Multisensor Image Registration and Optical Correlation

Yunlong Sheng[1], Xiangjie Yang[1], Daniel McReynolds[1], Piere Valin[2]
and Leandre Sevigny[3]

[1]Image Science group, Center for Optics, Photonics and Laser (COPL)
Dept. of Physics, Laval University, Ste-Foy, Quebec, Canada G1K 7P4;
[2]Lockheed Martin,6111 Royalmount Ave. Montreal Qc Canada H4P 1K6;
[3]Defence Research Establishment Valcartier, 2458 Boul. Pie XI Nord, C.P.
8800, Courcette Qc. Canada G0A 1R0.

Abstract

We present several approaches for visible and infrared video image sequence registration, useful for image fusion, target detection and recognition. Feature inconsistency and low contract and noise in the infrared image background consist of the principle difficulty in the image registration for the well separated spectral bands visible and IR images. Possibility of using and integrating the optical correlator into the operational systems is discussed.

Keywords: Image registration, correlation, Hausdorff distance

Contents

1999 Euro-American Workshop Optoelectronic Information Processing: A Critical Review,
edited by Philippe Refregier, Bahram Javidi, Proc. of SPIE Vol. 10296 (Vol. CR74), 1029609
· CCC code: 0277-786X/17/$18 · doi: 10.1117/12.365906

1. Introduction

Multiple video cameras of different spectral bands are widely used in the advanced vision systems to capture more information in biomedical imaging, remote sensing and other imaging applications. In many cases, images from multiple sensors need to be integrated into a single synthetic representation for human observers, who are in general under time constraints, stress and workload for interpretation, detection and decision making. Image registration is a key step before fusion. Image registration is also needed for interpreting time evolution of the images in remote sensing and medial diagnostics.

Image registration applied to a single video sequence is image video sequence stabilization. The primary means available for image stabilization is an electromechanical stabilizing platform, which is bulky and expensive. Its performance is degraded with vibration in the critical 0 - 20 Hz range. An automatic electronic image stabilizer should be able to first estimate components of the scene motion due to the camera movement and then eliminate those components by warping each frame into precise alignment with the next frame. Then, if a temporal filter is applied to compute the frame difference, then the background scene would be eliminated and the moving targets would be highlighted.

Multiple sensors can act in a synergistic manner. For instance, in two broad band visible/infrared battle field image sequences, soldiers and trucks can be hidden behind the smoke in the visible image, but they appear clearly as high contrast hot objects in the IR images. However, the contrast is extremely low in background of the IR images, so that one can not locate the hot objects within the background, and one needs to fuse the IR images with the corresponding visible images.

Challenge in the multiple sensor image registration is to align two images in spite of feature inconsistency. Some features in one image can do not show up in the image from another sensor. This is the feature inconsistency problem. Multiple sensors capture distinguished signatures from the input scene in different spectral bands. Multisensor imaging system should maximize independence of the acquired data. This is natural, since if one sensor captures images that are similar or correlated to the images already obtained by other sensors, then this sensor provides no additional information and should be removed from the system. The signature independence implies the features inconsistency.

Image registration and fusion are implemented in the practical image sensing systems. These are a potential application area of optical image processing. Optical correlators at the operating rate of 1000 correlations per second have been now built. Automatic target recognition and tracking using the optical correlators and optimally designed correlation filters have been demonstrated. However, many practical systems should accomplish complex tasks using a sequence of image processing algorithms such as detection, recognition, identification and tracking, which can be entirely automatic or assisted with human operator. In general, the tasks of practical and complex image processing systems can not be accomplished with a single short of correlation with an optical correlator. Thus, the high-speed optical correlator must be applied and integrated into the multi-step effective numerical image processing systems.

In this paper we present problematic of image registration, review the existing methods and show the algorithms that we developed to accomplish IR/visible battle field image registration. We show that the correlation is one of the basic operations very useful in many practical image processing systems, and the invariance problem is still the major issue associated in the correlation. Therefore, the algorithms developed for invariant optical pattern recognition filters could be useful in those systems. We choose the feature-based approach. We use multi-scale hierarchical edge detection, edge focusing and edge salience measure to extract salient edges from the low contrast and noisy IR image background. We use the Hausdorff distance measure for matching between the curves from two different modalities. We introduce the image partitioning technique in the Hausdorff distance matching, so that the global affine transformation is approximated by local translations. This approach speeds up significantly the computation of the Hausdorff distance.

2. 3-D projection

It is well known in the computer vision that 3-D space information may be obtained from two or more 2-D projection images of the same scene. When the camera is modeled as a pinhole camera, a point M in the 3-D space is projected into m_1 and m_2 in the two image planes, respectively, as shown in Fig.1, where c_1 and c_2 are the camera centers. One camera is rotated and translated in the 3-D space with respect to another. From the image point m_1 and camera center c_1 we know that the object point M should be on the projection line $m_1c_{1,}$ but its position on m_1c_1 is undetermined. If the image point m_2 of the same point M in image 2 is known, then the position of the point M in the 3-D space can be uniquely determined[1]. When M is move on the line m_1c_1 its image point m_2 would move along an epipolar line l_{m2}, in image 2. Reciprocally, when M moves along the projection line m_2c_2, its onto image 1 would move along an epipolar

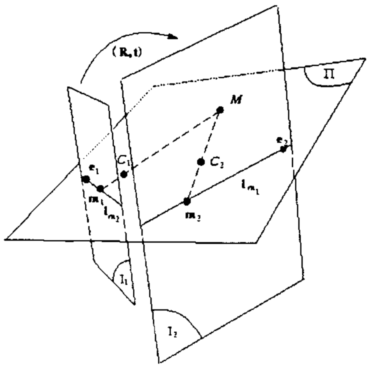

Fig.1 Epipolar Geometry

line l_{m1} in image 1. The epipolar lines l_{m1} and l_{m2} are the intersections between the plane Π formed by the object point M and two camera centers c_1 and c_2 and the image plans I_1 and I_2. The key is to establish correspondence between points m_1 and m_2 in images 1 and 2, such that they are the images of the same object point M.

Under the pinhole camera model a 3-D scene is projected into a 2-D image through the full perspective projection. Let $\vec{x}_s$ be the world coordinate system and $\vec{x}_c$ be the camera geometric coordinate system, whose origin is on the center and the axis z_c is on the optical axis of the camera, then the relation between the world system and the camera system is

$$\begin{pmatrix} x_c \\ y_c \\ z_c \\ 1 \end{pmatrix} = \begin{pmatrix} R & T \\ 0^T & I \end{pmatrix} \begin{pmatrix} x_s \\ y_s \\ z_s \\ 1 \end{pmatrix} \tag{1}$$

where **R** is the rotation matrix and **T** is the translations vector in the 3-D space of the world system with respect to the camera system. In the camera system an object point (x_c, y_c, z_c) is projected onto the image plane at (x, y) by

$$z_c \begin{pmatrix} x \\ y \\ 1 \end{pmatrix} = \begin{pmatrix} f & 0 & 0 & 0 \\ 0 & f & 0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix} \begin{pmatrix} x_c \\ y_c \\ z_c \end{pmatrix} \tag{2}$$

where f is the focal length of the camera. The combination of Eqs.1 and 2 describes the perspective projection from a 3-D space point (x_s, y_s, z_s) into its image point (x,y). For the sake of the simplicity we do not introduce the pixel coordinates of the camera here.

If the camera's position and orientation in the 3-D space and the camera intrinsic parameters, such as the focal length, aspect ratio and optical center position, are known (calibrated cameras), according to that shown in Fig.1 the object point can be reconstructed in the 3-D space from the two corresponding image points in the two images. This is a 3-D stereo vision problem. On the other hand, from a number of pairs of corresponding points, one can determine the camera rotation and translation matrices **R** and **T.** The is a motion estimation problem.

Image registration needs to determine the deformation of images due to the camera motion and view angle changes in order to realign the images. The registration function describes point (x_2, y_2) in image 2 as a function of point (x_1, y_1) in image 1, which can be derived by removing the object point $M(x_s, y_s, z_s)$ from Eqs.(1) and (2), that results in

$$\mathbf{m}_2^{\mathbf{T}}\mathbf{F}\mathbf{m}_1 = 0 \tag{3}$$

where $\mathbf{m}_i^{\mathbf{T}} = (x_i, y_i, 1)$ and **F** is known as the fundamental matrix of the two images, which depends on the position, orientation and intrinsic parameters of the camera. The fundamental matrix **F** has 9 unknown coefficients. If we can determine 8 matches between points in Images 1 and 2 we would be able, in general, to determine a unique solution for **F**, defined up to a scale factor, and then the registration function.

In the pinhole camera model, the perspective projection described in Eq.2 is non-linear. The coordinates (x,y) of the image point depend on the depth z_c of the object point. When the average distance from the camera to the center of mass of object L is much larger than the depth of the 3-D scene, $z_c = L + z_s$ and $L >> z_s$, one can replace z_c in the right-hand side of Eq.(2) by L, and then Eqs.(1) and (2) become linear, describing an approximation to the perspective projection. In this case, the registration function can be easily obtained as[2,3]

$$\begin{pmatrix} x_2 \\ y_2 \end{pmatrix} = \begin{pmatrix} p_1 & p_2 & p_3 \\ p_4 & p_5 & p_6 \end{pmatrix} \begin{pmatrix} x_1 \\ y_1 \\ 1 \end{pmatrix} + h(x_1, y_1) \begin{pmatrix} e_1 \\ e_2 \end{pmatrix} \tag{4}$$

where p_j with $j = 1, 2, ...6$ are the affine transformation coefficients relative to the 3-D rotation and translation of the camera, **e** is the epipolar vector and $h(x,y)$ is the height function of the scene, which introduces displacements of the image points (x_2, y_2) along the epipolar lines and proportional to the heights in the scene. The solutions of Eq.(4) for the six unknown coefficients p_j with $j = 1, 2, ...6$, $h(x,y)$ and **e** are not unique. Any linear polynomial $l(x,y) = ax + by + c$ can be added to $h(x,y)$, as long as $l(x,y)e$ is subtracted from the first term in the right-hand side of Eq.(4). The uniqueness of the solutions may be obtained if the solutions for p, h, and e satisfy a normalization constraints[2] that takes a planar approximation of the height function $h(x,y)$ and subtracts the planar approximate plane from Eq.(4). Thus, one first puts $h(x,y) = 0$ in Eq.(4) and then looks for the least mean square solution (LMS) to estimate the affine coefficients p_j. After the determination of p_j, the epipolar vector **e** and the scene height function $h(x,y)$ can be recovered from the residuals in the LMS fitting. The LMS solution for Eq.(4) needs to know a set of corresponding point pairs-i with $i = 1, ..., k$ with $k > 3$. When more correspondences k are available, it is advantageous to use all the points to improve the accuracy of the solution and detect outliers which are false correspondences in the set.

3. Image matching

To estimate the registration function Eqs.(3) or (4) one needs to determine a set of point matches between two images. The determination of point correspondences is the key step for 3-D stereo vision, motion analysis, image registration and model-based 3-D object recognition. When there are some characteristic points in the image, such as corners of 3-D objects or the landmarks in the remote sensing images, the point pairs are easy to determine. However, in outdoor battle field images the determination of point matches is a difficult task, especially for IR image registration, where one needs to detect and determine point matches in the static image background, where IR images have typically very low contrast due to the thermal equilibrium in the background.

3.1 Block matching

One approach is to partition the image into a number of sub-images, located in a regular grid as shown in Fig.2, and then define a central window in each sub-image as a template and correlate those block templates with their corresponding sub-images in another image[4]. These correlations can be implemented using the optical correlator.

The block matching results in a lattice of displacement vectors, which are evenly distributed over the image and are then useful to determine the transformation parameters for the registration function. However, the correlation can account only for translations. Other deformations, such as similarity or affine transform or local distortions[1] may be only approximated by the local translations of the blocks.

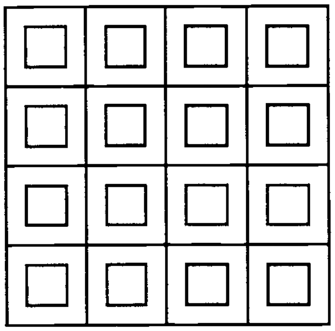

Fig.2 Block matching

The area-based approach for image registration utilizes full image information and can be applied to any images with rich or poor structure. Cross-correlation-based matched filter approach is optimal for the robustness against random noise. The drawback is that this approach can account only for small translations. The method can fail if the displacement of the bloc exceeds the size of the sub-image. Moreover, the higher the number of sub-images the higher precision of that approximation, however, the smaller translations that can be accounted for, and the less image features contained in the sub-images. The bloc matching is widely used in the video image processing and compression. With the small bloc size and small translations, the cross-correlation may be computed by electronic hardware in real-time of video rate. However, when the block size is large and its translation is large such that the search area is large, the computation of cross-correlation becomes very expensive. Optical correlator, such as the joint transform correlator, can be useful for the implementation of the bloc matching.

3.2 Feature matching

Feature-based image registration is to first detect image features, such as corners and edges, in the images and then determine the correspondences between the features in the two images, Finally, to fit the image transformation using the matches. This approach can account for any image deformation, and the processing speed is independent of image displacement. The feature based approach can be insensitive to multiple sensor modalities by selecting the structural salient features as will discussed below. It is fast to compute. Many powerful and versatile corner and edge detection, edge saliency techniques can be used. However, the feature-based approaches may fail to find matches in the structure-less areas. Its reliability depends on that of the feature extraction.

4. Image stabilization

Image stabilization is the registration of video sequence from a single camera. We choose the feature-based method which is more powerful to extract features in the IR image background and to account for image rotations and affine transformation[5].

We use the Harris-Stephens corner detector. The operation is to find the local maxima of the principal curvatures of the image local autocorrelation. The Harris-Stephens corner detector is efficient to compute and has been shown to be one of the best corner detectors in terms of repeatability with scale, illumination and view point changes. It is effective to extract corner points in the texture of the IR outdoor images. Typically, this corner detector returns hundreds corner points in an outdoor image.

Points are 2-D features, which are invariant to rotation and scaling and useful for fitting image transformations. However, points themselves do not carry image structural information. The corner points detected by Harris-Stephens corners detector can be the maximum curvature point on object edges, it can also be corners from noisy and textures. To establish point match, each corner point must be characterized by its local support and the cross-correlation of the local supports can be used for establishing initial matches.

4.1 Robust matching

A robust technique for point matching in two images from uncalibrated cameras is proposed by Zhang[1]. In one implementation, the local supports of the corners are of 15 x 15 pixels size and the search areas are of a quarter of the image. The larger the supports and the search areas, the more expensive the computational cost for the cross correlation. Thus, the optical correlator could be used in this operation. A thresholding of the correlation scores provides a set of candidate matches. However, in the results from the cross correlation and thresholding, one point in the first image may have correlation scores above the threshold with several points in the second image, and vice versa. This is the ambiguities in the correlation-based matches. For this reason, the point matches determined by correlation and thresholding are considered only as candidate matches.

To remove the ambiguities and determine the best matches, and determine the best matches among the candidate matches . One must find the matches A and B in images 1 and 2 such that when computing the cross correlations between local support of A and all point supports in image 2, the point B in image 2 gives the highest correlation score, and when computing the cross correlations between B and all the point in image 1, the point A in image 1 gives the highest correlation score. To find the best matches, a relaxation process is applied. The match strength is defined as a measure which depends not only on the correlation score but also on the distances to other candidate matches within neighborhood of a match. The relaxation is an iterative procedure with the "some-winners-take-all" strategy to minimize an energy function which is a summation of strengths of all candidate matches.

After the relaxation, we obtain the initial matches whose number is typically less than one hundred, with still many false matches. The initial matches are then used to estimate the epipolar geometry described in Eq.(3). A robust technique, namely the Least Median of Squares, is used in the estimation to discard false matches. The recovered epipolar geometry is again used to verify the matches with a stereo matching process. The resultant matches are finally used to determine the unknown coefficients in the registration function.

The experiments on the robust numerical image match algorithms show that although the cross correlation is widely used for image matching. Single shot of correlation followed by thresholding is usually not able to determine the correct matches in the practical applications. Additional operations such as the bi-directional correlation, relaxation, robust epipolar geometry estimation and stereo matching must be used. If the optical correlator will be used to accelerate the computation of cross correlation, It would be integrated into the numerical processing system. Some numerical processing of optical correlation data would imply.

4.2 Greylevel Differential Invariants

In the ground image registration the most deformations are local translations due to the camera movement and panning, so that the cross correlation-based method is effective. However, for aerial images with important rotation, scaling and affine distortion, the cross-correlation can fail to provide correct matches, because it is not invariant to rotation and scale changes. Moreover, for the corner points located on the depth discontinuity boundaries, whose local supports may be sheared with the viewpoint changes, the cross-correlation fails to find correspondences. We then use the Greylevel Differential Invariants (GDI) instead of the cross correlation for determining point matches[6]. The corners are still detected using the Harris-Stephens detector. Then the corner points are described by the GDI numerical features that are computed on a local support centered at the points. The GDI are nonlinear combinations of the low order derivatives (up to third order) of the image grey scale level. Those combinations are built to be invariant to rotation, due to the rotational symmetry of the GDI. Numerical differentiation of the digitized image is unstable, therefore the operation includes Gaussian smoothing of the image. For IR noisy image a large Gaussian of $\sigma = 7$ was used.

The GDI's are also invariant to scaling. However, the local supports around the corner points are not segmented from the image. Its range is fixed in the algorithms and can not change with the unknown image scaling. To obtain the scale invariance we compute the GDI's for a set of images with the scales $\sigma_i = (6/5)^i \sigma_0$ where $i = -n, .., -1, 0, 1, \ldots n$ and σ_0 is the reference scale of the original image. When n = 4, the scale factor ranges from 0.48 to 2.07 and 9 GDI's are computed for a keypoint. The is the multiscale representation that describes a keypoint in the scale-space.

The initial matches are determined in the GDI feature space. The Mahalanobis distance is used to determine the nearest neighbour using the k-d tree representation. The matches are verified in scale-space at multiple scales. We applied the GDI approach to an IR aerial image sequence with important rotations and scale changes. The number of matches returned in this approach was about 10 - 20, much less than that returned in the Zhang's approach.

From the initial matches, we first remove gross false matches using a test on the gradient angle at every matched point. The remaining matches are then used to determine the unknown coefficients in the registration function Eq.(4). For the aerial images we use the orthographic projection as a close approximation to perspective projection. Since the classical lease mean square method is sensitive to outliers, an M-estimation which is robust to outliers was used. By using the multiscale GDI matching, the registration is invariant to rotation and scale, and is computationally effective. However, when affine distortions, namely different scaling in horizontal and vertical directions, occur in the aerial image sequence, GDI failed to provide correct initial matches.

5. Multisensor image registration

Challenge in the multiple sensor image registration is the feature inconsistency. especially, for the images from two well separated spectral bands (visible and 8–12 μm IR bands). The radiometric data from IR passive sensors consist of 1) energy emitted by thermal radiation from the object bodies; 2) atmospheric emission reflected from object surfaces. In general, the gray-scale level of IR images depend on differences in body temperature, emissivity and reflectivity of the objects in the scene. The IR images have high contrast for hot objects in the scene, which are in most cases moving targets and are therefore not landmarks useful for registration. Image registration should rely on the static objects on the background of the scene, where, unfortunately, the IR outdoor images have very low contrast because the background objects have uniform temperature in the thermal equilibrium state. The background in the outdoor IR image is usually of very low contrast and noisy, or simply a dark region. Feature extraction and image registration based on the background are difficult.

There exist significant gray-level disparities between the IR and visible image. The thermal emitters are not necessarily good visual reflectors. A surface of high visual reflectivity (white surface) in visible band usually has low emissivity, so that the bright objects in the visible image may be dark in the thermal scene and vice versa. The sky is usually the brightest region in the visible image. It is, however, a dark region in the IR image because of the low temperature and the lack of reflectance. This is the reversal of contrast polarity between the visible and IR images.

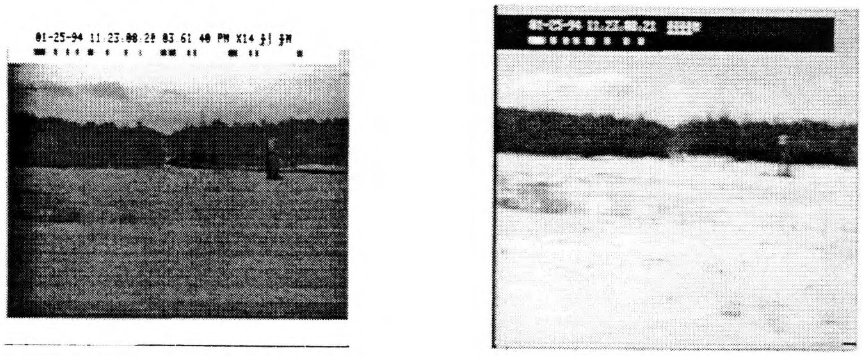

Fig.3 Contrast reversed IR image (left) and visible image (right)

The gray level disparity between the IR and visible images of real-world natural outdoor scene is much more complex than the simple contrast polarity reversal.

Figure 3 shows a contrast reversed IR image compared with the corresponding visible image of the same scene. The gray level distributions in most regions in the two images become similar, although there are still important gray-level disparities. Clouds are brighter than the sky in the visible image, owing to its higher reflectivity. They are also brighter than the sky in the IR image because of its higher reflectivity and emissivity. As a result, in the contrast polarity reversed IR images, the clouds are darker than the sky, whereas they are brighter than the sky in the visible image. Hence, a simple reversal of contrast polarity can not remove all the contrast reversal and gray level disparities. Also, shadows in the visible images are absent in the IR images.

Because of the gray scale level disparity and contrast polarity reversal the area-based block matching methods and the feature-based methods using the cross correlation or GDI feature matching of the gray scale levels in the local area supports, as described in Sections 4, failed to determine point matches. Several approaches have been proposed to bypass the feature inconsistency problem for multiple sensor image registration.

5.1 Laplacian pyramid

In some applications, one can transform two dissimilar multisensor images into similar. The intensities of the Laplacian pyramid images are insensitive to polarity reversals of contrast in the visible and IR images[7]. Figure 5 shows two step edges with opposite polarities of contrast. The edges are smoothed by the Gaussian filter. The Laplacian pyramid coefficient is the difference between the original edge and the smoothed edge. This is the Laplacian pyramid image at a resolution level determined by the size of the Gaussian smoothing function. When we take the absolute values of the Laplacian pyramid coefficients, the two Laplacian pyramid images become the same for the two contrast reversed step edges. Then, the area-based image registration can be applied to the Laplacian pyramid image intensities[7].

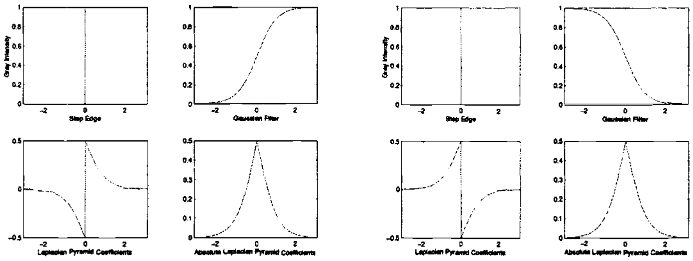

Fig.4 Two step edges with opposite polarities, smoothed by Gaussian filter, their Laplacian pyramid and absolute coefficients of the Laplacian pyramid.

The high frequency details in the image are lost in the Gaussian pyramid by the low-pass filtering and down sampling. One has to compute the difference between the averaged image and the original image in two successive pyramid levels. These difference images contain the detail information of the image. All the difference images form a new set of sequences that constitute another pyramid called the Laplacian pyramid. The original signal can be reconstructed exactly by summing the Laplacian pyramid. The Laplacian pyramid is a multiresolution representation and the Laplacian pyramid images represent detail information of the image. Because image gray-scale level disparities due to the sensor spectral responses are mostly low spatial frequencies, so that the Laplacian pyramid images contain also less gray level disparity. Hence the area-based image registration can be applied to images preprocessed by the Laplacian pyramid.

5.2 Phase matching

Images from multiple sensors have different radiometric intensity distributions due to the different spectral responses of the sensors. Those differences appear mostly as slow variations over wide regions in the image, such as sky, land and forest, which are usually represented with low spatial frequencies and are concentrated in a narrow low frequency band. In the Fourier transform-based registration[8], the displacement is found by cross-correlation between two images. The location of the cross-correlation peak mainly depends on the Fourier spectrum phase and is insensitive to Fourier spectrum energy. One can then whiten the Fourier spectrum and use the phase-only cross-correlation for the registration[4]. In this approach, the low and high frequencies contribute equally to the cross-correlation. Therefore, contribution of the high frequencies is greatly highlighted, compared with the conventional cross-correlation. The location of the cross-correlation peak would not change if the image intensity variations are limited to a narrow spatial frequency band. The Fourier phase correlation registration method is then relatively independent of the sensors.

5.3 Feature-based matching

Both the Laplacian pyramid and phase matching technique benefit from the use of high spatial frequencies of the image for bypass the feature inconsistency problem. The Laplacian pyramid represents detailed information, namely contours, in the image. It is well known from the optical phase-only filter experience that the whitening of the Fourier spectrum in the phase matching approach highlights the high spatial frequencies. However, the direct edge detector would be more powerful, precise and versatile than the Laplacian pyramid and whitening Fourier spectrum for the image feature extraction.

6. Multiscale edge detection

The feature-based multisensor image registration utilizes the fact that whatever the spectral responses of multiple sensors, two real world objects in the scene would always appear still differently. The boundaries of the objects may be then used as matching entities for multi-sensor image registration. In the 3-D real world scene, objects are separated from the background by depth discontinuities, which are usually manifest as intensity discontinuities in the 2-D images. Those edges and boundaries represent structures in the image, that are common for multiple image types and can be used for multiple sensor image registration.

Edges are defined as points where the modulus of gradient is a maximum in the gradient direction. Along an edge the image intensity can be singular in one direction while varying smoothly in the perpendicular direction. Edges can be created by occlusions, shadows, sharp changes of surface orientation, changes in reflectance properties, or illumination. In IR images of a 3-D scene, most edges represent occlusions and depth discontinuities between objects in the scene, which represent structural information in the image.

6. 1 IR image edge detection

A particular difficulty arises in the edge detection for IR/visible image registration. Image registration requires to extract common features which are static in the scene background. In most cases, the background objects in the IR images have the same thermal equilibrium temperature, so that the contrast in the IR image background is related to only the differences in the emissivities and reflectivities of the object surfaces and are therefore very low. Also, the IR images are typically noisy.

The first derivative of Gaussian is usually called Canny edge detector[9]. After the filtering, a non-maximum suppression process keeps the pixels where the values of the output are the local maximum in the direction of the gradient. The edge linking uses a hysteresis thresholding. We first determine edge pixels, which are above a high threshold. Then, among all other local maxima, which are above a low threshold, we keep only those pixels that are located in the neighborhood of the edge pixels. The parameters in the Canny edge detector are the width of first derivative of Gaussian filter σ and the low and high threshold values.

One problem of the Canny edge detector is its sensitivity to threshold values. The non-maximum suppression in the Canny detector is excessively reliant on the estimation of the gradient angle and so often fails to mark edge pixels at junctions, corners and even on some smooth curve portions where the contrast changes are too poorly defined. This is the reason for broken edges. When the response of an edge point is close to the detection threshold, a small change in edge strength or in the pixellation may cause a large change in edge topology, that makes the extracted edges suspicious, non-reliable, especially near the corners. The sensitivity to noise is another important problem in the edge detection. The noise in IR images occur as local fluctuations of the image brightness function, which have strong derivative magnitudes, but represent unnecessary image details unrelated to image structure. To extract structural edges from the noisy edge map we use the large Canny filter of $\sigma \geq$ 6-7, which corresponds to a filter size of 37 - 43 pixels. In this case the structural edge is a continuous curve, such as horizon in the ground images, so that the curve length thresholding can be applied to extract the horizon in the edge map. However, with a large σ , the extracted horizon line does not follow the real contour at high curvature. The larger the filter support σ, the less broken the edges are, and, however, the more image details are filtered out by the large size filter, resulting in a loss of edge localization. Therefore, the multi-scale edge detection is used to recover the localization in the coarse edges. We developed two multiresolution edge detection schemes: hierarchical edge detection and edge focusing.

6.2 Hierarchical Edge Detection

First, the horizon curve is detected at a coarse level with a large Canny edge detector which smoothes the images with a Gaussian of large support σ_0. The horizon is usually the longest curve in the image. For favoring continuity of the extracted curve, no thresholding on the gradient magnitude is applied, such that the horizon appears as a continuous curve or, at least, less broken. Then, the horizon is extracted from the noisy edge map by a curve length thresholding. In the cases where the horizon curves are still broken, we apply the edge saliency measure and combine both edge and region information in order to ensure the extraction of the horizon at the coarsest level, as explained in Section 6.

The coarse horizon is used to guide the search of edges at fine scale. We define a sub-image in the neighborhood of the coarse edge in the original image. The sub-image covers the region along the horizon with 40 pixels above and 10 pixels below each coarse horizon point. The choice of the sub-image size is according to the observation that the images of trees on the hill were cut by the smoothing at the coarse scale. To recover the top of trees we need a search in a large region above the horizon curve. We then apply the Canny edge detector with a small filter width σ within the sub-image. In the experiment, the fine Canny filter was with $\sigma = 0.7$ for visible and $\sigma = 1.5$ for IR images. The noise still exists after the Canny edge detection at the fine scale. However, this noise is within the sub-image zone and may be removed easily by a curve length thresholding, that results in a clearly defined horizon curve. A specific modification on the Canny edge detector was made to prevent the artificially defined sub-image boundaries from appearing as new edges.

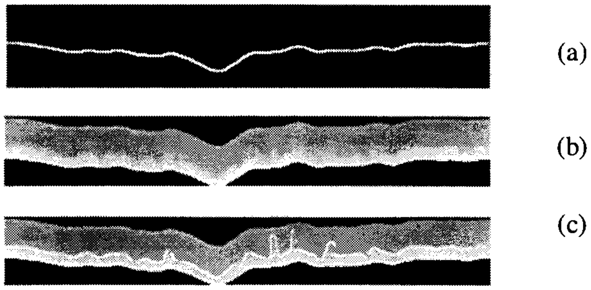

Fig.5 Results of Hierarchical Edge Detection

The coarse horizon extracted from an IR image is shown in Fig.5a, where $\sigma_0 = 7.0$, the minimum length threshold applied was 500 pixels. The sub-image is shown in Fig.5b. Figure 5c shows the fine edges obtained by applying a fine Canny edge detector with $\sigma = 1.5$.

The hierarchical edge detection is quit reliable and fast. Since at the fine scale the edge detection is guided by the coarse level edge, the search in large area is avoided, that reduces the computational cost. The shortcoming of the algorithm is the ad-hoc determination of sub-images.

6.3 Edge Focusing

Edge focusing is a coarse-to-fine edge tracking algorithm for recovering the edge points at the finest scale[10]. The scale-space tracking is implemented in a continuous manner. With continuous scaling, the edges are gradually focused by varying the resolution continuously, and moving in the scale space with sufficiently small steps, such that the edge element do not jump farther away than one pixel between successive steps. Our implementation of edge focusing is as following:

1. Detect edge using Canny Detector with the Gaussian smoothing σ_0 sufficiently large so that horizon curve is detected;
2. Extract the horizon using a threshold on the curve length; The horizon curve is denoted as $E(i, j, \sigma_0)$. If (i, j) is an edge point, then $E(i, j, \sigma) = 1$.
3. Detect edges $E(i, j, \sigma_k)$ in a window centered at each edge point $E(i, j, \sigma_{k-1})$, using the Canny edge detector of size $\sigma_k = \sigma_{k-1} - \Delta\sigma$ with $k = 1,2,3...$ and $\Delta\sigma = 0.5$. The window size is 7×7, when $\sigma_k > 2.0$, and is 5×5 when $1.0 \le \sigma_k \le 2.0$, and is3×3 when $\sigma_k < 1.0$.
4. Go on step 3) until a weak Gaussian smoothing of size σ_K.

In the successive Canny edge detection, after application of the first derivative of Gaussian filter the non-maximum suppression process is applied which keeps only the local maximum in the gradient direction. There is no threshold at finer resolution. The only threshold is on the curve length applied at the coarsest scale σ_0.

Bergholm[10] investigated the deformation of four elementary contour structures: step edge, corner, double edges and edge box. During the edge detection, those contours are generally deformed in four ways: rounding-off, expansion, transformation into circles, or merger, owing to the large Gaussian average operator which blurs the image. In each of the four cases, Bergholm showed that the displacement vector, describing the deformation of the edge contour, is normally of length within the range from 0 to $2|\Delta\sigma|$, where σ is the width of the Canny edge detector, $\Delta\sigma$ is the increment of size of the successive Canny filters. Therefore, if $|\Delta\sigma| = 0.5$, the displacement of the edge points would be normally less than one pixels, so that corners and junctions may be recovered with a precision less than one pixel.

Real world images contain mostly ramp edges instead of ideal step edges. It is easy to show that the Gaussian blurring operating on a ramp edge always yields smaller displacement than that yielded on a step edge as affirmed by Bergholm. A ramp edge may be modeled as a step edge smoothed by a Gaussian G whose size σ_1 depends on the imaging condition and on the camera. Let $r(x, y)$ denote the step edge and $f(x, y)$ the ramp edge in gray level image, then

$$f(x, y) = r(x, y) \otimes G(\sigma_1)$$

where $\otimes$ denotes the convolution. When we use Canny Edge Detector, the image is blurred again with a Gaussian smoothing whose size σ_2 depends on the scale of the edge detector. Let $g(x, y)$ denote the blurred ramp edge before computing the first derivative, then

$$g(x, y) = f(x, y) \otimes G(\sigma_2)$$

therefore

$$g(x,y) = r(x,y) \otimes G(\sigma_1) \otimes G(\sigma_2) = r(x,y) \otimes (G(\sigma_1) \otimes G(\sigma_2))$$

which is equal to

$$g(x,y) = r(x,y) \otimes G(\sqrt{\sigma_1^2 + \sigma_2^2})$$

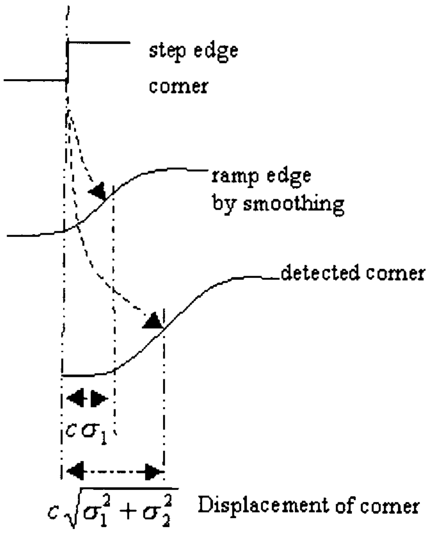

Fig.6 Rounding-off displacement for a ramp edge

Therefore, the length of rounding-off displacement ρ from the corner of ideal step edges to the detected corner is equal to $c\sqrt{\sigma_1^2 + \sigma_2^2}$, where c is a constant. However, the displacement from the center of the ramp corner to the detected corner would be proportional to $\sqrt{\sigma_1^2 + \sigma_2^2}$ - σ_1, as illustrated in Fig. 6 and would be less than σ_2. Therefore, if $|\Delta\sigma_2|$ = 0.5 in the edge focusing, the displacement of the ramp edge corner would be less than one pixels.

In our IR images the ramp edges of trees can be very slow of more than 20 pixels wide, corresponding to a large σ_1 more than 10. The edges around the trees were cut completely when a Canny edge detector of σ_2 = 7 was applied. This is because the large displacement of the corner $\sqrt{\sigma_1^2 + \sigma_2^2}$. However, using the edge focusing we were able to recover the edges and tops of the trees, which would be important for the image registration.

In our IR images the ramp edges of trees can be very slow of more than 20 pixels wide, corresponding to a large σ_1 more than 10. The edges around the trees were cut completely when a Canny edge detector of σ_2 = 7 was applied. This is because the large displacement of the corner $\sqrt{\sigma_1^2 + \sigma_2^2}$. However, using the edge focusing we were able to recover the edges and tops of the trees, which would be important for the image registration.

We implement the edge focusing algorithm with the filter size increment $\Delta\sigma = 0.5$ and varying size windows. We chose to use the window size larger than the usually used, 3 x 3, so that the gradient magnitude values can be evaluated at the two neighboring pixels, because in the non-maximum suppression the determination of an edge pixel requires to compare with at least two neighboring pixels. We believe that the length of rounding-off displacement ρ can be larger than one pixel, because the real ramp edges in our IR images were noisy and do not follow the theoretical model described in the precedent.

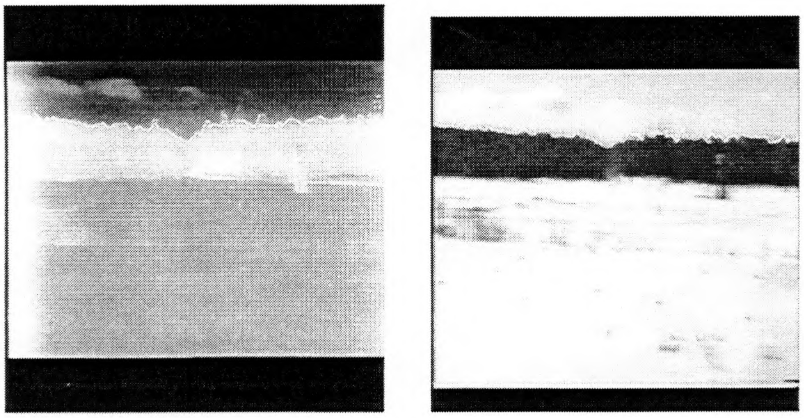

Fig.6. Experiment results of Edge Focusing. Right Visible image Left Infrared image.

For images shown in Fig.6, we first detected the coarse horizon with $\sigma_0 = 4.5$ for visible image and $\sigma_0 = 7.0$ for IR image using Canny Edge Detector. Then we applied the edge focusing with the scale step $\Delta\sigma = 0.5$ and the varying size windows. The final scale was $\sigma = 0.7$ for visible image and $\sigma = 1.5$ for IR image. Figure 6 shows the extracted edges which follow nicely the silhouette of the hill with some flat tops of trees recovered in both visible image and infrared image.

7. Hausdorff distance

Given a set of salient structural edges from each image, the next step is to determine the image transformation parameters that are useful for aligning those features. The search for the optimal image transformation can be implemented in several ways. Most feature matching methods determine the correspondences between the elements of the feature sets, and then determine the transformation parameters, as the two approaches described in Section 2. Once a set of correct matches is found the fitting to transformation is, in general, quick to compute. The drawback is the prohibitive cost of detecting and eliminating false matches. Its advantage is that once a set of correct matches is found the image transformation is, in general, quick to compute.

Transformation space methods is the direct search in the transformation space to match two given sets of features, where no explicit correspondences between features are given. The approach can be prohibitively expensive because the search space is generally very large. However, outliers are easily handled by using rank order statistics. A strategy for efficiently searching the parameter space is given by Huttenlocher *et al.*[11] In view of the large proportion of outliers in feature based multi-modal image alignment a transformation space method based on the directed Hausdorff distance was implemented. The size of the search space is reduced by partitioning the image into blocks and searching for translations that minimize the Hausdorff distance between corresponding blocks. The assumptions are that the motion can be locally approximated by simple translations of blocks, and the percentage of outliers and an error bound for the feature alignment are known approximately.

We are given by two binary feature sets consisting of points and edges. There are no explicit feature correspondences. The optimal match between the two sets may be found by cross-correlation. However, the points and edges have no geometric size in mathematical sense. This correlation would very sensitive to noise in the features sets, so that instead the distances between the two feature sets may be computed. The Hausdorff distance is defined by

$$A = \{a_1, \ldots, a_m\} \text{ and } B = \{b_1, \ldots, b_n\}$$

$$H(A,B) = \max(h(A,B), h(B,A))$$

$$h(A,B) = \max_{a \in A} \; \min_{b \in B} \|a-b\|$$

where A and B are point sets, H is the generalized Hausdorff distance and h is the directed Hausdorff distance. When the set B is aligned with the set A, the minimum distances between B and A result, and there is still one point in B whose distance to any point in A remains as the maximum among those minimum distances, that is defined as the directed Hausdorff distance from B to A. Inversely, when the set A is aligned with the set B, there is one point in A whose distance to any point in B (not necessarily to the point in B that gives the directed Hausdorff distance h from B to A) remains as the maximum among the minimum distances, that is defined as the directed Hausdorff distance from A to B. The generalized Hausdorff distance is the maximum among the directed and inverse Hausdorff distances.

To compute the Hausdorff distance between the image A and model B, we first build a distance transform map of A in which the value at every pixel represents the distance from that pixel to a nearest point in the set A. Then we superimpose the model point set B with its all possible distorted versions on the distance transform map of A. The minimum distances from every point in B to the point set A can be then easily sorted in a decreasing order.

In the presence of outliers the Hausdorff distance will return the greatest distance which is likely due to an outlier. To handle outliers, the partial directed Hausdorff distance is introduced as

$$h_k(A,B) = \underset{\mathrm{a} \in \mathrm{A}}{k\text{th}} \; \min_{b \in B} \|a-b\|.$$

which allows to accept *(100-k)%* of outliers and evaluates the *k%* ranked distances for determining the Hausdorff distance.

The Hausdorff distance can be prohibitively expensive because the search space is generally very large if all the possible distortions of the model must be considered. We reduce the size of the search space by partitioning the image into blocks and searching for translations that minimize the Hausdorff distance between corresponding blocks. The idea is similar to that of the block matching described in Section 3 for estimation of displacements of the block centers based on the image areas. The assumptions are that the motion can be locally approximated by simple translations of blocks, and the percentage of outliers and an error bound for the feature alignment are known approximately.

The alignment method using Hausdorff distances proceeds as follows for a pair of images after extraction of the salient edges. We found the horizons extracted by multiscale edge detection do not suffice to align two images, because the horizon features are not evenly distributed over the image. Thus, extracting structural features in the other parts of image is necessary. Figure 7 shows the edge features used for the Hausdorff distance matching

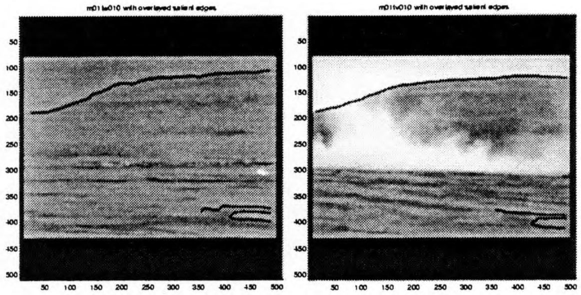

Fig.7 Edges extracted in IR and Visible battle field images.

The process steps are as following

1) Compute a quadtree partition of each edge image such that no block without edge points is further subdivided. The partition with fewer blocks is retained for both images. Define a set of model edge points for the first block in image 1 from the edge points that lie within that block. Create a model image from these edge points.
2) Define a set of subimage edge points from the corresponding block in image 2 from the edge points within the block extended by a border whose dimensions correspond to the largest expected vertical and horizontal displacements. Create a target image from these edge points.
3) Compute the directed partial Hausdorff distance under a translation transformation from the model image to the target image. The translation which minimizes the *k*th ranked distance is retained.
4) Repeat steps 2 and 3 for the remaining non-empty blocks. If at least 3 blocks provide local translation estimates from step 3 then the global affine transformation is estimated, the nonreference image is resampled according the global affine transformation and the images are fused. The image fusion is accomplished by an appropriately weighted combination of the aligned images brightness values.

Figure 7 shows a scene taken simultaneously by a daylight and IR camera at Defense Research Establishment Valcartier. The viewpoints of the two cameras are displaced slightly and there is a slight relative rotation about the optical axis which would yield a very poor fused image if no alignment is made. The quadtree decomposition stops at the first level, i.e., there are 4 blocks. The salient edges include the silhouette of the hill and some ground structure, which are overlayed on the images.

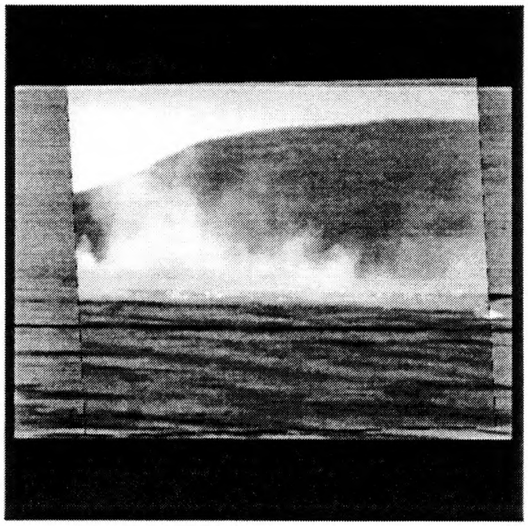

Fig.8 . Aligned and fused IR and visible images. Fusion is by weighted combination of image brightness values after alignment

Finally, Figure 8 shows the fused aligned images. The salient edges are registered in each of 3 blocks, the fourth block contains no edge points. The Hausdorff distance is used to find the optimal displacement assuming 5 percent outliers for the blocks covering the hill edge and 10 percent outliers for the edge in the lower right block. The specified search strategy finds the translation for each block such that 90 percent of the visible image edge points are no more than 5 pixels from some IR image edge point for the corresponding block. The local displacements are then used to determine the global affine transformation to register the two images. The estimated (x,y) displacements for the blocks upper left, upper right and lower right that are supplied to the global affine estimator for aligning the visible image to the IR image are (33,-3), (-11,-7) and (-8,-11) respectively.

The estimated affine transformation parameters that map point p in the visible image to the point p' in the IR image such that $p' = \mathbf{M}p+\mathbf{t}$ are

$\mathbf{M}$ = [0.8239 0.0544] and

$\mathbf{t}$ = (17.1925, -6.2471)^T.
[-0.0179 0.9897]

Note that the image coordinate system origin is top left with positive x to the right and positive y down.

Table 1 Algorithms for image matching

	Robust Epipolar Estimation	**Differential Invariants**	**Block Matching**	**Hausdorff Distance**
Features	Harris-Stephens Corners Intensity of local supports		Sub-image intensity	Binary edges, points
Similarity metric	**Correlation** scores	Differential Invariants	Correlation scores	Hausdorff Distance
Search Strategy	Compute transform parameters By optimization		Compute transform parameters by optimization	
Methods	M-estimation Least Median Square		LMS, Local Displacement Vectors	Cell decomposition, Early rejection ...
Algorithm	Initial matches → Epipolar Estimation →Stereo verification→New Estimate		Direct correlation	Distance Trans. + sorting
Transformations	Translations, Perspective Projection	Rotation, Scaling Orthographic Projection	Translations Approx. Affine	Affine
Results	Epipolar Geometry	Registration function	Sub-pixel matching	Ranked distances
Robustness	Robust estimation with initial matches		Outlier rejection	Partial Distance
Multi-sensor registration	Corner detection & local supports are sensitive to gray level disparity		Lapacian pyramid energy etc.	Structural salient edges
Cost	Fast	more expensive	Real-time	Expensive

8. Conclusion

We have analyzed the problematic in the real world visible/IR image registration with the image sequences from well separated spectral bands. After a preprocessing with the Laplacian pyramid the area-based approaches may be still applied. However, multiscale extraction of structural edges followed by feature matching using the Hausdorff distance measures are more powerful to process very poor quality IR images.

Because the common features extracted from images of two modalities can be still different in detail, the transformation space match methods with the Hausdorff distance measures were used which are more suitable than the direct feature matching methods for dealing with outliers in the extracted fature sets. We have introduced image quadtree partition technique to the Hausdorff distance matching, that dramatically reduces the size of the search space into that of the search for translations which minimize the Hausdorff distance between corresponding blocks.

We have shown image registration of visible/IR real world images of battle fields. The key point is to extract salient features from the real world images using local, regional and global information. Mulitsensor image registration and fusion is one realization of advanced computer vision systems. Multiple sensors, multiple spectra and color cameras, 3-D perspective projection image formation and time video image sequences are widely used in the advanced computer vision systems.

Cross correlation is one of the basic operations used in image registration algorithms for determining candidate point matches. The invariance to image rotation, scale and view angle changes is an important issue for feature matching. Table 1 gives a list and comparison of four algorithms described and implemented in this paper. High speed optical correlator could be useful as an efficient hardware for implementation of image registration algorithms, and the optimal correlation filter designs could enhance the performance of the image processing systems. However, most optical processing hardware have the lack of flexibility. The optical correlator should be integrated into the numerical systems, such that powerful robust algorithms can be used to processing the correlation output data and to remove the ambiguity of the correlation and threshoulding outputs.

References

[1] Z. Zhang,, R. Deriche, O. Faugeras, L. Quang-Tuan, "A robust technique for matching two uncalibrated images through the recovery of the unknown epipolar geometry", Artificial Intelligence, vol.78, no.1-2, 87-119, (1995).

[2] M. D. Pritt., "Image registration with use of the epipolar constraint for parallel projections", J. Opt. Soc. Am. A. Vol. 19, No. 10, 2187-2192 (1993).

[3] L. S. Shapiro, A. Zisserman., M. Brady, "3D motion recovery vis affine geometry", Int. J. of Comp. vision, 16, 147-182 (1995).

[4] Brown L. G., "A survey of image registration techniques". ACM comput. surveys, vol.24, 325-376 (1992).

[5] . D. McReynolds, P. Marchand, Y. Sheng, L. Gagnon, L. Sévigny, "Stabilization of Infra-red Aerial Image Sequences Using Robust Estimation," Vision Interface 99, Trois-Rivieres, Qc, 288-295, (1999).

[6] Schmid, C., Mohr, R., "Matching by local invariants", Rapport de Recherche, N 2644, INRIA, (1995).

[7] Sharma R. K. and Pavel M., "Registration of video sequences from multiple sensors", Proc. Image Registration Workshop, NASA Goddard Cenetr, 361-364 (1997)

[8] Brown L. G., "A survey of image registration techniques". ACM comput. surveys, vol.24, 325-376 (1992).

[9] Canny J., "A computational approach to edge detection", Trans. IEEE PAMI-8, No.6, 679-698 (1986).

[10] Bergholm F., "Edge focusing", IEEE Trans. PAMI-9, No.6, 726-741 (1987).
[11] Huttenlocher, D.P., Klanderman, G.A., Rucklidge, W.J., "Comparing images using the Hausdorff distance," IEEE Trans. PAMI-15, No.9, 850-863 (1993).