic Algorithm based Hierarchical Correspondence Search

We use Genetic Algorithm (GA) to solve the optimization problem in Equation (8). GA provides a learning method motivated by an analogy to biological evolution. Rather than search from general-to-specific hypotheses, or from simple-to-complex hypotheses, GA generates successor hypotheses by repeatedly mutating and



Color Background Image — IR Background Image

Color Edge Image — IR Edge Image

Figure 2: Different object appearances in color and IR images are due to the phenomenological differences between the image formation process of color and IR video sensors. The images are at different resolutions and the field-of-views (FOVs) of the two sensors overlap (Color video sensor FOV contains IR video sensor FOV).

recombining parts of the best currently known hypotheses. At each step, a collection of hypotheses called the current population is updated by replacing some fraction of the population by offspring of the most fit current hypotheses. After a large number of steps, the hypotheses having the best fitness are considered as solutions. However, a single genetic algorithm is not appropriate to estimate the subpixel location of corresponding points in given search windows. The accurate subpixel corresponding point search requires longer bit length for each coordinate value of each point. As a result, the population size of GA need to be large to reduce the probability of falling into a local maxima. Due to the costly fitness function (8), the large population size is not desirable. In this paper, we propose a Hierarchical Genetic Algorithm (HGA) based search scheme to estimate the model parameters within a series of windows with adaptively reduced size as shown in Figure 3. The model parameters are coordinate values of the two points in the color image plane, corresponding to the two reference points in the thermal image plane.

We choose the estimated human silhouette centroids (as mentioned in Section 2.2) from two IR images in the same IR video as 2 reference points in the IR image plane. Let $\mathbf{p} = [x_1, y_1, x_2, y_2]^T$ be the model parameter to be estimated, where $\mathbf{x}_1 = (x_1, y_1)$ and $\mathbf{x}_2 = (x_2, y_2)$ are correspondence coordinates to be searched. The estimated 2 centroids from the two corresponding color images, $\mathbf{x}_{1,0} = (x_{1,0}, y_{1,0})$ and $\mathbf{x}_{2,0} = (x_{2,0}, y_{2,0})$, are chosen as the initial correspondence in the color image plane. At each search level of the HGA based search scheme, GA is applied to estimate the two corresponding coordinates according to Equation (8). The center and size of search windows are both determined by the previous three estimates of corresponding coordinates. In the kth search level, the centers of the two search windows for the two corresponding points are chosen as follows:

$$\mathbf{c}_{i,k} = \begin{cases} \mathbf{x}_{i,0}, & ifk = 1; \\ (\mathbf{x}_{i,0} + \mathbf{x}_{i,1})/2, & ifk = 2; \\ (\mathbf{x}_{i,k-3} + \mathbf{x}_{i,k-2} + \mathbf{x}_{i,k-1})/3, & ifk > 2, \end{cases} \tag{9}$$

where $i = 1, 2$, $\mathbf{x}_{i,j}$ is the new estimate of $\mathbf{x}_i$ after the jth search level. The square length of the search windows are chosen as follows:

$$w_k = \begin{cases} w_0, & ifk = 1; \\ \max_{j=1}^{4}\{|p_{j,0} - p_{j,1}|\}, & ifk = 2; \\ \max_{j=1}^{4}\{\max\{p_{j,k-3}, p_{j,k-2}, p_{j,k-1}\} & \\ \quad - \min\{p_{j,k-3}, p_{j,k-2}, p_{j,k-1}\}\}, & ifk > 2, \end{cases} \tag{10}$$

where w_0 is the preselected initial length of search windows. This iterative procedure is repeated until the search w_k is lower than a pre-selected lower limit w_l.

In the proposed approach, the code length of parameters in each

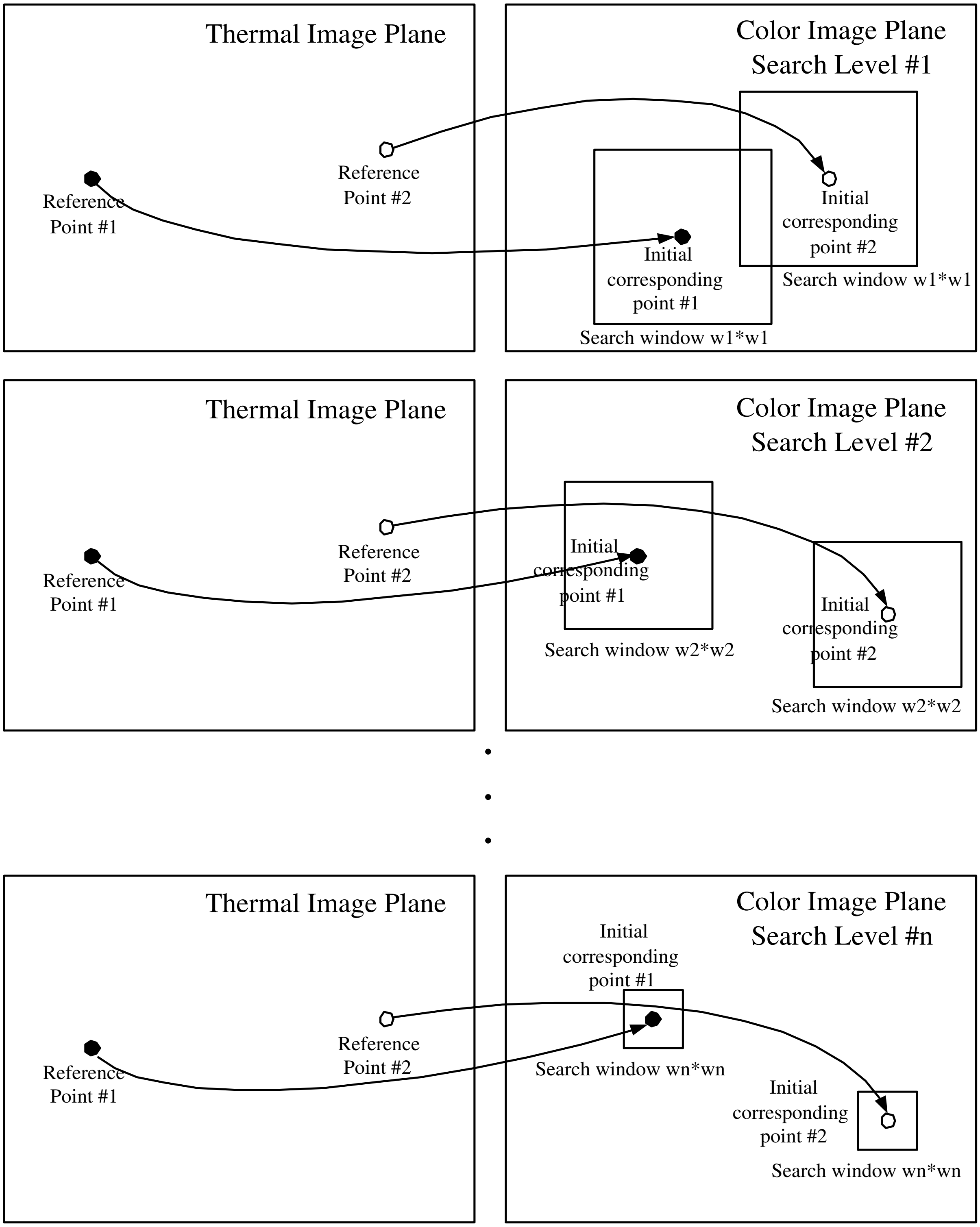


Figure 3: HGA-based search scheme to estimate the two 2-D point locations (black and white) simultaneously in color images. This graph only illustrates the procedure of correspondence estimation. The search windows here are only for illustration whose sizes are much larger than the real window sizes.

Figure 4: Examples of registration results: first row - original color images, second row - original IR images, third row - transformed color images.

GA can be short without decreasing the final estimation accuracy. Considering the costly objective function in our application, the population size cannot be large. Short code length is desired because a GA with high ratio of code length over population size has a high probability of falling into the local maximum. Generally, the window size will be adaptively reduced until reaching the lower limit. Even if the real correspondence exists outside of the initial search window, the approach still has the possibility to find a good estimate because the new window might cover areas outside of the initial window. After the correspondences in the color image plane are located, the transformation is uniquely determined for this pair of color and IR image sequences, and it will be used to transform color images into the plane of IR images.

3. EXPERIMENTAL RESULTS

The image data used in our experiments are real human walking data recorded by the two video sensors in the same indoor environment. Color images are recorded by a PC camera with image size of 240×320 as shown in the first row of Figure 4. IR images are recorded by a long-wave IR video sensor with image size of 240×320 as shown in the third row of Figure 4. Both video sensors have fixed but different focal lengths. The IR video sensor has a higher resolution and less distortion than the PC camera, and is, therefore, used as the base video sensor in our experiments.

Three color and IR images are selected for matching in Equation (8). The initial search window is set as 16×16 pixels ($w_1 = 16$), and the final search window is 0.1×0.1 pixels ($w_l = 0.1$). In the GA at each search level, we use 6 bits to represent each coordinate value (totally 24 bits for 4 coordinate values); fitness function is the similarity between image pairs in Equation (8); population size is 100; crossover rate is 0.9; crossover method is uniform crossover; mutation rate is 0.05; the GA will terminate if the fitness values have not changed for 5 successive steps.

Figure 5 shows examples of estimated transformation results from good initial corresponding points at different search levels, while Figure 6 shows results from bad initial corresponding points. Even though the original transformation 6(c) is far away from the true transformation, the transformation results are improved gradually at successive search levels and finally converged around the real transformation. The variations of fitness values from good and bad initial correspondence are shown in Figure 7 and 8, respectively. The vertical line corresponds to the last generation at each search level. The curve between two adjacent vertical lines indicate the variation of GA fitness values in a search level. In the GA at each search level, the populations are randomly generated, leading to the drop of the fitness value at the beginning of each search level. We do not use the population from the previous search level because we hope to estimate transformation parameters more accurately as the window size decreases and diversify the population to avoid premature convergence. In general, our image registration approach is not sensitive to the location of initial correspondence if it is located inside or slightly outside of the initial search window depending on the size of the initial search window.

Figure 4 shows the comparison of original color images, transformed color images, and original IR images. To evaluate the registration performance, we define the registration precision as $P(A, B) = (A \cap B)/(A \cup B)$, where A and B are manually labeled human silhouette pixel sets from the original IR image and the transformed color image, respectively. According to this definition, the registration precision for the 3 image pairs in Figure 4 is 78%, 80% and

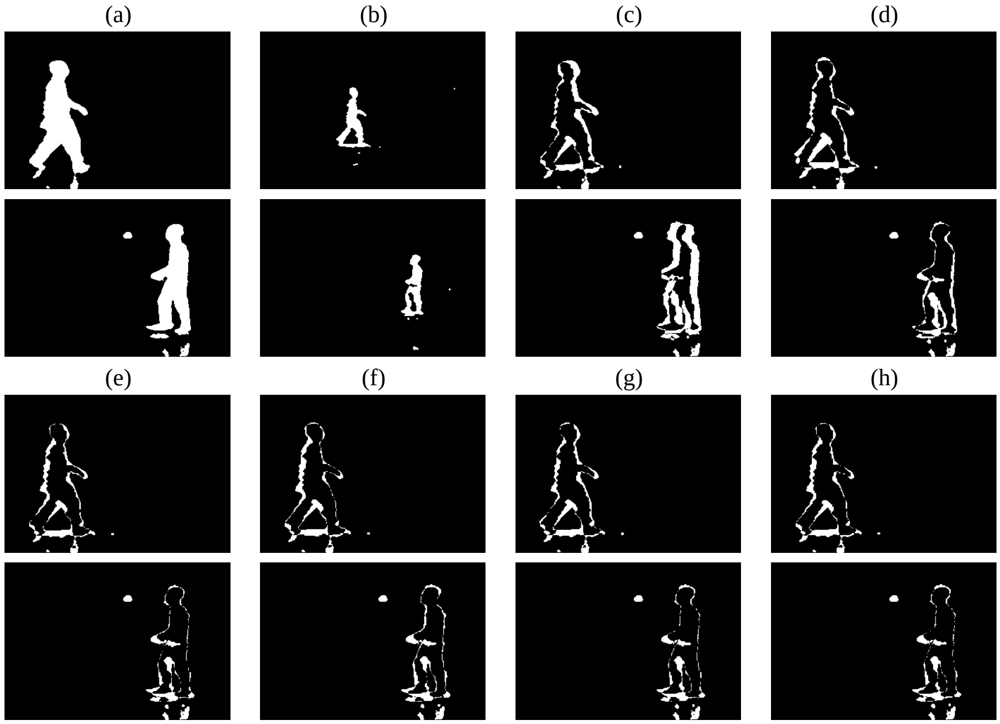


Figure 5: Example of estimated transformation results from good initial correspondence: (a) original silhouettes from thermal images, (b) original silhouettes from color images, (c) matching error of the initial transformation, (d) matching error after the first search level, (e) after the second search level, (f) after the third search level, (g) after the fourth search level, (h) after the 12th search level.

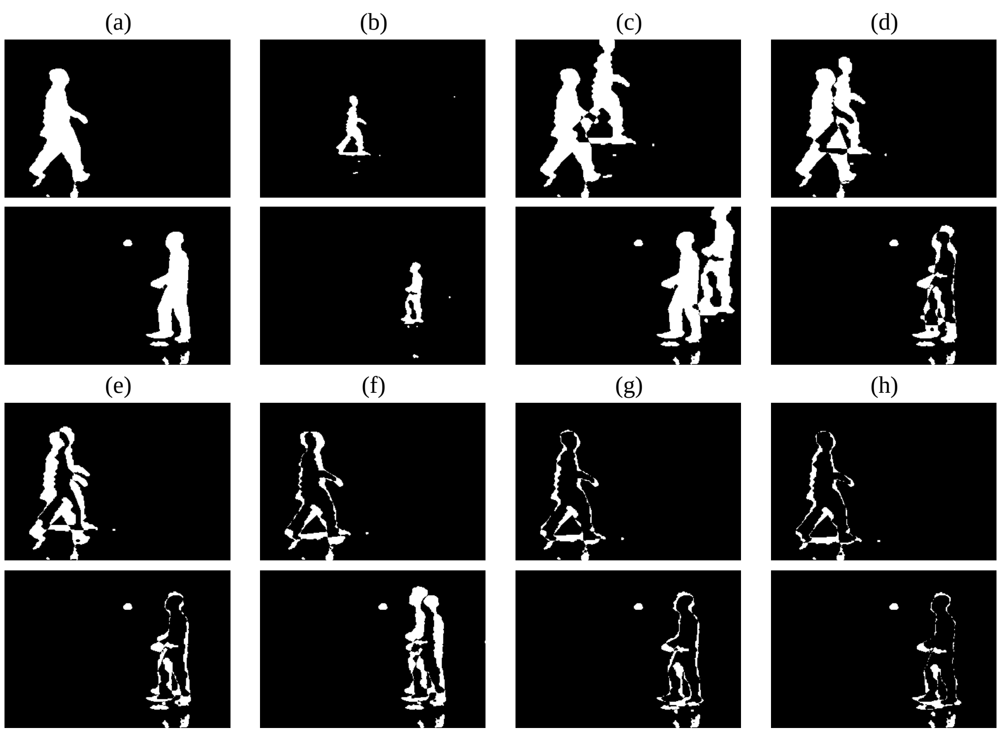


Figure 6: Example of estimated transformation results from bad initial correspondence: (a) original silhouettes from thermal images, (b) original silhouettes from color images, (c) matching error of the initial transformation, (d) matching error after the first search level, (e) after the second search level, (f) after the third search level, (g) after the fourth search level, (h) after the 23th search level.

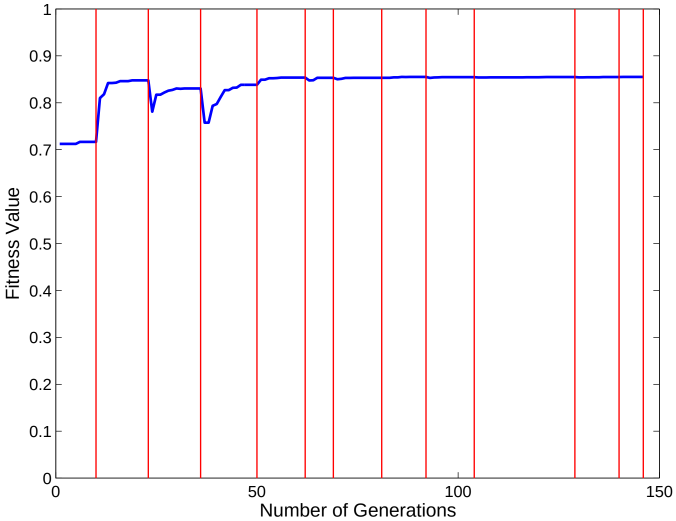


Figure 7: Variation of fitness values from good initial correspondence. The vertical line corresponds to the last generation at each search level. The curve between two adjacent vertical lines indicate the variation of GA fitness values at a search level.

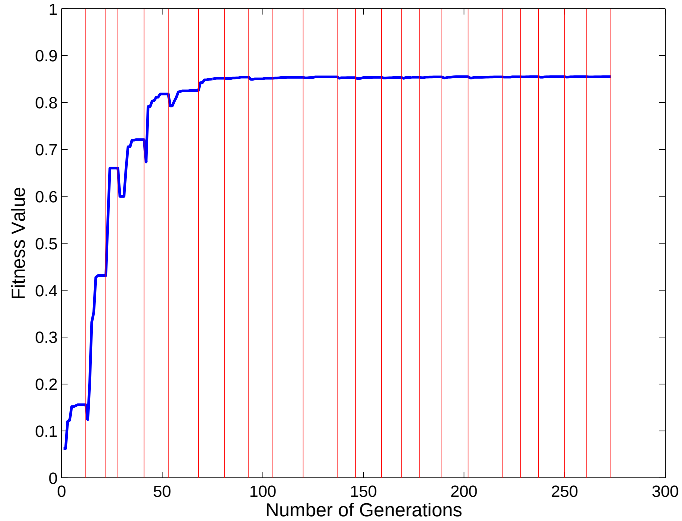


Figure 8: Variation of fitness values from bad initial correspondence. The vertical line corresponds to the last generation at each search level. The curve between two adjacent vertical lines indicate the variation of GA fitness values at a search level.

85%, respectively. Considering that the color and IR image pairs are not exactly synchronized, and there are labeling errors due to the physical difference between color and IR signals, our image registration still achieves good results. This experiment is a nice case study with successful results. The proposed approach will be applied on different data sets to test the generality.

4. CONCLUSIONS

In this paper, we propose a Hierarchical Genetic Algorithm (HGA) based scheme to automatically find correspondence so that the preliminary silhouettes form the color and IR images are well matched. The computed correspondence and the corresponding transformation are used for image registration in the same scene. Experiments show that the proposed approach is not sensitive to the location of initial correspondence if it is located inside or slightly outside of the initial search window, and achieves good performance for image registration between color and IR image sequences.

5. REFERENCES

[1] Clark, G., Sengupta, S., Buhl, M., Sherwood, R., Schaich, P., Bull, N., Kane, R., Barth, M., Fields, D., Carter, M.: Detecting buried objects by fusing dual-band infrared images. 1993 Conference Record of The Twenty-Seventh Asilomar Conference on Signals, Systems and Computers **1** (1993) 135–143

[2] Perez-Jacome, J., Madisetti, V.: Target detection from coregistered visual-thermal-range images. Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing **4** (1997) 2741–2744

[3] Li, H., Manjunath, B., Mitra, S.: A contour-based approach to multisensor image registration. IEEE Transactions on Image Processing **4** (1995) 320–334

[4] Inglada, J., Adragna, F.: Automatic multi-sensor image registration by edge matching using genetic algorithms. Proc. IEEE International Geoscience and Remote Sensing Symposium **5** (2001) 2313–2315

[5] Ali, M., Clausi, D.: Automatic registration of SAR and visible band remote sensing images. Proc. IEEE International Geoscience and Remote Sensing Symposium **3** (2002) 1331–1333

[6] Dare, P., Dowman, I.: Automatic registration of SAR and spot imagery based on multiple feature extraction and matching. Proc. IEEE International Geoscience and Remote Sensing Symposium **7** (2000) 2896–2898

[7] Zheng, Q., Chellappa, R.: A computational vision approach to image registration. IEEE Transactions on Image Processing **2** (1993) 311–326

[8] Li, H., Zhou, Y.: Automatic EO/IR sensor image registration. Proc. International Conference on Image Processing **3** (1995) 240–243

[9] Li, H., Zhou, Y., Chellappa, R.: SAR/IR sensor image fusion and real-time implementation. Record of the Twenty-Ninth Asilomar Conference on Signals, Systems and Computers **2** (1995) 1121–1125

[10] Mandava, V., Fitzpatrick, J., Pickens, D.I.: Adaptive search space scaling in digital image registration. IEEE Transactions on Medical Imaging **8** (1989) 251–262

[11] van den Elsen, P., Pol, E.J., Viergever, M.: Medical image matching - a review with classification. IEEE Engineering in Medicine and Biology Magazine **12** (1993) 26–39

[12] Yao, J.: Image registration based on both feature and intensity matching. Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing **3** (2001) 1693–1696