A comparison of image features for registering LWIR and visual images

Jaco Cronje
Council for Scientific and Industrial Research, Pretoria, South Africa
Email: jcronje@csir.co.za

Jason de Villiers
Council for Scientific and Industrial Research, Pretoria, South Africa

***Abstract*—This paper presents a comparison of several established and recent image feature-descriptors to register long wave infra-red images in the 8–14 μm band to visual band images. The feature descriptors were chosen to include robust algorithms, SURF and SIFT — and fast algorithms, BRISK and BFROST. To evaluate the feature-descriptors a ground truth was created by determining the intrinsic and extrinsic camera calibration parameters for the cameras and using this to photogrammetrically relate pixel positions between the images. The inlier results of each feature descriptor for the top 20%, 50% and 100% of the matches (based on match strength) were used to create a homography. The average pixel error between the homography reprojected feature points and the photogrammetric reprojection was used as the error. The results show that none of the descriptors perform well in standard form, with BFROST faring slightly better than the other algorithms. This suggests a need to modify the algorithms to detect physical/structural features and de-emphasise textural features.**

I. Introduction

A. Relevance of cross spectral registration

Long Wave Infra Red (LWIR) imagery in the 8–14 μm wavelength band, also known as thermal imagery, has several advantages over visual band imagery [1]. Among these are decreased sensitivity to atmospheric aerosols and scintillation, superior performance in low (visual) light conditions and easy detection of many objects of interest such as vehicles with an internal combustion engine. This is due to the majority of light in this spectrum being emitted by the objects being surveyed rather than being reflected light.

There are several disadvantages to LWIR imagery too. Of particular interest is that intensity of objects in LWIR imagery is solely due to their surface temperature and emissivity, this implies that distinguishing marks such as colour, insignia and serial/licence numbers are generally not visible. In addition, current LWIR cameras typically have significantly lower resolution than visual cameras (e.g. see Sections III-A and III-B) yet cost significantly more. To illustrate these phenomena Figure 1 shows LWIR photos of the authors, it is much more difficult to distinguish between them.

Registering the images of the two bands, that is determining the pixel correspondence between a LWIR and visual image, would allow both the easy determination of objects of interest (using the LWIR band) and their identification (in the visual band). Other benefits may be found such as the haze mitigation of visual images via incorporating a Near Infra-Red (NIR) channel [2].

B. Related Work

Many examples of image feature detector/descriptors have been developed for matching features between visual images. The Geographical Information Systems (GIS) field yields some papers on cross-spectral feature detection. Firmenich *et al.* [3] describe how the Scale-Invariant Feature Transform (SIFT) [4] was modified to perform better in matching between the visual and NIR channels by making it insensitive to reversal in the image gradient. Hasan *et al.* [5] also improved upon SIFT for visual-NIR matching by constraining the portion in the second image on which a match for a feature in the first image is searched. This was done by using two strong matches — which include both spatial and orientation information — to predict where each other feature will be and their scale. Teke and Temezel [6] applied this scale restriction method to the Speeded Up Robust Features (SURF) [7] algorithm. Their results show a worst case matching between the NIR and Blue channels, with results of between 77% and 85% depending on the implementation of SURF and whether or not the scale restriction is applied. Equivalent results for red channels are 86% through 91%.

Brumby *et al.* [8] investigate the supervised evolution of feature extraction kernels by combining primitive image processing operations in order to extract the desired features (such as roads, crop types and rivers) from pre-registered hyperspectral images extending from the visual to short wave infra red (SWIR).

This work is different from that described above in that LWIR is used instead of NIR, a difference of over tenfold in wavelength. This results in a further decrease in feature mapping performance due to the greater dissimilarity between the bands.

C. Axis and notation definition

The mathematical notation used in this paper is as follows: A 3D vector, V_{bac}, is a vector from point a directed towards point b expressed in terms of its projections on orthogonal coordinate system c's axes. V_{bac} is used when the magnitude of the vector is unknown or unimportant. T_{bac} represents the translation or displacement of point b relative to point a.

(a) Author 1 (b) Author 2

Fig. 1. LWIR images of the authors

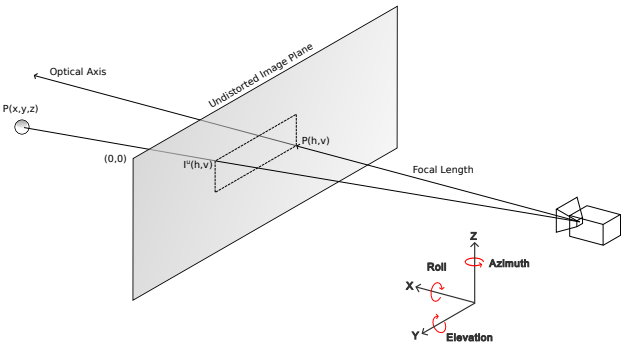


Fig. 2. Axis definition.

R_{ab} is a 3-by-3 Euler rotation matrix expressing the rotation of an orthogonal axis system a relative to (and in terms of its projections on) an orthogonal axis system b. Individual elements of 3 dimensional vectors are referred to as x, y or z whereas 2 dimensional (2D) vector's elements are referred to as horizontal (h) and vertical (v) to avoid confusion. Figure 2 defines the axis system used and the directions of positive rotation.

D. Paper organisation

The rest of this paper is organised as follows: Section II describes the basic workings of the feature detectors. Section III describes the equipment used in this comparison. Section IV details procedure used to objectively compare the different feature metrics. Section V provides the results of the comparison. Section VI summarises the results and places them in context.

II. Feature Descriptor

This section describes the feature detectors used in this comparison. Two floating point and two binary feature point descriptors were evaluated.

A. Scale-Invariant Feature Transform

The SIFT [4] detector searches for stable features across multiple scales by searching for local extrema features over a set of Difference-of-Gaussian (DoG) images. An orientation histogram is constructed by sampling gradient orientations around the feature. The highest peak in the histogram is used as the feature orientation.

The region around the feature is divided into 4 by 4 sample areas. An orientation histogram is calculated for each of the sampling areas. A Gaussian weighting is then applied to the magnitudes before they are accumulated into the histogram. The values of all the histograms are placed into the feature vector. The normalised feature vector forms the 128 floating point value feature descriptor. SIFT is robust to almost all common image transformations.

The match strength between two SIFT features is defined as the L2-Norm: i.e. the length of the difference between the two feature vectors. Smaller values are better.

B. Speeded Up Robust Features

SURF [7] was inspired by SIFT [4], with the main goal to improve the execution speed of the detector and descriptor. SURF depends mainly on an integral image to approximate and speed-up the execution time.

The detector relies on the determinant of the Hessian matrix. The Hessian matrix is approximated by sampling rectangular regions that approximate the Gaussian derivatives. The local extrema from the approximate determinant of the Hessian matrix is located across different scales. Haar wavelets are used to calculate the orientation of sampling points around the feature. The feature orientation is detected by examining the magnitude of the orientations within a sliding arc window. The arc direction with the highest resulting magnitude is chosen as the dominant orientation.

The region surrounding the feature is divided into 4 by 4 sub-regions. Haar wavelet responses for each sub-region are accumulated to form the 64 element floating point feature vector.

The match strength between two SURF features is also defined as the L2-Norm: the length of the difference between the two feature vectors.

C. Binary Robust Invariant Scalable Keypoints

Binary Robust Invariant Scalable Keypoints [9] (BRISK) is a binary feature extractor, the feature detection part uses the

improved version of the Features from Accelerated Segment Test [10] (FAST) detector, namely Adaptive and Generic Accelerated Segment Tests [11] (AGAST) to detect key-points. The feature detection phase tries to detect features by searching in different scale-spaces. Local image gradients are calculated between sampling point pairs surrounding the feature. The sum of all gradients is used as the feature rotation.

The binary descriptor is built by comparing pairwise, smoothed pixel intensities from sampling points surrounding the feature. Each bit is set when the first pixel intensity is greater than the second pixel intensity. The resulting bits are concatenated to form the 512 bit descriptor.

The match strength between two binary features is defined by the number of elements that differ between the two binary vectors, i.e. the Hamming distance. Smaller values are better.

D. Binary Features from Robust Orientation Segment Tests

Binary features from robust orientation segment tests [12] (BFROST) is a fast feature extractor designed for the Graphics Processing Unit (GPU). BFROST uses the same continuous pixel-set criteria as the FAST detector to detect features with an additional 16 possible feature rotation estimations based on the median of the continuous pixel-set segment.

The feature descriptor describes an area around a detected feature point with a 256 bit binary vector. The descriptor is built by comparing the average pixel intensities of regions surrounding the feature. An integral image is used to speed-up the intensity calculations performed on the sampling pattern.

BFROST is scalable, rotation and translation invariant and robust to noise. The match strength between two features is also defined as the Hamming distance.

III. Equipment

One visual and one LWIR camera, as described below, were rigidly mounted relative to each other. Their intrinsic and extrinsic parameters were then determined (see Section IV-A) to allow for photogrammetric registration.

A. Visual Cameras

Prosilica GT1920 cameras, which have a 3MP resolution of 1936×1456, were used in this work. Pentax lenses with 8mm focal length were used, and provided a field of view (FOV) of $\pm 50^\circ$ horizontally by $\pm 40^\circ$ vertically.

B. Long Wave Infra Red Cameras

Xenics Gobi 640GigE microbolometers were used in this comparison. The cameras have a large 10.88mm by 8.17mm Charge Coupled Device (CCD) offering a resolution of 640×480 pixels. Combined with a 10mm lens, this provided an FOV of $\pm 60^\circ$ horizontally by $\pm 48^\circ$ vertically.

IV. Experimentation Methodology

A. Generating the ground truth

In order to quantifiably compare the different feature descriptors, a ground truth registration was sought. This was obtained by photogrammetrically calibrating the cameras. The lens distortion and inverse distortion was determined as described de Villiers *et al.* [13] using five radial, three tangential parameters and the optimal distortion center. The focal length and the extrinsic parameters of the camera were then determined as per de Villiers [14].

Once these parameters are known, the position that a pixel from Camera B should be placed in Camera A's image is determined by first calculating the the point where the distortion-corrected vector associated with each pixel of Camera B meets the stitching surface (assumed here to be a sphere [14]). This point is then back projected through to Camera A's image plane, where it was redistorted and scaled to determine the pixel position.

In order to calculate the point on the stitching sphere associated with each pixel, one first recalls the cosine rule:

$$a^2 = b^2 + c^2 - 2bc\cos\theta_{bc} \tag{1}$$

where:

$a, b, c =$ the lengths of the side of a triangle, and

$\theta_{bc} =$ the angle between sides b and c.

Now for a pixel i of Camera B, assign the corners of a triangle to be the known center of the sphere in some reference system (i.e. T_{SRR}), the position of camera B expressed in the same reference system (i.e. T_{C_BRR}) and the point where the pixel's vector intersects the sphere. This then infers that side a is equal to the stitch radius (R), and that side b is the distance between the camera and sphere center, or $\|T_{SC_BR}\|$ where $T_{SC_BR} = T_{SRR} - T_{C_BRR}$. All that is required is to determine the vector associated to each pixel and the cosine between it and T_{SC_BR}.

First one creates a vector in Camera B's axis using the focal length and intrinsic distortion parameters:

$$\begin{aligned}
I_i^u &= f_B^{undistort}(I_i^d), \\
V_{P_iBB} &= \begin{bmatrix} FLen_B \\ (P_h^B - I_{i_h}^u)pix_w_B \\ (P_v^B - I_{i_v}^u)pix_h_B \end{bmatrix}, \\
U_{P_iBB} &= \frac{V_{P_iBB}}{\|V_{P_iBB}\|}, \\
U_{P_iBR} &= R_{BR}U_{P_iBB}
\end{aligned} \tag{2}$$

where:

$I_i^d =$ the image coordinate of pixel i,

$f_B^{undistort} =$ the predetermined lens undistortion characterization function [14] for camera B,

$(P_h^B, P_v^B) =$ the principal point of camera B,

$(I_{i_h}^u, I_{i_v}^u) =$ the undistorted pixel position of pixel i,

pix_w_B = the width of the pixels on camera B's CCD,
pix_h_B = the height of the pixels on camera B's CCD,
R_{BR} = rotation of camera B relative to the ref. axis (known from the extrinsic parameters), and
U_{P_iBR} = desired pixel unit vector in reference axis.

Now, recalling that the dot product of two vectors is equal to the product of their magnitudes multiplied by the cosine of the angle between them, Eq. 1 can be rewritten as:

$$R^2 = \|T_{SC_BR}\|^2 + c^2 - 2c \times T_{SC_BR} \bullet U_{P_iBR} \quad (3)$$

which can be rewritten as:

$$0 = c^2 + c(-2 \times T_{SC_BR} \bullet U_{P_iBR}) + \|T_{SC_BR}\|^2 - R^2 \quad (4)$$

This is a quadratic in standard form, and if the camera is inside the stitch sphere will yield a positive and a negative real solution. The positive solution is the desired answer, which yields the point on the stitch radius as

$$T_{iRR} = T_{C_BRR} + c \times U_{P_iBR} \quad (5)$$

Once this point is known it is projected onto camera A's image plane, scaled to the pixel domain and then converted from the undistorted to distorted pixel domains to determine the corresponding pixel from Camera A. This process is exactly the same as that described in Sections III-B through III-D of de Villiers [14].

B. Creating the homography

OpenCV [15] was used to perform the homography calculation using the specified top percentage of the matches. The Random Sample Consensus option was selected to reject outlier matches. The percentage of inlier matches was recorded and used as further indication of the robustness of the homography determined with that particular feature descriptor and match strength.

C. Comparison metric

The metric used is the average error of the inlier features used to create the homography as described in Section IV-B. The error is the distance in pixels between the features in camera B reprojected onto camera A as determined by the homography of Section IV-B and photogrammetric calibration of Section IV-A. This is expressed mathematically as:

$$Error = \frac{1}{N} \sum_{j=0}^{j<N} \left(\|P_j^H - P_j^P\| \right) \quad (6)$$

where:

N = the number of inlier features used,
P_j^H = homography based pixel position of feature j, and
P_j^P = photogrammetrically based coordinate of feature j.

D. Image Scenes

Figure 3 shows the first scene used for this evaluation, it is an urban outdoor scene containing man-made structures with strong edges and texture. Figure 4 shows the outdoor scene used which contains natural vegetation. Both scenes appear, subjectively, to contain rich texture in the visual band.

V. Results

A. Intra-band registration

Table I provides the results of registering between visual images, the values are the number of inliers that agree with the best fit homography. Each scene is registered three times using only the top 20%, 50% or 100% of the matches respectively. The inlier percentage is the percentage of these top matches that were used. Table II provides the same results for registering the LWIR images.

The high percentage of agreement gives confidence on the correctness of the implementations of the four feature-detector algorithms. This is further supported by Figures 5 and 6, which show features correctly being matched within each band. Inlier matches are shown with a green line, while outliers are shown by the blue lines.

The BRISK algorithm performs poorly when 50% or 100% of the matches are used as many of the matches are weak and erroneous. It performs comparably to SIFT and BFROST when only the top matches are used. BFROST performs poorly on the LWIR Urban scene, but is comparable to SIFT in terms of performance when only the top 20% of the matches are used. SURF is consistently worse than SIFT and only marginally better than BRISK.

TABLE I
VISUAL TO VISUAL REGISTRATION INLIER PERCENTAGES

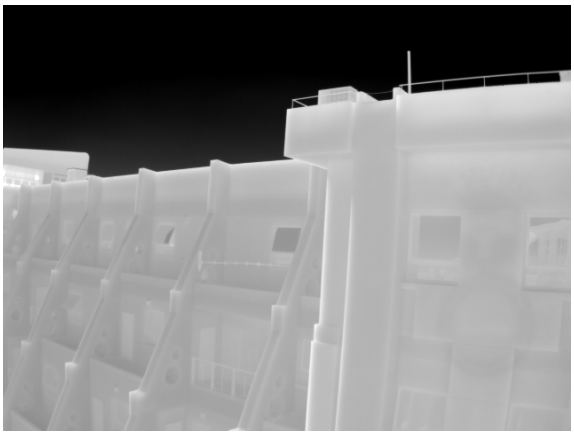
Feature Descriptor	Scene 1			Scene 2		
	20%	50%	100%	20%	50%	100%
SIFT	95.00	99.34	88.15	96.01	99.26	78.86
SURF	65.05	73.60	73.84	94.44	91.26	75.50
BRISK	96.89	85.33	51.75	91.61	71.59	44.12
BFROST	94.11	93.02	86.58	98.92	86.69	65.23

TABLE II
LWIR TO LWIR REGISTRATION INLIER PERCENTAGES

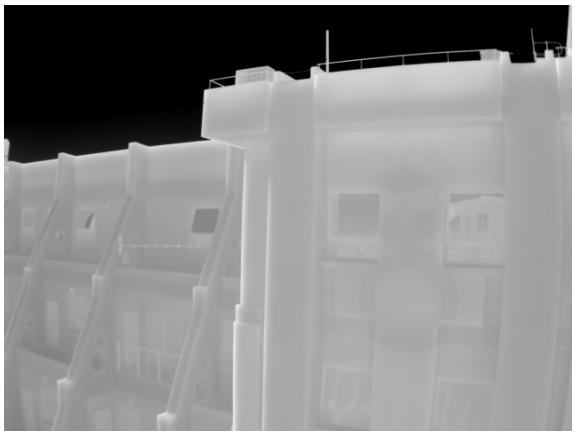
Feature Descriptor	Scene 1			Scene 2		
	20%	50%	100%	20%	50%	100%
SIFT	100.00	100.00	81.48	92.30	94.11	66.17
SURF	75.00	68.47	66.30	87.17	85.71	75.00
BRISK	91.48	81.19	64.25	100.00	69.29	50.78
BFROST	71.42	77.35	81.13	100.00	80.00	70.37

B. Inter-band registration

Table IV provides the results of registering the LWIR images onto the visual images, the values are as per Eq. 6. Table III provides the percentage of inlier features from generating the best fit homography. Figure 7 helps put these numbers

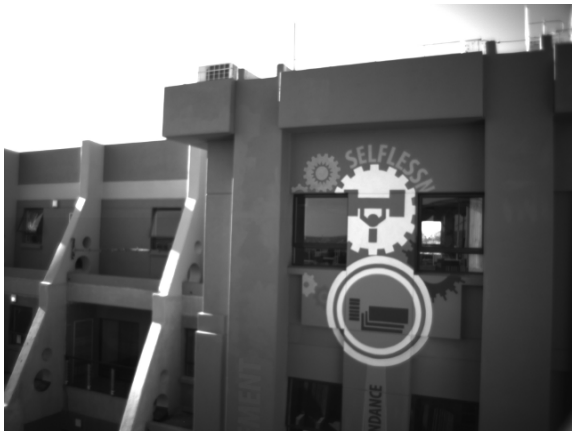


(a) LWIR image 1

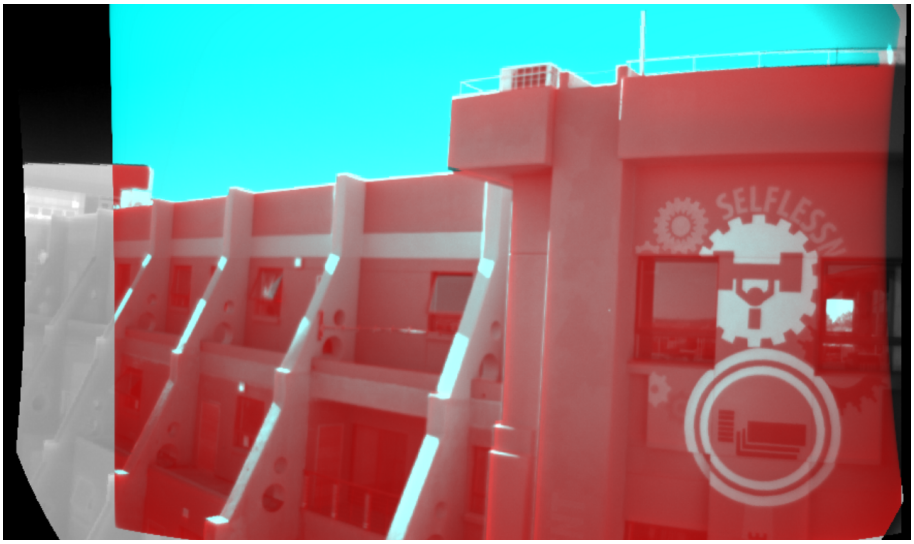


(b) LWIR image 2

(c) Visual image 1

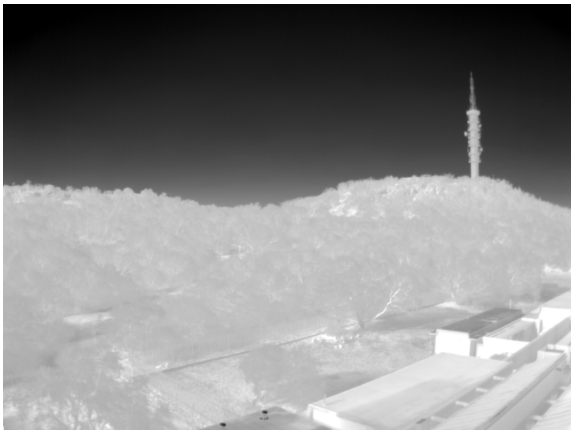


(d) Visual image 2

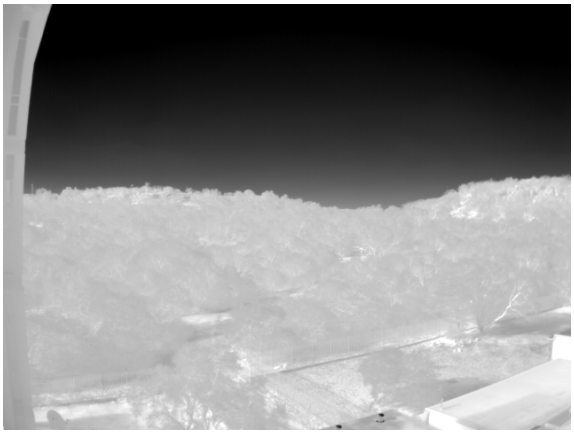


(e) Photogrammetrically stitched image

Fig. 3. Scene 1, Urban landscape



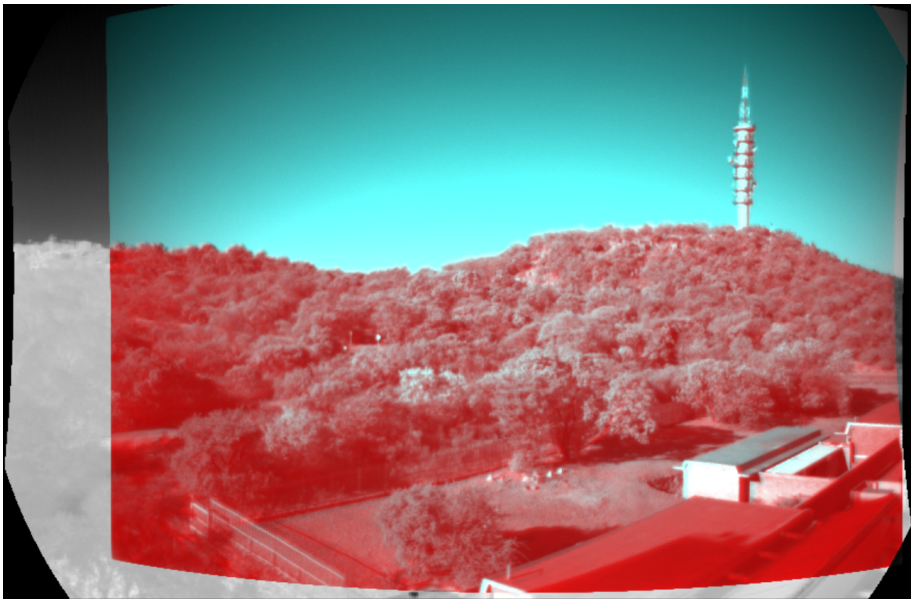
(a) LWIR image 1



(b) LWIR image 2

(c) Visual image 1

(d) Visual image 2



(e) Photogrammetrically stitched image

Fig. 4. Scene 2, Natural landscape

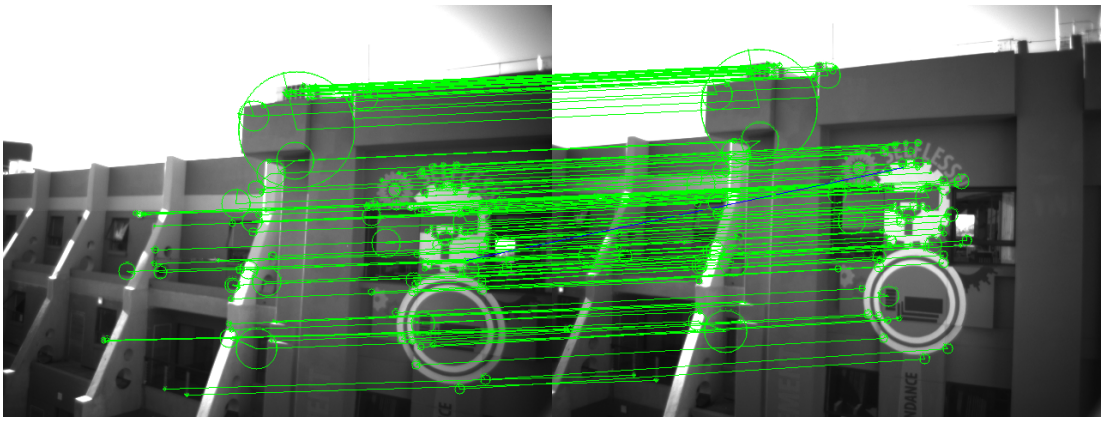


Fig. 5. SIFT feature matches between visual and visual of Scene 1.

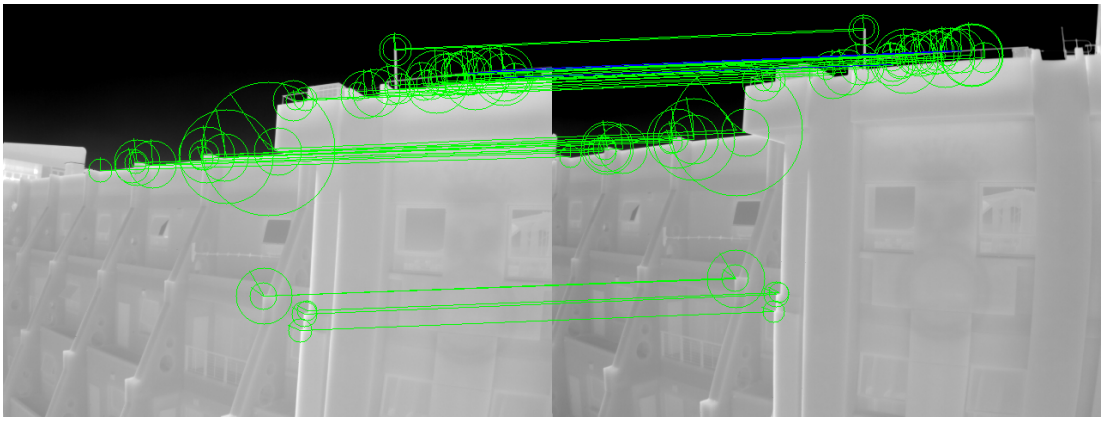


Fig. 6. BRISK feature matches between LWIR and LWIR of Scene 1.

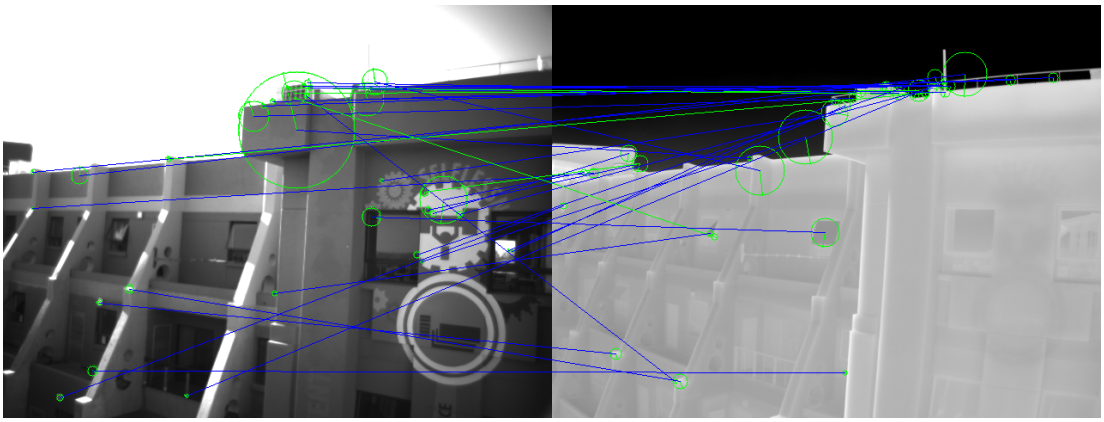


Fig. 7. SIFT feature matches between visual and LWIR of Scene 1.

in context by displaying the features matched between the thermal and visual bands.

Inspection of the values clearly shows that none of the descriptors were able to successfully register the LWIR and visual images. All the algorithms obtained errors of several hundred pixels (expressed in the 3MP AVT Camera's image space) in Table IV, this is further shown by the extremely low agreement percentages in Table III. Often it was not possible for OpenCV to find more than 6 points (the minimum is 4) that agreed to create a consensus homography.

SIFT and BFROST had the greatest number of inliers in the urban and natural scenes respectively, and the second greatest agreement in the other scene. However BFROST had, in almost all the tests, a noticeably lower error than all the other feature descriptors.

A final verification of the correctness of the photogrammetric procedures (in addition to generating Figures 3(e) and 4(e)) was performed. Ten points were crudely selected in each band in each scene, and their equivalent error was calculated. These results are given in the final row of Table IV and confirm the correctness of the photogrammetric procedures. These errors being in the order of 10 pixels, are due to the non precise manual feature selection (which is magnified by the difference in resolutions) and the poor image quality of the Pentax lenses, whose soft focus in the peripheries of the FOV adversely affected the calibrations.

LWIR–visual registration based on canonical features does not perform well. This is due to different keypoints being identified in each band which is compounded by the descriptions of correctly identified matching keypoints frequently being different too.

Further work on feature based matching ma focus on contour alignment and modification of feature descriptors to better cater for cross band matching.

VI. Conclusions

This paper tested four popular feature descriptors for the purpose of registering LWIR and visual imagery. The feature descriptors were used in unmodified canonical form6 to facilitate the selection of which descriptor should modified for LWIR-visual registration. In addition to the standard calculation of number of inlier matches, a quantified error based on comparison to photogrammetric calibration and stitching was performed.

It was found that none of the algorithms were able to register across the two bands, although all the algorithms registered well within either of the bands. This finding is consistent with Firmenich *et al.* [3] who speculated that a new feature extractor may need to be developed for LWIR imagery registration.

SIFT and BFROST significantly outperformed SURF and BRISK for inter band registration. BFROST was significantly quicker than SIFT, and so it is recommended for future modification for LWIR-visual registration.

VII. Acknowledgements

This work was supported by the Armaments Corporation of South Africa.

TABLE III
LWIR to visual registration inlier percentages

Feature Descriptor	Scene 1			Scene 2		
	20%	50%	100%	20%	50%	100%
SIFT	57.14	27.77	16.21	33.33	15.21	7.60
SURF	23.07	12.30	5.38	12.00	4.80	1.99
BRISK	28.57	17.75	9.81	25.80	23.22	12.25
BFROST	44.44	22.72	13.33	37.50	17.07	8.43

TABLE IV
LWIR to visual registration errors

Feature Descriptor	Scene 1			Scene 2		
	20%	50%	100%	20%	50%	100%
SIFT	926.6	488.6	638.2	604.8	812.2	611.4
SURF	743.1	599.1	798.7	471.8	810.1	574.2
BRISK	888.8	765.3	781.9	531.3	763.4	741.0
BFROST	684.6	438.0	620.7	929.2	371.3	441.0
Manual	11.0			11.0		

References

[1] L. Biberman, *Electro-Optical Imaging: System Performance and Modeling*. SPIE Press, 2000.

[2] L. Schaul, C. Fredembach, and S. Susstrunk, "Color image dehazing using the near-infrared," in *Image Processing (ICIP), 2009 16th IEEE International Conference on*, nov. 2009, pp. 1629 –1632.

[3] D. Firmenich, M. Brown, and S. Ssstrunk, "Multispectral interest points for rgb-nir image registration." in *ICIP*, 2011, pp. 181–184.

[4] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.

[5] M. Hasan, X. Jia, A. Robles-Kelly, J. Zhou, and M. Pickering, "Multispectral remote sensing image registration via spatial relationship analysis on sift keypoints," in *Geoscience and Remote Sensing Symposium (IGARSS), 2010 IEEE International*, july 2010, pp. 1011 –1014.

[6] M. Teke and A. Temizel, "Multi-spectral satellite image registration using scale-restricted surf," in *Pattern Recognition (ICPR), 2010 20th International Conference on*, 2010.

[7] H. Bay, T. Tuytelaars, and L. V. Gool, "SURF: Speeded up robust features," in *In ECCV*, 2006, pp. 404–417.

[8] S. Brumby, P. Pope, A. Galbraith, and J. Szyinanski, "Evolving feature extraction algorithms for hyperspectral and fused imagery," in *Information Fusion, 2002. Proceedings of the Fifth International Conference on*, vol. 2, 2002, pp. 986–993.

[9] S. Leutenegger, M. Chli, and R. Y. Siegwart, "BRISK: Binary robust invariant scalable keypoints," in *International Conference on Computer Vision 2011 - ICCV2011*, 2011.

[10] E. Rosten, R. Porter, and T. Drummond, "Faster and better: A machine learning approach to corner detection," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 32, no. 1, pp. 105–119, 2010.

[11] E. Mair, G. Hager, D. Burschka, M. Suppa, and G. Hirzinger, "Adaptive and generic corner detection based on the accelerated segment test," *Computer Vision–ECCV 2010*, pp. 183–196, 2010.

[12] J. Cronje, "BFROST: binary features from robust orientation segment tests accelerated on the GPU," in *Proceedings of the 22nd Annual Symposium of the Pattern Recognition Society of South Africa*, ser. PRASA2011, 2011.

[13] J. P. de Villiers, F. W. Leuschner, and R. Geldenhuys, "Centi-pixel accurate real-time inverse distortion correction," in *Proceedings of the 2008 International Symposium on Optomechatronic Technologies*, ser. ISOT2008, vol. 7266, 2008, pp. 1–8.

[14] J. P. de Villiers, "Real-time stitching of high resolution video on COTS hardware," in *Proceedings of the 2009 International Symposium on Optomechatronic Technologies*, ser. ISOT2009, vol. 9, 2009, pp. 46–51.

[15] G. Bradski, "The opencv library," *Dr. Dobb's Journal of Software Tools*, 2000.