# Registration of IR and EO Video Sequences based on Frame Difference

Zheng Liu and Robert Laganière
School of Information Technology and Engineering
University of Ottawa
800 King Edward Ave.
Ottawa, ON K1N6N5 Canada
E-mail: laganier@site.uottawa.ca

## Abstract

*Multi-modal imaging sensors are employed in advanced surveillance systems in the recent years. The performance of surveillance systems can be enhanced by using information beyond the visible spectrum, for example, infrared imaging. To ensure correctness of low- or high-level processing, multi-modal imagers must be fully calibrated or registered. In this paper, an algorithm is proposed to register the video sequences acquired by an infrared and an electro-optical (CCD) camera. The registration method is based on the silhouette extracted by differencing adjacent frames. This difference is found by an image structural similarity measurement. Initial registration is implemented by tracing the top head points in consecutive frames. Finally, an optimization procedure to maximize mutual information is employed to refine the registration results.*

## 1. Introduction

Vision systems working beyond visible spectrum are becoming affordable assets to advanced surveillance systems. The performance of these systems can be enhanced through taking full advantage of the information available across the electromagnetic spectrum. This makes the surveillance system more robust and reliable under different conditions, such as noisy and cluttered background, poor lighting, smoke, and fog. The technique to achieve this goal is known as information or sensor fusion. Depending on the requirements, the fusion of multi-modal images can be implemented at different levels using various fusion algorithms [1, 4].

The infrared (IR) camera uses thermal detector to measure the difference in infrared radiation of different objects, i.e. the variance of thermal emissivity properties. The electro-optical (EO) sensor, e.g. CCD or CMOS cameras, captures the reflective light properties of objects [6]. Therefore, the visual and IR imagery provide the complementary information about the scene [6]. Multiple cues provided by the two imaging modalities can be used to achieve detection, tracking, and content analysis for the surveillance applications. However, preceding to any further processing, the EO and IR images from the video sequences should be registered so that the corresponding pixels in the two images are associated with the same physical points in the scene. This ensures the correctness of pixel- and high-level processing.

The image registration consists of four basic steps: feature detection, feature matching, mapping function design, and image transformation and resampling [12]. Li et al. registered multi-sensor images with image contours [7]. In another publication of Li et al. [8], they used a wavelet-based approach to detect image contour and located feature points by using local statistics of image intensity. The feature points were matched with a normalized correlation method. Coirs et al. matched the triangles formed by grouped straight line segments extracted from the IR and EO images [3]. However, the physical correspondences may not be fully detected with matchable contours or lines. The same scene may appear totally different in two image modalities. Han et al. suggested using the silhouette of a moving human body to register IR and EO images. They found the silhouette by classifying a pixel as belonging to either foreground or background based on the background Gaussian distribution [5]. The centroid and head top points in two pairs of images were used as control points. A genetic algorithm was employed to minimize the registration error function. In [11], Ye et al. proposed using zero-order statistics to detect moving object in a video sequence. Through tracking the feature points, an iterative registration algorithm is implemented. Related work was also reported by Maes et al. and Chen et al. in [9, 2], where the registration is carried out based on maximizing mutual information of two image regions. However, in [2] the images

must be roughly registered with some prior knowledge in the surveillance application.

The moving object detection, which is also known as the background maintenance, still remains a challenge for surveillance applications. For the millimeter wave (MMW) video sequence, such detection can be more difficult due to it blurry nature [2]. In this paper, we propose a registration method that uses the silhouette of the frame difference instead of the silhouette of moving objects; therefore, this method does not reply on the success of foreground detection and can be applied to any imaging modality. The frame difference can be steadily detected with the structural similarity measurement. Instead of extracting feature points in one image, the trajectory formed by the head top points in consecutive frames is used for initial registration. A refining process is implemented based on the maximum mutual information method [2].

The rest of the paper is organized as follows. The detailed procedure for registration is described in section (2). The whole process consists of two steps, i.e. initial registration and parameter refinement. Experimental results are presented in section (3). Discussion and conclusion can be found in section (4) and (5) respectively.

## 2. Registration based on Frame Difference

The proposed registration process can be implemented in two steps. In the first step, the head top points are detected from the silhouette of frame difference. The initial parameters can be estimated by matching the trajectories in IR and EO sequences. The second step is to refine the registration parameters by directly registering two regions of interest with the mutual information maximization method. In the current description of our approach, we assume that only one person is present in the video sequences as a moving object, although it is possible to deal with multiple points in one frame.

### 2.1. Image Similarity Measurement

The simplest way to find the difference between two images $x$ and $y$ is the subtraction operation. However, the threshold may vary with different video clips. In this work, we use the structural similarity (SSIM) measurement to detect the difference between consecutive frames. The SSIM is defined as [10]:

$$SSIM\left(x,y\right)=\left[l\left(x,y\right)\right]^{\alpha}\left[c\left(x,y\right)\right]^{\beta}\left[s\left(x,y\right)\right]^{\gamma} \quad (1)$$

where there are:

$$l\left(x,y\right)=\frac{2\mu_x\mu_y+C_1}{\mu_x^2+\mu_y^2+C_1} \quad (2)$$

$$c\left(x,y\right)=\frac{2\sigma_x\sigma_y+C_2}{\sigma_x^2+\sigma_y^2+C_2} \quad (3)$$

$$s\left(x,y\right)=\frac{\sigma_{xy}+C_3}{\sigma_x\sigma_y+C_3} \quad (4)$$

By setting $\alpha=\beta=\gamma=1$ and $C_3=C_2/2$, equation (1) becomes:

$$SSIM\left(x,y\right)=\frac{\left(2\mu_x\mu_y+C_1\right)\left(2\sigma_{xy}+C_2\right)}{\left(\mu_x^2+\mu_y^2+C_1\right)\left(\sigma_x^2+\sigma_y^2+C_2\right)} \quad (5)$$

In equation (1), there are three components. The first one, $l(x,y)$, measures how the mean luminance differs between the two images while the second ($c(x,y)$ estimates the contrast. The third one $s(x,y)$ is the correlation.

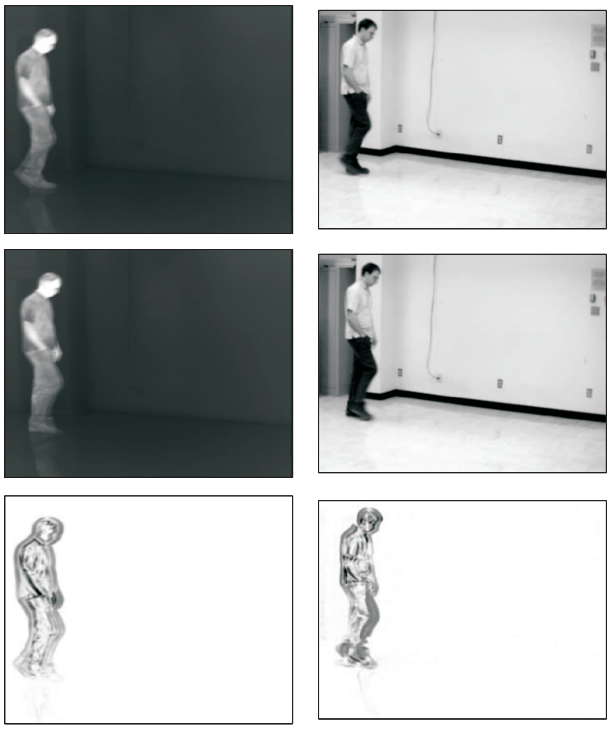


**Figure 1. The SSIM measure. On the left column are the IR images. Right column is from EO camera. Two adjacent frames and their SSIM map are from the top to bottom.**

An example of applying SSIM to find the frame difference is given in Figure 1. The SSIM maps are generated from two adjacent frames for IR and EO sequences respectively. The mean value of the SSIM map gives an index value, which indicates how different the two images are. In our application, we use the SSIM map.

### 2.2. Silhouette Extraction

Once the SSIM maps are obtained. The detection of the frame difference is straightforward. Simply applying a fixed threshold value to both SSIM maps, two binary images can be obtained. After morphologic operations, the binary images are scanned from top to bottom and filled with "1" between the left and right edges as shown in Figure 2.

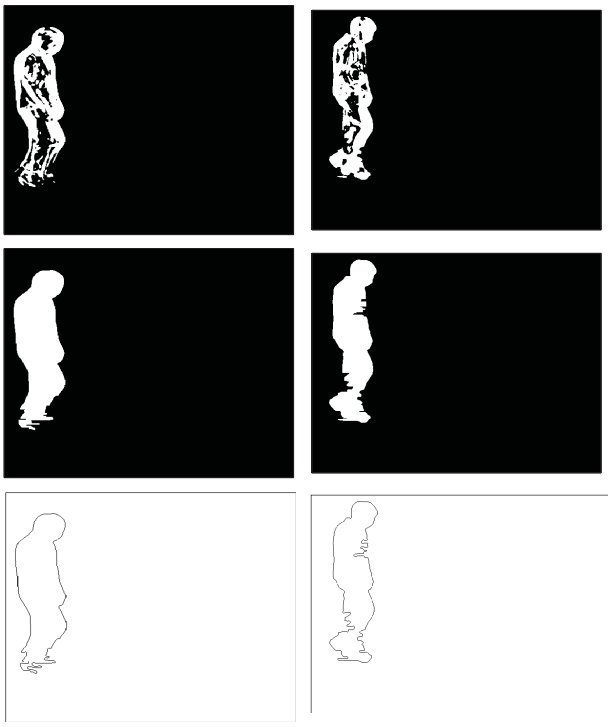


**Figure 2. The thresholded binary images from SSIM maps are on the top, the postprocessed results on middle, and on bottom are the extracted contours extracted.**

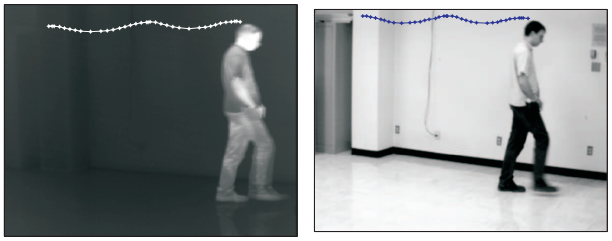
In the experiment, we set the threshold value as $0.6$ for both the IR and EO images. The contour of the silhouette is detected with zero-cross based edge detection. The top head points are searched from each frame and used for initial parameter estimation.

### 2.3. Parameter Estimation and Refinement

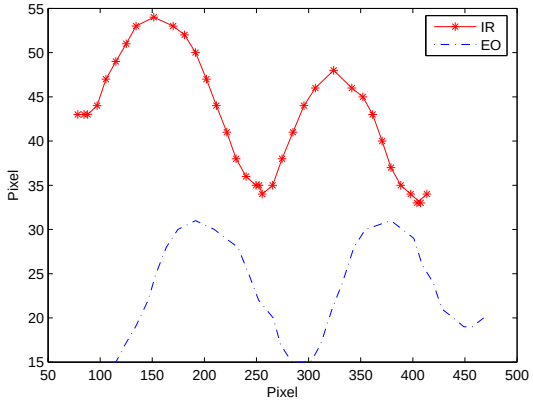
It is reasonable to assume that the IR and EO cameras are mounted in parallel, which means we can omit the rotation between the frames acquired by the two cameras. Therefore, a 2-D homogeneous transform can describe the geometric relation between the two frames. If IR image is used as a reference, there is:

$$\begin{cases} X_{IR} = kX_{EO} + \Delta X \\ Y_{IR} = kY_{EO} + \Delta Y \end{cases} \tag{6}$$

where $k$ stands for the scaling parameter and $\{\Delta X, \Delta Y\}$ are the translating parameters. There are three parameters to be found in total.



(a) The top head points.




(b) The trajectory of top head points.

**Figure 3. The top head points in two video sequences.**

Assuming the top head points in IR image correspond to the head top points in EO image, we can solve equation (6) with the least square method. Figure 3 shows the trajectory of head top points from IR and EO sequences. The initial estimation can be obtained by solving the equation (6) given the corresponding head points. However, these points may not be exactly matched. The initial registration can be further refined by applying a mutual information based registration approach [2, 6].

We can use the binary maps in Figure 2 to find the region of interest (ROI) easily as shown in Figure 4. Note that binary map can extract the corresponding ROI for any two adjacent frames.

The definition of mutual information (MI) for two discrete random variables $X$ and $Y$ is:

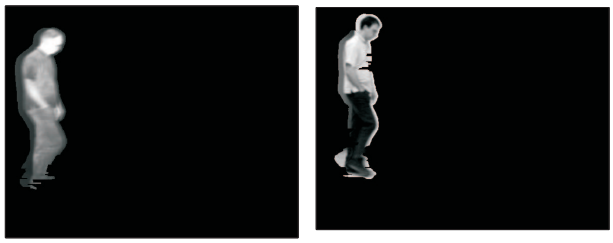


**Figure 4. The regions of interest from two frames.**

$$I(X;Y) = \sum_{y \in Y} \sum_{x \in X} p_{XY}(x,y) \log_2 \frac{p_{XY}(x,y)}{p_X(x)\, p_Y(y)} \quad (7)$$

where $p_{XY}(x,y)$ is the joint probability distribution function of $X$ and $Y$, and $p_X(x)$ and $p_Y(y)$ are the marginal probability distribution functions of $X$ and $Y$ respectively. Actually, MI quantifies the distance between the joint distribution of $X$ and $Y$, i.e. $p(x,y)$, and the distribution associated to the case of complete independence $p_X(x)p_Y(y)$ [9, 2]. For the IR and EO image, the joint probability distribution can be obtained from the image's histogram. In equation (7), $p_{XY}(x,y)$ is replaced by the normalized joint grey level histogram of the IR and EO image, that is:

$$p_{XY}(x,y) \leftarrow h_{IE}(l,m) = \frac{g(l,m)}{\sum_{l,m} g(l,m)} \quad (8)$$

where $g(l,m)$ is the joint histogram of IR and EO image. We can use 256 for both $L$ and $M$. The marginal probabilities are represented by normalized marginal histogram of IR and EO image. There are:

$$p(x) \leftarrow h_I(l,m) = \sum_l h_{IE}(l,m) \quad (9)$$

$$p(y) \leftarrow h_E(l,m) = \sum_m h_{IE}(l,m) \quad (10)$$

Mutual information can be equivalently expressed with joint ($H(L,M)$) and marginal entropies ($H(L), H(M)$):

$$I(L;M) = H(L) + H(M) - H(L,M) \quad (11)$$

where there are:

$$H(L) = -\sum_l h_I(l,m) \log_2 h_I(l,m) \quad (12)$$

$$H(M) = -\sum_m h_E(l,m) \log_2 h_E(l,m) \quad (13)$$

$$H(L,M) = -\sum_{l,m} h_{IE}(l.m) \log_2 h_{IE}(l,m) \quad (14)$$

The registration is to transform the EO image to the coordinate of the IR image. When the transformed image is aligned with the reference, the MI value is maximized. Thus, searching the transform parameters that maximize MI give the registration result. Similarly, we use simplex search method as proposed by Chen et al. [2]. The implementation of the simplex search algorithm is available in Matlab® as a function named "fminsearch".

## 3. Experimental Results

In the experiment, we register two clips from IR and EO video sequence (30fps). The threshold value to get the binary image is set as $0.6$ for both the clips. The grey level for IR and EO images is rounded to $0 \sim 255$. The initial estimation of the registration parameters from head top trajectory are $\{k = 0.9495; \Delta X = 20.943; \Delta Y = -28.9725\}$. The refinement of this result is carried out for the thirty-five frames in the two clips. The results are shown in Figure 5. Table 1 lists the mean, maximum, and minimum value of the parameters.

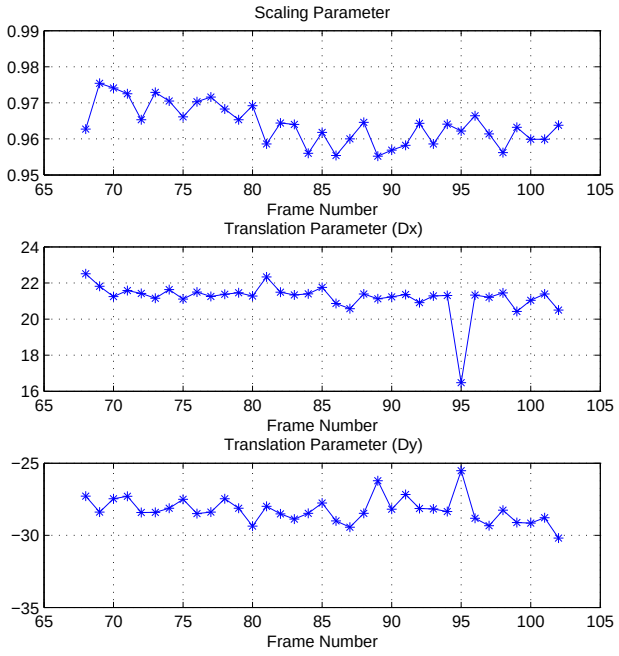



**Figure 5. The refined registration results based on MI.**

Using the mean value of the refined registration parameters, the EO frames are transformed and registered with IR frames as shown in Figure 6. The human body is segmented

**Figure 6. The registration results. Top: IR frames; $2^{nd}$ row: EO frames; $3^{th}$ row: transformed EO frames; bottom: the synthesized images.**

**Table 1. The registration parameters obtained by maximum MI.**

| | Mean | Max | Min |
|---|---|---|---|
| $k$ | 0.9640 | 0.9754 | 0.9552 |
| $\Delta X$ | 21.19 | 22.51 | 16.47 |
| $\Delta Y$ | -28.25 | -25.52 | -30.19 |

and embedded in the EO frames. The synthesized images indicate how well the two sequences are registered.

## 4. Discussion

The centroid of human body could be another feature point for registration as described in [5]. One precondition is that a "clear" silhouette of human body should be obtained. In the proposed method, the centroid points are not used for registration, because the shadow on the floor makes the bottom boundary indistinct and the centroid point cannot be steadily detected.

The translation parameters obtained from $95$ frame appears to be a outlier. There are two possible reasons contributing to such variance. The first one is that detected head top point may not be accurate. The second is that the shapes of the two silhouette are not the same.

Although we assume that there is no rotation between two frames, such angular difference may be considered

when the registration is refined with maximum MI. In our case, the rotation parameter searched by maximum MI is around 0.0003 rad; therefore, we do not have to consider it.

The registration of multi-modal video sequences does not have to be implemented in real time, only if the configuration of the cameras does not change dynamically. In this paper, the accuracy of the registration is not studied. It is much meaningful to discuss the accuracy when further processing is considered. How the accuracy will affect the result should be investigated in future work.

## 5. Conclusion

In this paper, a registration method for multi-sensor video sequences is proposed. The approach is based on registering trajectory of the head top points detected from the silhouette of frame difference, which is found by the structural similarity measurement. Such difference can be used to find the region of interest. The refinement of the initial registration is implemented by maximizing the mutual information of the detected regions of interest. The advantage of this technique is that it is not necessary to segment the exact silhouette of the moving object from the video sequence, which is difficult for imaging modality like millimeter wave. Secondly, the proposed method tries to use individual feature point in multiple frames rather than matching multiple points from one image. This makes the registration process easily implemented and the initial result is close to the refined one.

## 6. Acknowledgement

The experiments described in this paper were validated with video sequences recorded for the MONNET project which was funded by Precarn Inc and leaded by the Computer Vision and Systems Laboratory at Laval University.

## References

[1] R. S. Blum, Z. Xue, and Z. Zhang. An overview of image fusion. In R. S. Blum and Z. liu, editors, *Multi-Sensor Image Fusion and Its Applications*, chapter 1, pages 1–35. Taylor and Francis, 2005.

[2] H. M. Chen, P. K. Varshney, and M. A. Slamani. On registration of regions of interest (ROI) in video sequences. In *Proceedings of the IEEE Conference on Advanced Video and Signal Based Surveillance*, pages 313–318, Los Alamitos, CA, USA, 2003.

[3] E. Coiras, J. Santamaria, and C. Miravet. Segment-based registration technique for visual-infrared images. *Optical Engineering*, 39(1):282–289, January 2000.

[4] G. L. Foresti and L. Snidaro. A distributed sensor network for video surveillance of outdoors. In G. L. Foresti, C. S. Regazzoni, and P. K. Varshney, editors, *Multisensor Surveillance Systems*. Kluwer Academic Publishers, 2002.

[5] J. Han and B. Bhanu. Detecting moving humans using color and infrared video. In *Proceedings of IEEE Conference on Multisensor Fusion and Integration for Intelligent Systems*, pages 228–233, 2003.

[6] S. Krotosky and M. Trivedi. Multimodal stereo image registration for pedestrian detection. In *Proceedings of Intelligent Transportation Systems*, pages 109–114, Toronto, Canada, September 2006.

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

[8] H. Li and Y.-T. Zhou. Automatic Visual/IR image registration. *Optical Engineering*, 35(2):391–400, February 1996.

[9] F. Maes, A. Collignon, D. Vandermeulen, G. Marchal, and P. Suetens. Multimodality image registration by maximization of mutual information. *IEEE Transaction on Medical Imaging*, 16(2):187–198, April 1997.

[10] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli. Image quality assessment: from error measurement to structural similarity. *IEEE Transactions on Image Processing*, 13(1), 2004.

[11] G. Ye, J. Wei, M. R. Rickering, M. R. Frater, and J. F. Arnold. Simultaneous tracking and registration in a multisensor surveillance system. In *Proceedings of ICIP*, volume 1, pages 933–936, Sept 14-17 2003.

[12] B. Zitova and J. Flusser. Image registration methods: A survey. *Image and Vision Computing*, 21:977–1000, 2003.