11-3

Detecting Moving Humans Using Color and Infrared Video

Ju Han and Bir Bhanu
Center for Research in Intelligent Systems
University of California, Riverside, California 92521, USA
{jhan,bhanu}@cris.ucr.edu

Abstract

We approach the task of human silhouette extraction from color and infrared video using automatic image registration. Image registration between color and thermal images is a challenging problem due to the difficulties associated with finding correspondence. However, the moving people in a static scene provide cues to address this problem. In this paper, we propose a hierarchical scheme to automatically find the correspondence between the preliminary human silhouettes extracted from color and infrared video for image registration. Next, we discuss some strategies for probabilistically combining cues from registered color and thermal images. It is shown that the proposed approach achieves good results for image registration and human silhouette extraction. Experimental results also show a comparison of sensor fusion strategies and demonstrate the improvement in performance for human silhouette extraction.

1. Introduction

Current human recognition methods, such as fingerprints, face or iris biometrics, generally require a cooperative subject, views from certain aspects and physical contact or close proximity. These methods can not reliably recognize non-cooperating individuals at a distance in real-world changing environmental conditions. Moreover, in many applications of personnel identification, many established biometrics can be obscured. Gait, which concerns recognizing individuals by the way they walk, can be used as a biometric without the above-mentioned disadvantages.

The initial step of most of the gait recognition approaches is human silhouette extraction [1, 2, 3, 4]. Most gait recognition approaches use electro-optical (EO) sensors such as CCD cameras. However, it is very likely that some part of human body or clothing has similar color as background colors. In this case, human silhouette extraction usually fails on this part. Moreover, the existence of shadows is a problem for EO sensors. In addition, EO sensors do not work under low lighting conditions such as night or indoor environment without lighting. The top row in Figure 1 shows human silhouette extraction results from a color image.

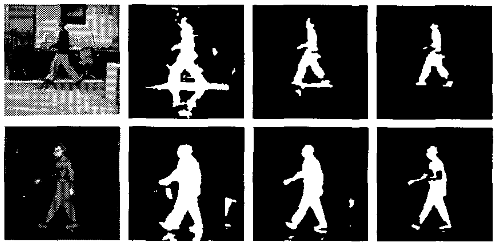

Figure 1. Human silhouette extraction results from color image (top row) and thermal image (bottom row) using background subtraction method with increasing thresholds from left to right (the leftmost image is the original image).

To avoid the disadvantages of using EO sensors, we investigate the possibility of using infrared (IR) sensor for gait analysis [5]. Unlike a regular camera which records reflected visible light, a long wave ($8 - 12\mu m$) IR camera records electromagnetic radiation emitted by objects in a scene as a thermal image whose pixel values represent temperature. In a thermal image that consists of humans in a scene, human silhouettes can be easily extracted from the background regardless of lighting conditions and colors of the human surfaces and backgrounds, because the temperatures of the human body and background are different in most situations [6]. Although the human silhouette extraction results from IR sensors are generally better than that from EO sensors, human silhouette extraction is unreliable when some part of the human body or clothing has the temperature similar to background temperature. In addition, human body occurs obvious projection on smooth surfaces such as smooth floor. The bottom row in Figure 1 shows human silhouette extraction results from a thermal image.

In recent years, some sensor fusion approaches have already been employed for human detection and recognition. Wilder et al. [7] compare the effectiveness of EO and IR imagery for detecting and recognizing faces, and expect the improvement of face detection and recognition algorithms that fuse the information from the two sensors. Yoshitomi et al. [8] propose a integrated method to recognize the emotional expressions of a human using both voice and facial

0-7803-7987-X/03/$17.00 ©2003 IEEE.

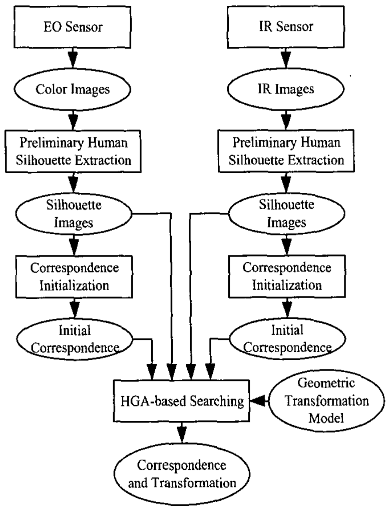
EO Sensor
IR Sensor
Color Images
IR Images
Preliminary Human Silhouette Extraction
Preliminary Human Silhouette Extraction
Silhouette Images
Silhouette Images
Correspondence Initialization
Correspondence Initialization
Initial Correspondence
Initial Correspondence
HGA-based Searching
Geometric Transformation Model
Correspondence and Transformation

Figure 2. Diagram of the proposed hierarchical Genetic Algorithm based multi model image registration approach.

expressions. The recognition results show that the integration method for recognizing emotional states gives better performance than any of isolated methods.

Notice that the unreliably extracted body parts from one sensor might be reliably extracted from the other sensor. In this paper, we propose a human silhouette detection approach by combining cues from both EO and IR sensors.

2. Technical Approach

Objects in color and thermal images appear different due to the phenomenological difference between the image formation process of two sensors. This makes image registration between color and thermal images a challenging problem. However, in the application of gait recognition, human motion provides enough cues for image registration between color and thermal images assuming the background is static.

In this paper, we propose a Genetic Algorithm (GA) based hierarchical search approach for image registration between color and thermal image sequences as shown in Figure 2. First, human motion in a scene is recorded simultaneously by both EO and IR sensors. Next, a simple background subtraction method is applied on both color and thermal images to extract preliminary body silhouettes from the background. Silhouette centroid and head position are then computed from the body silhouettes as the initialized correspondence between color and thermal images. A hierarchical Genetic Algorithm (HGA) based scheme is employed to search correspondence so that the silhouettes form the color and thermal images are well matched. The correspondence so obtained and the corresponding transformation are used for image registration in the same scene. Finally, registered thermal and color images are combined by probabilistic strategies to obtain better body silhouette extraction results.

Mandava et. al proposed an adaptive search space scaling GA approach for medical image registration with manually selected region-of-interest [9]. Compared with their approach, our approach employs the similar concept of hierarchical search space scaling in GA. However, the two approaches are different in strategies, implementation and applications which will be explained in details in the following Sections.

2.1. Image Transformation Model

We use one EO camera and one IR camera for sensor fusion. We locate the EO and IR camera as close as possible without interference, and adjust their camera parameters so that the fields of view of both cameras contain the desired scene where human motion occurs. Such a geometric transformation can be represented by a 3-D linear transformation and a 3-D translation. We transform points in color images into points in IR images because the IR images have higher resolution and smaller view field. 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:

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

$$\begin{pmatrix}X'\\Y'\end{pmatrix} = \begin{pmatrix}f'x'/z'\\f'y'/z'\end{pmatrix} \quad \text{and} \quad \begin{pmatrix}X\\Y\end{pmatrix} = \begin{pmatrix}fx/z\\fy/z\end{pmatrix}, \tag{1}$$

where (x, y, z) and (x', y', z') are the 3-D location of the points in EO and IR camera coordinates, respectively; $(\Delta x$, Δy, $\Delta z)$ is the 3-D displacement vector of two cameras in the world coordinate system; f and f' are focal length of two cameras.

According to the degree of elasticity of the transformations, they can be rigid, affine, projective, or curved [10]. Our geometric transformation is more complex than the rigid, affine, and projective models. However, the geometric transformation for planar objects can be strictly represented by a projective transformation [11] as follows. A projective transformation can be represented by a 3-D linear transformation. The 2-D point (X, Y) in the first image is transformed into the 2-D point (X', Y') as follows:

$$\begin{pmatrix}X'\\Y'\end{pmatrix} = \begin{pmatrix}\frac{x'}{z'}\\\frac{y'}{z'}\end{pmatrix} \text{ and } \begin{pmatrix}x'\\y'\\z'\end{pmatrix} = \begin{pmatrix}a_{11}&a_{12}&a_{13}\\a_{21}&a_{22}&a_{23}\\a_{31}&a_{32}&a_{33}\end{pmatrix}\begin{pmatrix}X\\Y\\1\end{pmatrix} \tag{2}$$

where z' represents the extra homogeneous coordinate. A minimum correspondence of 8 points (4 pairs) is required in projective transformation.

Assuming that the human surface from the camera view is approximately planar compared with the long distance between the cameras and people, and the background pixels are not considered, the projective transformation model is appropriate for image registration in this application.

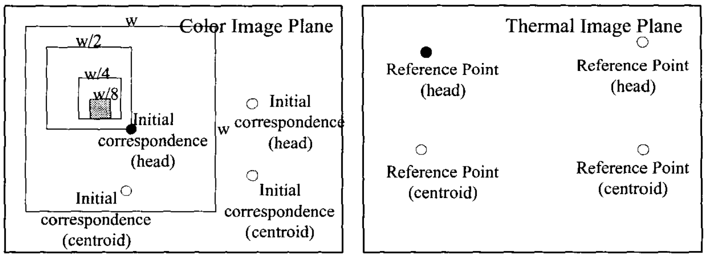
w
~~Color~~ Image Plane
w/2
w/4
w/8
Initial correspondence (head)
Initial correspondence (centroid)
w
Initial correspondence (head)
Initial correspondence (centroid)
Thermal Image Plane
Reference Point (head)
Reference Point (head)
Reference Point (centroid)
Reference Point (centroid)

Figure 3. HGA-based search scheme to estimate the four 2-D point locations in color images.

2.2. Preliminary Human Silhouette Extraction and Correspondence Initialization

Assuming that human is the only moving object in the scene, it can be extracted by a simple background subtraction method. To model the background, we choose some frames from color image sequences which only contain 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

$$\begin{array}{rrcl} & |r(X,Y) - \mu_r(X,Y)| & > & \beta\sigma_r(X,Y) \\ & |g(X,Y) - \mu_g(X,Y)| & > & \beta\sigma_g(X,Y) \\ \text{or} & |b(X,Y) - \mu_b(X,Y)| & > & \beta\sigma_b(X,Y) \end{array}$$

where r, g and b represent pixel color values of the input image; 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 arbitrary selected threshold.

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

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

where t represents the pixel thermal value in the input thermal 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 thermal images so that the extracted body silhouettes can be compared at the same level.

After body silhouettes are extracted from each color image and its corresponding thermal image, the centroid of the silhouette region and the extreme top point of the head region are computed as the initial correspondence between each color image and its corresponding thermal image.

2.3. Automatic Image Registration

In the projective transformation model, 4 pairs of correspondence are required. For each pair of color and thermal images, we only have two pairs of initial correspondence, i.e., centroid and head top point. In our approach, we choose 4 pairs of initial correspondence from two image pairs in the given color and thermal image sequences. For image registration between two image sequences using projective model, the assumption of planar object surface must be satisfied. That is, human should walk along the same direction in the scene so that the human body surface from the camera view in each frame lies on the same plane over the whole sequence.

Due to the physical difference of objects in EO and IR spectrum, the initial correspondence is not accurate to compute transformation model parameters. Instead, the preliminary extracted body silhouette regions provide more reliable information. Therefore, we propose a method to perform a least squares fit of the transformed color silhouette to the thermal silhouette. That is, we estimate the set of model parameters $\mathbf{p}$ to minimize

$$error(\mathbf{p}; I_i; C_i) = \sum_{i=1}^{N} \sum_{X,Y \in I_i} (T_{C_i;\mathbf{p}}(X,Y) - I_i(X,Y))^2, \tag{3}$$

where I is the silhouette binary image obtained from thermal images, C is the silhouette binary image obtained from color images, $T_{C;\mathbf{p}}$ is the transformed binary image of C by projective transformation with parameter set $\mathbf{p}$, N is the number of color and thermal image pairs, and X, Y are image plane coordinates in I.

A Genetic Algorithm (GA) is appropriate to solve this optimization problem. However, the GA cannot be implemented without the knowledge the range of 8 parameter values associated with the projective transformation model in Equation (2). If we fix 4 points in a thermal image, the 2-D coordinates of 4 corresponding points in the corresponding color image can be used as parameters of the projective transformation model. Because the image size is limited, the parameter ranges can be easily determined. In this paper, we proposed a Hierarchical Genetic Algorithm (HGA) based search scheme to estimate the model parameters (four 2-D point coordinates) as shown in Figure 3.

First, we choose the estimated human silhouette centroids and head top points (as mentioned in Section 2.2) from two thermal images in the same infrared video as 4 reference points. The corresponding 4 points estimated from the two corresponding color images are chosen as the initial correspondence. At the first iteration of the scheme, a GA is applied to estimate the 4 correspondence coordinates according to Equation 3 within a $w \times w$ window centered at each initial correspondence location, where w is the square side length in pixels. After the GA is converged, we obtain a new estimate of correspondence coordinates. At the second iteration, a GA is applied within a reduced $w/2 \times w/2$ window centered at each of the estimated location from the first iteration. Similarly, at next iteration, a GA is applied within a window whose side length is reduced to the half from the previous iteration, and the center of each window is the estimated location from the previous level. This procedure is repeated until the window size is lower than a pre-selected threshold.

In the proposed approach, the bit length of parameters in each GA can be small 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 minimum. Even if the real correspondence is located out of the initial $w \times w$ window, the approach still have the possibility to find a good estimate because the new window might cover areas out of the initial window. After the correspondence in color images are located, the transformation is uniquely determined, and will be used to transform color images onto the thermal image plane.

2.4. Sensor Fusion

To improve the accuracy of human silhouette extraction, we need to combine the information from the registered color and thermal images. If the human silhouette extraction is viewed as a classification procedure, the commonly used classifier combination strategies can be employed here. Kittler et al. [12] demonstrate that the commonly used classifier combination schemes can be derived from a uniform Baysian framework under different assumptions and using different approximations. Similar strategies can be applied in body silhouette extraction by combining registered color and thermal images as follows:

- Product rule: $(X,Y) \in S$,
  if $P(\mathbf{c}(X,Y) \in S)P(t(X,Y) \in S) > \tau$
- Sum rule: $(X,Y) \in S$,
  if $P(\mathbf{c}(X,Y) \in S) + P(t(X,Y) \in S) > \tau$
- Max rule: $(X,Y) \in S$,
  if $\max\{P(\mathbf{c}(X,Y) \in S), P(t(X,Y) \in S)\} > \tau$
- Min rule: $(X,Y) \in S$,
  if $\min\{P(\mathbf{c}(X,Y) \in S), P(t(X,Y) \in S)\} > \tau$

where (X,Y) represents the 2-D image coordinate, S represents the human silhouette, $\mathbf{c}$ represents the color value vector, and t represents the thermal value. The estimate of probability is computed as

$$P(\mathbf{c}(X,Y) \in S) = 1 - e^{-||\mathbf{c}(X,Y) - \mu_c(X,Y)||^2} \quad (4)$$

$$P(t(X,Y) \in S) = 1 - e^{-|t(X,Y) - \mu_t(X,Y)|^2}, \quad (5)$$

where μ_c represents the mean background color value vector, and μ_t represents the mean background thermal value.

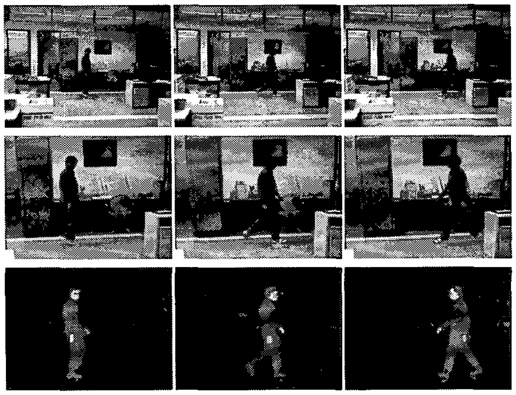

Figure 4. Examples of registration results: first row - original color images, second row - transformed color images, third row - original thermal images.

3. Experimental Results

The image data used in our experiments are real human walking data recorded 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. Thermal images are recorded by a long-wave IR camera with image size of 240 × 320 as shown in the first row of Figure 4. Both cameras have fixed but different focal lengths. The IR camera has a higher resolution and less distortion than the PC camera, and is, therefore, used as the base camera. The color images are transformed and then fused with the original thermal images in our experiments.

3.1. Image Registration Results

Three color and thermal images are selected for least square fitting in Equation 3. The initial search window is set as 80 × 80 pixels, and the final search window is 10 × 10 pixels. In GA at each level, we use 4 bits to represent each coordinate(totally 32 bits for 8 coordinates); fitness function is the least square error in Equation 3; 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 20 successive steps. Figure 5 shows examples of estimated transformation results at different levels. Even though the original transformation 5(c) is far away from the real transformation. The transformation results are improved gradually at successive levels and finally converged around the real transformation. Figure 4 shows the comparison of original color images, transformed color images, and original thermal 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 thermal image and the transformed color image, respectively. According to this definition, the registration precision for the 3 image pairs in Figure 4 is 77%, 80% and

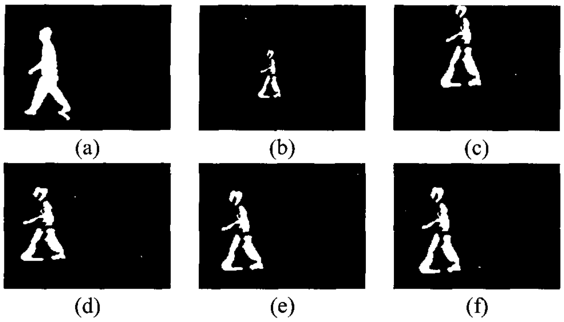

Figure 5. Examples of estimated transformation results at different levels: (a) original silhouette from the thermal image, (b) original silhouette from the color image, (c) initial transformation of (b), (d) after the first iteration, (e) after the second iteration, (f) after the final iteration.

85%, respectively. Considering that the color and thermal image pairs are not exactly simultaneously obtained, and there are labeling errors due to the physical difference between color and thermal signals, our image registration still achieves good results.

3.2. Sensor Fusion Results

We evaluate the human silhouette detection performance by the receiver operating characteristic (ROC) curves [13]. Let N be the number of moving object pixels in the Ground Truth image, α be the number of moving object pixels that the algorithm did not detect, B be the number of background pixels in the Ground Truth image, and β be the number of background pixels that were detected as foreground. The Ground Truth image in our experiments are manually labeled from the original thermal images. The confusion matrix is given in Table 1. We can define the Probability of detection and Probability of false alarms as

$$Pd = (N - \alpha)/N \quad \text{and} \quad Pf = \beta/B. \tag{6}$$

If we evaluate M images as a whole, the equations become

$$Pd = \sum_{i=1}^{M}(N_i - \alpha_i)/\sum_{i=1}^{M} N_i \quad \text{and} \quad Pf = \sum_{i=1}^{M}\beta_i/\sum_{i=1}^{M} B_i. \tag{7}$$

Equation (7) are used to obtain the ROC curves for detection performance evaluation of different fusion strategies which are shown in Figure 6.

This figure shows that the product, sum and max fusion rules achieve better results than only using color or thermal classifiers. Among these rules, the sum rule achieves the best results. Considering that the image resolution of the thermal camera is higher than that of the EO camera, and thermal cues are generally more reliable than color cues, the thermal classifier has much higher confidence than the color classifier. We believe that the main reason for the good performance achieved by sum rule is its robustness to errors (or noise) especially from the color classifier [12]. Product rule considers more color information, so it is sensitive to the noise from color classifier especially when the false alarm is low. Max rule considers less color information with low confidence, so its performance is higher than that of the thermal classifier but lower than sum rule. The performance of min rule is even worse than that of using thermal only because it mainly focus on the color information with low confidence. Figure 7 shows the human silhouette extraction results by combining color and thermal image pairs with different strategies. The thresholds for these rules are chosen as the smallest values such that shadows in both color and thermal images are eliminated.

	Ground Truth Foreground	Ground Truth Background
Detected Foreground	$N - \alpha$	β
Detected Background	α	$B - \beta$

Table 1. Confusion matrix.

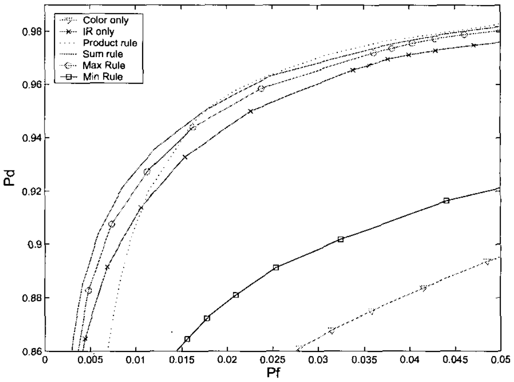

Figure 6. ROC curves for detection performance evaluation of different fusion strategies for silhouette detection.

4. Conclusions

In this paper, we approach the task of human silhouette extraction from color and thermal image sequences using automatic image registration. A hierarchical Genetic Algorithm (HGA) based scheme is employed to find correspondence so that the preliminary silhouettes form the color and thermal images are well matched. The obtained correspondence and corresponding transformation are used for image registration in the same scene. Registered color and thermal images are combined by probabilistic strategies to obtain better body silhouette extraction results. Experiments show that the proposed approach achieves good performance for image registration between color images and thermal image sequences. It is also shown that each of the product, sum, max and min fusion rules achieves better performance on

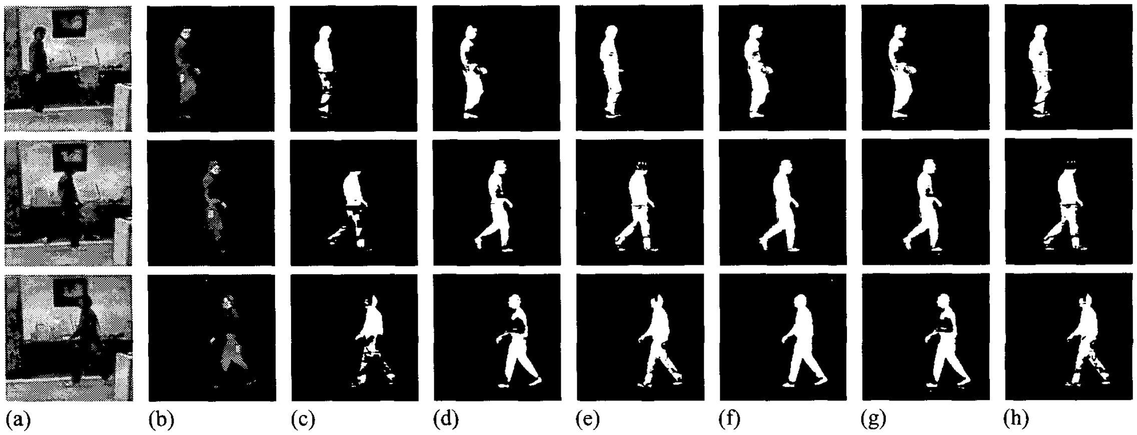

Figure 7. Examples of fusion results: (a) transformed color images, (b) original thermal images, (c) silhouette from (a), (d) silhouette from (b), (e) silhouette from product rule fusion, (f) silhouette from sum rule fusion, (g) silhouette from max rule fusion, (h) silhouette from min rule fusion.

silhouette detection than color or thermal images used individually. Among these rules, sum rule achieves the best results.

5. Acknowledgment

This work was supported in part by grants F49620-02-1-0315 and DAAD19-01-0357; the contents and information do not necessarily reflect the position or policy of U.S. Government.

References

[1] S.A. Niyogi and E.H. Adelson, "Analyzing and recognizing walking figures in xyt," in *Proc. IEEE Conference on CVPR*, pp. 469–474, 1994.

[2] J.J. Little and J.E. Boyd, "Recognizing people by their gait: the shape of motion," *Videre: Journal of Computer Vision Research*, vol. 1, no. 2, pp. 1–32, 1998.

[3] H. Murase and R. Sakai, "Moving object recognition in eigenspace representation: gait analysis and lip reading," *Pattern Recognition Letters*, vol. 17, no. 2, pp. 155–62, 1996.

[4] P.S. Huang, C.J. Harris, and M.S. Nixon, "Recognizing humans by gait via parameteric canonical space," *Artificial Intelligence in Engineering*, vol. 13, pp. 359–366, 1999.

[5] B. Bhanu and J. Han, "Kinematic-based human motion analysis in infrared sequences," in *Proc. IEEE Workshop on Applications of Computer Vision*, pp. 208–212, 2002.

[6] H.D. Arlowe, "Thermal detection contrast of human targets," in *Proc. IEEE International Carnahan Conference on Security Technology*, pp. 27–33, 1992.

[7] J. Wilder, P.J. Phillips, Cunhong Jiang, and S. Wiener, "Comparison of visible and infra-red imagery for face recognition," in *Proc. Internatinal Conference on Automatic Face and Gesture Recognition*, pp. 182–187, 1996.

[8] Y. Yoshitomi, Sung-Ill Kim, T. Kawano, and T. Kilazoe, "Effect of sensor fusion for recognition of emotional states using voice, face image and thermal image of face," in *Proc. IEEE International Workshop on Robot and Human Interactive Communication*, pp. 178–183, 1996.

[9] V.R. Mandava, J.M. Fitzpatrick, and D.R. III Pickens, "Adaptive search space scaling in digital image registration," *IEEE Transactions on Medical Imaging*, vol. 8, no. 3, pp. 251–262, September 1989.

[10] P.A. van den Elsen, E.-J.D. Pol, and M.A. Viergever, "Medical image matching-a review with classification," *IEEE Engineering in Medicine and Biology Magazine*, vol. 12, no. 1, pp. 26–39, March 1993.

[11] J. Yao, "Image registration based on both feature and intensity matching," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 3, pp. 1693–1696, 2001.

[12] J. Kittler, M. Hatef, R. Duin, and J. Matas, "On combining classifiers," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 3, pp. 226–239, 1998.

[13] S. Nadimi and B. Bhanu, "Multistrategy fusion using mixture model for moving object detection," in *Proc. International Conference on Multisensor Fusion and Integrationfor Intelligent Systems*, pp. 317–322, 2001.