MOVING HOT OBJECT DETECTION IN AIRBORNE THERMAL VIDEOS

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES
OF
MIDDLE EAST TECHNICAL UNIVERSITY

BY

UTKU KABA

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR
THE DEGREE OF MASTER OF SCIENCE
IN
ELECTRICAL AND ELECTRONICS ENGINEERING

JUNE 2012

Approval of the thesis:

MOVING HOT OBJECT DETECTION IN AIRBORNE THERMAL VIDEOS

submitted by **UTKU KABA** in partial fulfillment of the requirements for the degree of **Master of Science in Electrical and Electronics Engineering Department, Middle East Technical University** by,

Prof. Dr. Canan Özgen _____
Dean, Graduate School of **Natural and Applied Sciences**

Prof. Dr. İsmet Erkmen _____
Head of Department, **Electrical and Electronics Engineering**

Prof. Dr. Gözde Bozdağı Akar _____
Supervisor, **Electrical and Electronics Engineering Dept., METU**

Examining Committee Members:

Prof. Dr. Yasemin Yardımcı _____
Graduate School of Informatics, METU

Prof. Dr. Gözde Bozdağı Akar _____
Electrical and Electronics Engineering Dept., METU

Prof. Dr. Aydın Alatan _____
Electrical and Electronics Engineering Dept., METU

Assoc. Prof. Dr. İlkay Ulusoy _____
Electrical and Electronics Engineering Dept., METU

Dr. Sait Kubilay Pakin _____
ASELSAN A.Ş.

Date: **July 11, 2012**

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Name, Last name : Utku KABA

Signature :

ABSTRACT

MOVING HOT OBJECT DETECTION IN AIRBORNE THERMAL VIDEOS

Kaba, Utku

M. Sc., Department of Electrical and Electronics Engineering

Supervisor: Prof. Dr. Gözde Bozdağı Akar

July 2012, 94 pages

In this thesis, we present an algorithm for vision based detection of moving objects observed by IR sensors on a moving platform. In addition we analyze the performance of different approaches in each step of the algorithm. The proposed algorithm is composed of preprocessing, feature detection, feature matching, homography estimation and difference image analysis steps. First, a global motion estimation based on planar homography model is performed in order to compensate the motion of the sensor and moving platform where the sensors are located. Then, moving objects are identified on difference images of consecutive video frames with global motion suppression. Performance of the proposed algorithm is shown on different IR image sequences.

Keywords: Homography, Harris, SIFT, RANSAC, Image Registration, Thermal, Motion.

ÖZ

HAVA PLATFORMLARINA ENTEGRE TERMAL KAMERALARLA HAREKETLİ SICAK NESNELERİN TESPİTİ

Kaba, Utku
Yüksek Lisans, Elektrik ve Elektronik Mühendisliği Bölümü
Tez Yöneticisi: Prof. Dr. Gözde Bozdağı Akar

Temmuz 2012, 94 sayfa

Bu tezde, sabit olmayan hava platformlarına bütünleşik IR almaçlarla gözlemlenen hareketli nesnelerin tespitinde kullanılan, görüntü tabanlı bir yöntem sunulmaktadır. Önerilen yöntem ilk işleme, öznitelik çıkarımı, öznitelik eşleştirmesi, homografi modellemesi ve fark resmi analizi ara kademelerinden oluşmaktadır. İlk olarak, almaçların ve almaçların bütünleşik olduğu platformun hareketini telafi etmek için, düzlemsel homografi bulma yöntemiyle arka plan hareketi modellenir. Sonrasında, arka plan hareketi bastırılmış ardışık görüntülerin fark resimlerinde, hareketli nesneler tespit edilir. Bahsi geçen algoritmanın performansı çeşitli IR görüntü setleri üzerinde gösterilmektedir.

Anahtar Kelimeler: Homografi, Harris, SIFT, RANSAC, Görüntü Çakıştırma, Termal, Hareket.

To those who contribute to this thesis...

ACKNOWLEDGEMENTS

I would like to express my gratitude to my supervisor Prof. Dr. Gözde Bozdağı Akar for her supervision, advice, criticism, encouragement and insight throughout the research. I would like to thank Prof Dr. Aydın Alatan, Prof Dr. Yasemin Yardımcı, Assoc. Prof. Dr. İlkay Ulusoy and Dr. Sait Kubilay Pakin for serving in my committee and sharing their opinions.

I would also like to thank Tübitak for their financial support during my graduate education.

I would like to thank my Mom, Dad and sister Duygu for their love. I feel their support always with me throughout my life and I suppose this is the source of all achievements of my life.

I would also like to thank to my colleagues for their assistance and patient. And of course to my company ASELSAN and my colleagues Erdem, Hamza and Murat for their guidance and support at all level of this study.

I would like to express my special and deepest thanks to my friends, Şeyma, Eda, Fatih and Kadir, for their encouragements, supports and friendships.

TABLE OF CONTENTS

LIST OF FIGURES

FIGURES

LIST OF TABLES

TABLES

CHAPTER 1

INTRODUCTION

Moving object detection systems are used in large variety of applications from traffic monitoring to military surveillance. In order to be able to work at day and night conditions, IR cameras are preferred for these applications. However IR cameras provide poor resolution therefore features are not as discriminative as in day cameras. Also noise levels of thermal videos are greater than day cameras. They often contain artifact on brightness which fades out at the end of scan line or local sensor errors [2]. However, thermal video characteristics bring some advantages by means of motion detection. Usually moving objects are hot compared to background and their motion with respect to scene can be extracted by using contrast difference.

Motion detection algorithms in videos obtained by stationary cameras are simpler compared to moving cameras. Since background is also moving, local motion can only be extracted if global motion of the scene is modeled. There are a lot of researches about motion detection by moving cameras. Woelk and Koch try to measure entire optical flow for day TV using constant intensity assumption of objects at consecutive frames [8]. However, its computational cost is quite high for real time systems [1]. Strehl and Aggarwal make affine global motion assumption and correlates entire frames [2]. In order to decrease complexity, they use pyramidal structure where correlation starts from the lowest resolution and finally affine transformation model is found at full resolution. However, affine model for global motion is suitable for nadir view. It is based on the assumption that rotational and translational camera motion is allowed but parallel lines remain parallel which is

violated by perspective changes. Moreover, exploiting the entire frame for global motion model results in high computational complexity.

Global motion suppression algorithms also differ by the scene depth characteristics. For the cases, where two or more planar surfaces exist due to varying depth, camera motion can be modeled by motion analysis of layers. Irani and Anandan select a plane for registration and apply a parallax based rigidity constraint over the registered image for scenes where depth is not constant [25]. They show their algorithm's performance on electro optic videos. However, computational cost of this algorithm is very high. Even global motion suppression for an image pair takes a few minutes with available processors [26].

Chellappa and Qian propose a method based on sequential importance sampling for moving object detection by moving electro optic cameras [27]. Their method is based on detecting feature points in the first image of the sequence and tracking these feature points to find an approximate sensor motion model. The algorithm uses segmentation for detection of feature points belonging to the moving objects. Their method works both with 2D and 3D scenes, however, feature selection is inherently problematic and the proposed algorithm has an off-line character [26].

Background modeling is used frequently for motion detection by stationary cameras. Since background for moving cameras is also changing, such an approach is not directly applicable to moving camera scenarios. Yu and Medioni [31] first apply image warping based on planar homography model to 91 consecutive frames in a sliding window for each frame. In order to prevent error propagation, warping of 91 frames into base frame is repeated for every frame. Then, they detect moving objects by conventional background modeling as in stationary cameras and shown to be successful on electro optic videos. However, registration of 91 frames is computationally costly and they try to solve this problem by hardware optimization of GPUs [31].

Kirchhof and Stilla use planar homography approach for global motion modeling which is reasonable because platforms that the IR sensors are mounted are generally very far from observed regions. They extract Harris corners at each frame and match

them by normalized correlation. Then, projective planar homography is estimated after outlier elimination by RANSAC. Transformation matrix of planar homography is used for registration of consecutive frames after applying difference operation to find local motion. However, if the outliers' ratio is high, more iteration must be performed in RANSAC which increases computational load. In this thesis, we propose a similar feature-based planar homography modeling where computational complexity of RANSAC algorithm is decreased by early filtering of the worst matches.

1.1 Scope of Thesis

The aim of this thesis is to develop a moving object detection algorithm that can work real time on systems with thermal airborne cameras. Proposed system has 4 main parts: 1) Preprocessing, 2) Feature detection and matching, 3) Global motion modeling and 4) Local motion identification. We also discuss, test and show the results of different approaches at each step on sample FLIR sequences with resolution 256x320 given by [30].

In preprocessing step, consecutive images are smoothed to get rid of noise and deinterlacing if necessary. Then, Harris corners are located on both images as features and matched between frames with similarity measure, normalized correlation. SIFT and other robust features are not selected due to their high computational complexity and Harris performance is proven to be comparable with SIFT for usual thermal airborne videos.

From matched features, a homography with eight parameters is modeled for global motion. Since generally feature detection and matching algorithm gives more than a hundred of pairs, over determined solution is found by normalized direct linear transform. However, matched features have not only measurement noise and but also outliers which corrupts the mean like solutions. Therefore outliers are eliminated by iterative random sample consensus (RANSAC). In order to decrease number of iterations and thus complexity of system, very distinctive outliers are eliminated

prior to homography estimation. Then, final inliers are used to construct a homography model, which is simply 3x3 transformation matrix defined in homogenous coordinates.

Finally, homography model is used to warp images and difference of warped and base image is obtained. Since moving objects are hot compared to background, two blobs are seen on difference image, one negative and positive. Blob structure is exploited to remove erroneous alarms by morphological operations considering the fact that there should be an anti-blob for each non-zero region close to itself.

In this work, performance of the proposed algorithm is evaluated on several videos with different parameter selections in each subsection of the algorithm.

1.2 Outline of Thesis

In Chapter 2, related studies on the global motion suppression are analyzed in two subsections, intensity-based and feature-based. In Chapter 3, proposed algorithm is presented on a sample thermal image pair by showing each intermediate results of the motion detection system. Moreover, different alternatives to proposed system are compared and results are discussed. Finally in Chapter 4, possible future works and conclusion are given.

CHAPTER 2

LITERATURE SURVEY

The most crucial part of the moving object detection in videos where background is not stationary is global motion estimation. If global motion is modeled well, local motion can be easily discriminated. Several global motion estimators are used in literature for different applications which can be grouped as intensity-based and feature-based. Intensity-based algorithms use complete images to extract global motion parameters. However, feature-based algorithms exploit a number of key points/regions to fit a model for global motion which results in low computational complexity.

2.1 Intensity-Based Global Motion Estimation

2.1.1 Phase Correlation

Phase correlation methodology exploits the frequency spectrum properties of images. If a 2D signal is shifted, it results in a phase shift on frequency domain. If we model global motion by pure translation, relation between two images can be expressed with Equation (2–1).

$$f_2\ x,y\ = f_1\ x+m,y+n \tag{2–1}$$

where $f_1(x,y)$ and $f_2(x,y)$ are two images and (m,n) is the displacement in 2D. Equation (2-2) is expressed in frequency domain as follows:

$$F_2\ w_1, w_2\ = F_1\ w_1, w_2\ exp\ -j2\pi\ w_1 m + w_2 n \tag{2–2}$$

Then, frequency domain correlation image is defined with Equation (2–3) where denominator is added to normalize.

$$C_{f_1,f_2}\ w_1, w_2\ = \frac{F_1(w_1,w_2)F_2^*(w_1,w_2)}{|F_1\ w_1,w_2\ F_2^*\ w_1,w_2\ |} \tag{2–3}$$

Taking inverse Fourier transform of correlation image, translation vector is obtained.

$$F^{-1}\{C_{f_1,f_2}(w_1, w_2)\}\ =\ c_{f_1,f_2}(x, y)\ =\ \delta(x - m, y - n) \tag{2–4}$$

Therefore, translation between two images can be calculated by Equation (2–4). Location of peak in image $c_{f_1,f_2}(x, y)$ gives translation in x and y directions.

If also local motion exists between images, there will be some other impulses with lower magnitudes. However, global motion impulse will be dominant in magnitude which makes global motion modeling possible with local motion. An example is given in Figure 2–1 with one directional motion. When there is only global motion, Figure 2–1b shows that phase correlation method gives an impulse at the translation point. However, Figure 2–1c and Figure 2–1d have local motions which do not bias main motion impulse [7].

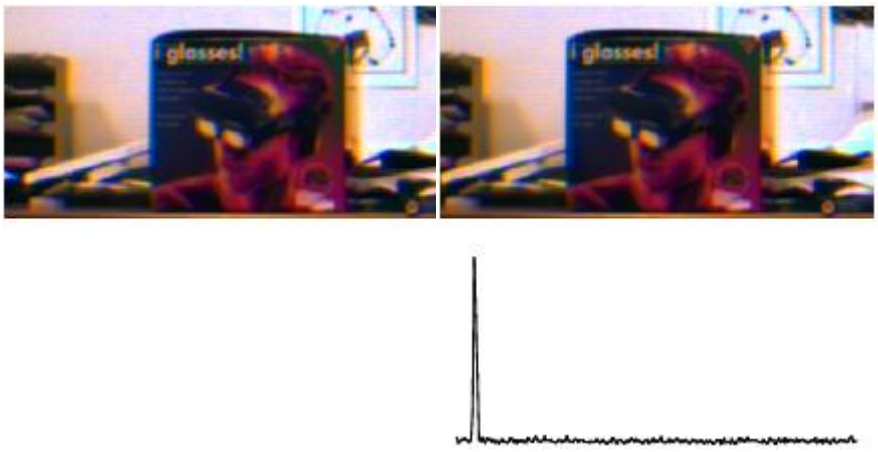


Figure 2–1 One dimensional phase correlation result (a) original image, (b) global translation and phase shift , (c) and (d) global and local motion coexist [7]

However, if rotational and scale changes are significant, motion between two images cannot be modeled with simple translation. In such cases, phase correlation algorithm does not give promising results [7].

2.1.2 Optical Flow Estimation

Optical flow estimation method exploits correlation between temporal and spatial derivatives of images at each pixel. Assuming that illumination does not change between images, relation between two consecutive images for each pixel can be defined with Equation (2–5)

$$f(x, y, t) = f(x + \Delta x, y + \Delta y, t + \Delta t) \tag{2–5}$$

where Δx and Δy corresponds to pixel movement and f(x,y,t) is pixel intensity at location (x,y) and time t. If Δx and Δy are small, optical flow can be approximated by Taylor series expansion as follows:

$$f(x + \Delta x, y + \Delta y, t + \Delta t) \cong f(x, y, t) + \frac{\partial f \Delta x}{\partial x} + \frac{\partial f \Delta y}{\partial y} + \frac{\partial f \Delta t}{\partial t} \tag{2–6}$$

which results in final optical flow equation.

$$\frac{\partial f}{\partial x}\frac{\partial x}{\partial t} + \frac{\partial f}{\partial y}\frac{\partial y}{\partial t} = -\frac{\partial f}{\partial t} \tag{2–7}$$

or notation can be simplified as follows:

$$f_x v_x + f_y v_y = -f_t \tag{2–8}$$

where v_x and v_y are two unknowns stands for pixel motion velocity. Since there is only one equation, two unknowns cannot be solved alone. In order to introduce more equations, some other constraints should be used [6]. Lucas-Kanade assumes that pixel velocity vector of each pixel is same in its neighborhood. Then for neighboring pixels p_1, p_2 …. p_n, a set of equations are obtained. These equations give over determined solution for v_x and v_y. In order to find best model for pixel movement, least squares solution is found.

$$f_x\ p_1\ v_x + f_y\ p_1\ v_y\ = -f_t\ p_1$$

$$f_x(p_2)v_x + f_y(p_2)v_y = -f_t(p_2)$$

.

.

.

$$f_x(p_n)v_x + f_y(p_n)v_y\ = -f_t(p_n) \tag{2–9}$$

These set of equations can be written in matrix form as follows:

$$A = \begin{matrix} f_x(p_1) & f_y(p_1) \\ f_x(p_2) & f_y(p_2) \\ \vdots & \vdots \\ f_x(p_n) & f_y(p_n) \end{matrix} \tag{2–10}$$

$$v = \begin{matrix} v_x \\ v_y \end{matrix} \tag{2–11}$$

$$t = \begin{matrix} f_t(p_1) \\ f_t(p_2) \\ \vdots \\ f_t(p_n) \end{matrix} \tag{2–12}$$

$$A.v = t \tag{2–13}$$

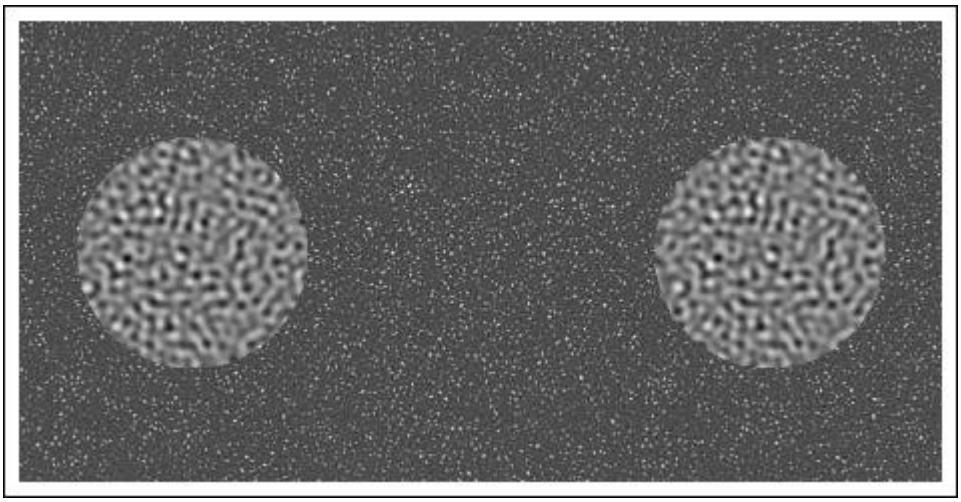
Least squares solution for v is given in Equation (2–14)

$$v_{LS} = (A^T A)^{-1} A^T t \tag{2–14}$$

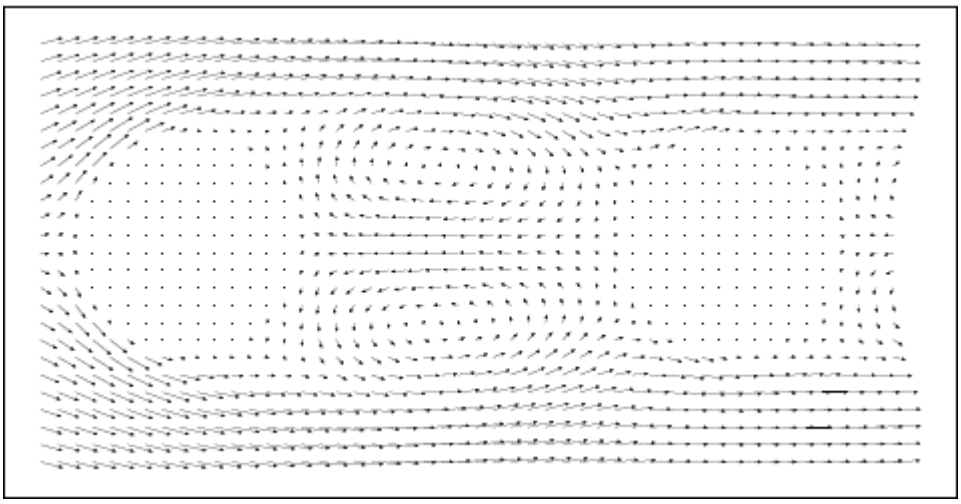
Least squares solution gives same importance to all pixels inside defined window. However, a better approach emphasizes the center pixel. This is achieved by a Gaussian windowing matrix. Center pixel emphasized least squares solution is given below:

$$v_{LS\_W}\ =\ (A^T W A)^{-1} A^T W t \tag{2–15}$$

Where W is 2D Gaussian matrix which results in solution of least squares dominated by center pixel solutions of Equation (2–15)



(a)



(b)

Figure 2–2 Optical flow example
(a) original image (b) corresponding optical flow [20]

Woelk and Koch constructs complete optical flow model for image sequences and extracts global motion. However, computing optical flow for each pixel is

computationally costly. Therefore, optical flow based approach is not used in this work.

2.1.3 Affine Global Motion Estimation

Instead of computing translational motion, affine models are also used in motion estimation [2]. Affine global motion estimation starts with affine image transformation model assumption which has 6 degrees of freedom. It assumes that rotational, translational and scaling effects are possible between images but parallel lines remain parallel which hold for cameras with nadir view. Affine transformation in homogenous coordinates is modeled with Equation (2–16).

$$M\ \boldsymbol{x},\theta\ = \begin{matrix}\theta_1 & \theta_2 & \theta_3\\ \theta_4 & \theta_5 & \theta_6\end{matrix} \cdot \begin{matrix}x\\ y\\ 1\end{matrix} \tag{2–16}$$

where unknown parameters are

$$\theta = \ \theta_1 \quad \theta_2 \quad \theta_3 \quad \theta_4 \quad \theta_5 \quad \theta_6\ ^T \tag{2–17}$$

Unknown parameters in Equation (2–17) define global motion. Aggarwal estimates these parameters by iterative registration of warped images where motion parameters are updated in each step [2]. He correlates complete images at each step, estimates registration error and updates unknown transformation vector to minimize error.

At each iteration, θ is updated with following equations:

$$\theta_{n+1} = \theta_n + \delta_n \tag{2–18}$$

and increment at each iteration is:

$$\delta_n = -(\ \ {}_xP^T(\nabla I^{t-\tau})\ \nabla I^{t-\tau}\ {}^TP)^{-1}(\ \ {}_xP^T(\nabla I^{t-\tau})\Delta I_n) \tag{2–19}$$

$$P = \begin{matrix}x & y & 1 & 0 & 0 & 0\\ 0 & 0 & 0 & x & y & 1\end{matrix} \tag{2–20}$$

ΔI_n is the pixel wise difference between reference image and registered image. $\nabla I_{t-\zeta}$ is the gradient image of the previously registered image. Iterations are performed until θ is slowly changing.

In order to compensate only global motion and left with local motion, full image correlation is performed. Since dominant motion between images is global motion, final transformation matrix does not suppress local motion. However, in order to extract scene global motion, full image correlation is necessary which results in significant computational cost. Aggarwal uses pyramidal correlation as in Figure 2–3 to decrease computational cost. Transformation matrix found in lowest scale is used as initial starting condition for upper a scale. At the end, complete images are correlated [2].

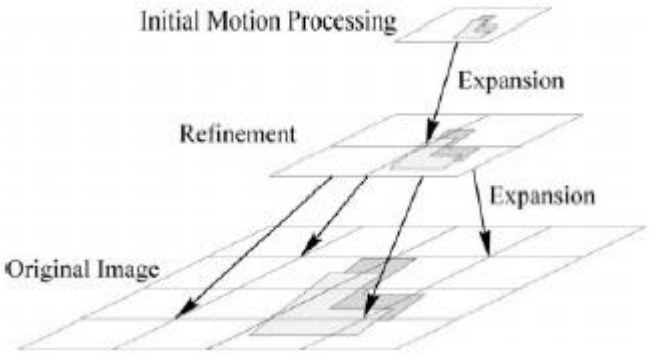


Figure 2–3 Pyramidal registration [2]

2.2 Feature-Based Global Motion Estimation

Intensity-based global motion estimation algorithms use complete images to model motion. Therefore, this results in high computational load. Instead of it, some key points, called as features can be identified to describe image motion characteristics. Such an approach would decrease computational complexity. Feature-based algorithms try to find matches between features of consecutive frames. Then from matched features, transformation matrix is calculated for global motion suppression.

Features can be either selected on both images and matched by similarity criteria or tracked on second image. Therefore, feature selection is quite crucial. Features shall be easily tracked, localized well and invariant to global motion. Moreover, since motion detection algorithms generally work on real time systems, computational load of feature extraction and matching/tracking shall be low. Therefore several features extraction and matching algorithms are investigated in following subsections.

2.2.1 Feature Extraction Methods

(i). Moravec Corner Detection

Moravec corner detection algorithm defines a window for a candidate pixel and moves this window in multiple directions to measure the patch similarity under the window [24]. If tested pixel is on a uniform region, no significant changes are observed when window is moved in any direction. If it is on an edge, when window is moved parallel to edge, there is no change. However, perpendicular movement of window results in significant patch change. On the other hand, if patch center is on a corner, in all direction, similarity changes significantly.

Since an edge can be in any direction, Moravec's algorithm necessitates movement of window in any direction for each pixel which is inefficient by means of computational complexity.

(ii). Harris Corner

Harris corner detection can be considered as an improved version of Moravec corner detection [3]. Harris exploits the gradients of candidate pixels. Gradient of a pixel is very low if it is on a uniform region. On exact location of a corner, gradient is not well defined but in its neighborhood, gradients are distinct and directions differ abruptly. On an edge pixel however, gradients are perpendicular to edge direction in its neighborhoods.

Harris corner detection over an image starts with obtaining derivatives in two dimensions namely I_x and I_y. I_x is constructed by convolving original image with [1

0 -1] mask. Similarly I_y is constructed with [1 0 -1]. Then structure tensor computed [3].

$$S_T = \sum_u \sum_v w(u,v) \begin{matrix} {I_x}^2 & I_x I_y \\ I_x I_y & {I_y}^2 \end{matrix} \quad (2\text{–}21)$$

S_T is the structure tensor matrix for pixel at (x,y). To increase sensitivity for cornerness criteria, averaging in a window is performed for each (u, v). As in stated in [23], window size of S_T is not so critical. However, if it is too high, cornerness information can be smoothed and if it is too low, noisy result are obtained. Selecting an intermediate value is feasible.

If 2 eigenvalues of matrix S_T are high positive values, a corner is marked. However, computing these eigenvalues is computationally costly. Better approach is computing the following equation:

$$M_C = \lambda_1\lambda_2 - \mathrm{K}(\lambda_1 + \lambda_2)^2 = \det(S_T) - \mathrm{K}\, trace^2(S_T) \quad (2\text{–}22)$$

$\mathrm{M_C}$ is used as a measure of cornerness. That is for higher $\mathrm{M_C}$ values, stronger corners are identified. K is empirical constant and usually chosen as 0.04.

If features are selected from an image from higher $\mathrm{M_C}$ values, they are concentrated on corners because Mc is high at exact corners and their neighborhoods. Therefore, in order to localize a point; non-maximum suppression is performed in a defined window.

Harris corners are used widely as features in literature because of the fact that computational complexity of feature extraction and matching is low compared to others. Theoretically, Harris corners are invariant to rotational and translational changes but not scale, affine or projective transformations.

(iii). Minimum Eigen Value Method

Minimum eigenvalues method also relies on the structure tensor S_T. Different from Harris corner detector, it calculates directly eigenvalues of matrix S_T and measure of feature is constructed by minimum of eigenvalues. It is also best feature by means of traceability and sometimes called as KLT feature detector [9]. Practically there is no

significant difference between Harris and minimum eigen value features. However, eigen value computation is costly compared to Harris measure [12]. Therefore, we only evaluate Harris performance but not the Minimum Eigen Value method.

(iv). Harris Laplace Detector

Harris Laplace detector is the scale invariant version of Harris corner detector [23]. It simply constructs images at different scales by convolving Gaussian masks with different variances as given by Equation (2–23). Then, Harris corners are identified with Equation (2–22). However, smoothing weakens the features and makes detecting corner difficult at higher scales. Therefore, normalization is performed at each scale. Normalized Gaussian mask can be found in [23].

(v). SIFT

Scale-invariant feature transform [10] has been introduced by David Lowe and used frequently since 1999. The reason that makes SIFT so popular is that detected features are scale and orientation invariant and partially invariant to illumination and affine changes that makes matching of features at different orientation, scale and illumination levels possible.

Feature extraction algorithm over an image f(x,y) starts with creating an octave of images where each image is formed by filtering f(x,y) with Gaussian masks at different scales as given in Figure 2–4. Filtered image is given by:

$$L(x, y, ki\sigma) = f(x, y) * G(x, y, ki\sigma) \quad (2\text{–}23)$$

In Equation (2–23), kiσ produces Gaussian masks at different scales. After creating images at different scales, DoGs (Difference of Gaussians) are formed. A DoG image, represented by D(x,y,kiσ) is found by following equation:

$$\mathrm{D}(\mathrm{x}, \mathrm{y}, \mathrm{ki}\sigma) = \mathrm{L}(\mathrm{x}, \mathrm{y}, \mathrm{ki}\sigma) - \mathrm{L}(\mathrm{x}, \mathrm{y}, \mathrm{kj}\sigma) \quad (2\text{–}24)$$

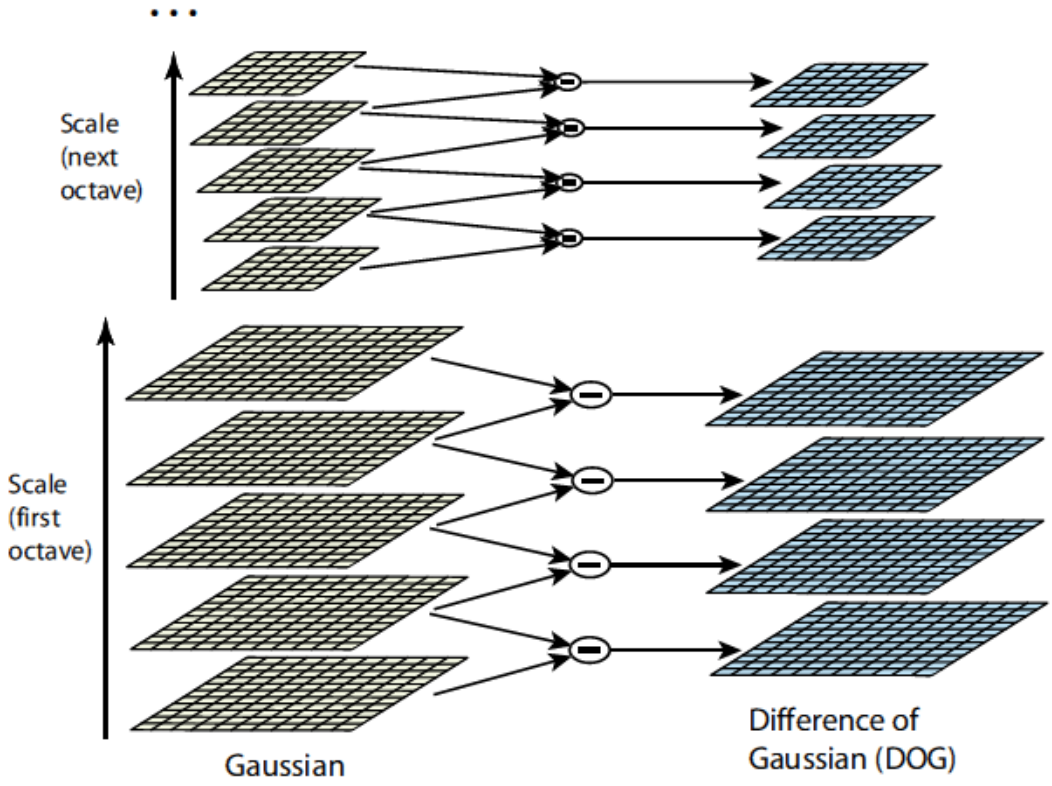


Figure 2–4 DoG images at different scales [10]

A pixel at location (x,y) is selected as feature if D(x,y,kiσ) is the maximum or minimum of its 8 neighbors at same scale and 9 neighbors at the upper and lower scale. In total, each candidate feature pixel is compared with 26 neighborhood pixels. This gives rough locations of features. In order to increase sensitivity, sub-pixel locations of features are found by approximating D(x). However, some features may be located on edges or low contrast regions. Since low contrast regions are sensitive to noise and features are poorly localized on edges, detected features on edges and low contrast regions are eliminated. After localization of features, descriptor is formed on scale kiσ which provides scale independence. In order to ensure orientation independence, gradient image is formed and each feature is oriented in the direction of dominant gradient.

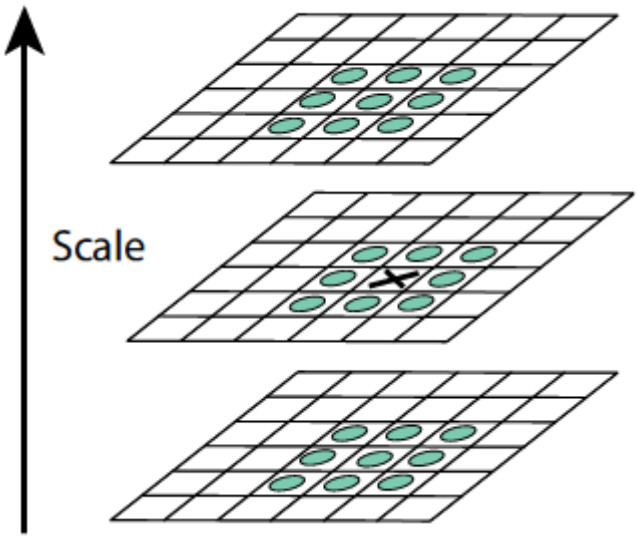


Figure 2–5 Marked pixel is tested with 26 neighborhood pixels [10]

Descriptor of detected feature is created from gradient image in a window. Then histogram of gradients is formed in predefined bin number. If bin number is b, gradient directions are represented by 360/b degrees. For a window of w*w, feature descriptor vector has a dimension of w*w*b. In general, SIFT feature descriptor is formed from 8 bins of histogram in 4*4 grids which results in 128 dimensional vector. At the end, descriptor vector is normalized to eliminate illumination change effects.

SIFT detects more features than other algorithms described above. Main reason of that feature detection is performed over multiple scales. Mikolajczyk and Schmid shows that SIFT outperforms most of the feature detectors where image deformations exist [13].

Figure 2–6 SIFT used for object recognition; small rectangles show matched features where one of the trains is oriented and occluded [10]

(vi). SURF

SURF (Speeded Up Robust Features) has been developed on the idea of SIFT with similar feature quality properties but has better performance by means of computational complexity [11].

It is based on approximation of Hessian matrix which is defined with Equation (2–25) where σ corresponds to second order Gaussian derivative scale and *x* is a pixel in image f(x,y).

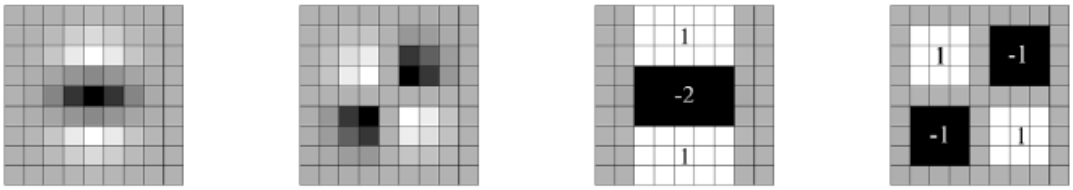
$$H(x,\sigma) = \begin{matrix} L_{xx}(x,\sigma) & L_{xy}(x,\sigma) \\ L_{xy}(x,\sigma) & L_{yy}(x,\sigma) \end{matrix} \tag{2–25}$$

$$L_{xx}\ x,\sigma\ = \frac{\partial^2}{\partial^2 x^2} g(\sigma) \tag{2–26}$$

$$L_{xy}\ x,\sigma\ = \frac{\partial^2}{\partial x \partial y} g(\sigma) \tag{2–27}$$

$$L_{yy}\ x,\sigma\ = \frac{\partial^2}{\partial^2 y^2} g(\sigma) \tag{2–28}$$

Gaussian second order derivative mask is approximated by rectangular integral operators as given Figure 2–7. Since operators are chosen to speed up the process with rectangular approximation, it is called as "Fast-Hessian" detector.



(a) (b) (c) (d)

Figure 2–7 Gaussian derivatives and approximations, (a) Gaussian y derivative, (b) Gaussian xy derivative, (c) and (d) are box approximations [10]

To provide scale independence, instead of smoothing and down sampling the image, Fast-Hessian operation is performed for different scales. Initial mask size is chosen to be 9x9 as in Figure 2–7 which corresponds to σ=1.2 (also initial scale). Then for each octave, filter size is doubled. That is, Gaussian second derivative masks of size 15x15, 21x21 and 27x27 are used without changing the filter constants. Then, non-maximum suppression over 3-3-3 neighborhood is applied for higher determinants of Fast-Hessian Matrix over octaves. In order to be more precise, determinants are interpolated to find sub-pixel resolution locations.

Orientation independence is provided by similar orientation assignment with SIFT. Instead of gradients, Haar-wavelet responses are found and patches are oriented at the dominant direction.

Feature descriptor is created from a region of 20s where s corresponds to scale of detected feature. Then 20s region is divided into 4*4 square smaller regions. In each region, magnitude of Haar wavelet responses are used to create a feature vector v. Here *dx* and *dy* represents the magnitude of Haar wavelet responses in x and y direction, respectively. In total, dimension of SURF descriptor is 64. Compared to 128 of SIFT, SURF has lower dimension and lower computational cost for feature matching.

$$v = (\quad dx, \quad dy, \quad |dx|, \quad |dy|) \tag{2–29}$$

Features matching criteria is based on descriptor distance. Mahalanobis or Euclidean distance measures are calculated for similarity and lower distance means higher similarity.

(vii). ORB

SIFT is as mentioned before, considered as the most robust feature detector by means of robustness to orientation, scaling and noise deformations. However, its computational complexity is quite high that makes usage of SIFT difficult or expensive for real time systems. ORB (Oriented Fast and Rotated BRIEF) is proposed as an alternative to SIFT with similar quality but with better performance by means of computational complexity [4]. It is based on FAST feature detector and BRIEF (Binary Robust Independent Elementary Features) descriptors [16]. For details, refer to [4].

2.2.2 Feature Matching Algorithms

Features detected on consecutive frames are matched by a similarity measurement. Different similarity measures are used in literature and their general approach is based on distance measurement between all possible pairs and matching the closest pair. Frequently used algorithms are mentioned below.

2.2.2.1 Simple Pixel Intensity Based Matching

If one image patch is shifted without changing illumination, rotation and scale, simple block matching can be applied. If patches to be compared for similarity have d pixels, distance between them can be defined with Equation (2–30).

$$D\ P_1, P_2 = \sum_{i=1}^{w} \sum_{j=1}^{w} | P_1\ i,j - P_2\ i,j | \qquad (2–30)$$

P1 and P2 are two patches and D is the corresponding distance measure. For a patch, the most similar patch from a set can be chosen with lowest distance.

Pixel based matching is not generally used because it is very sensitive to noise, illumination and geometric changes.

2.2.2.2 Normalized Correlation

Normalized correlation for two image patches is defined with Equation (2–31).

$$R\ P_1, P_2\ = \frac{\sum_{i=1}^{w}\ \sum_{j=1}^{w}\ (P_1(i,j) - \mu_1)\ (P_2(i,j) - \mu_2)}{\sigma_1 \sigma_2} \qquad (2–31)$$

R defines correlation between two patches P_1 and P_2 that is if R is high, two patches are considered as similar.$\mu_1$and μ_2 are the means and σ_1 and σ_2 are variances of two patches.

Normalized correlation more robust than simple pixel comparison and used frequently [1]. However rotational and scale changes may result in erroneous matches.

2.2.2.3 Descriptor Distance Based Matching

Features represented by descriptor vector can be compared for similarity using a norm. SIFT descriptors are generally compared according to Euclidean distance. In order to decrease false matches, Low checks best matches with second best checks. If feature P_1 matches with P_2 and second best match is between P_1 and P_3, Low calculates the ratio d(P1, P2)/d(P1, P3) and if it is greater than 0.8, rejects the match [10].

ORB uses on the other hand, bit stream of intensity comparisons. The similarity measurement criterion of ORB is Hamming distance where as the number of different bits in a compared pair increases, similarity decreases.

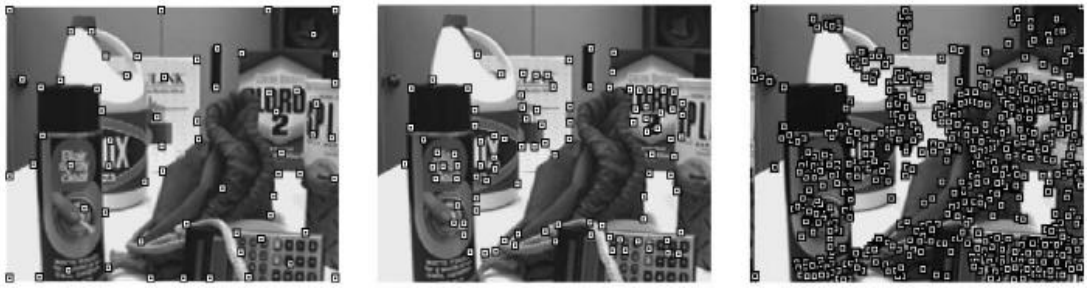
2.2.2.4 Feature Tracking

Instead of feature extraction and matching on both images, features detected on first image can be tracked to find correspondence points. KLT is frequently used to track features [6]. However, image deformations lead to bad feature tracking and drift which are seen frequently in IR videos [1]. Therefore, we do not propose to use feature tracking.

2.2.3 Comparison of Features

Feature matching performance significantly effects global motion estimation quality thus moving object detection. Several frequently used feature extraction and matching algorithms are mentioned before one of which is selected for our proposed algorithm. Among these, SIFT, SURF and ORB have similar claims being good in rotation, scale, affine and noise invariance. Harris corners, on the other hand, are translation and rotation invariant but not so robust under affine and scale changes. However, feature extraction and matching of Harris corners is very fast compared to SIFT, SURF, ORB and BRIEF.

Since Morevac's algorithm is primitive version of Harris, we did not evaluate its performance. Moreover, minimum Eigen value method (or KLT feature detector) is computationally costly but detects almost same features with Harris. Therefore, we discarded minimum Eigen value method from our experiments.



(a) (b) (c)

Figure 2–8 Feature detection result of (a) Harris, (b) KLT and (c) SIFT [13]

Bulla compares SIFT, SURF, ORB and BRIEF by means of correct feature matching ratio. For this, he uses 1000 object images of Amsterdam Library of Object Images under aforesaid distortions [18]. He chooses one of the images as a test image and compares the correct match probability for SIFT, SURF and ORB descriptors. Features are matched if distance ratio between best second match and first match is greater than 0.8. Others are eliminated. And the test results show that "The SIFT descriptor, despite of its age, still outperforms modern descriptors in many tested scenarios" [18].

One can also combine different feature detection algorithms with descriptors. For example, Harris features can be defined with SIFT descriptors. Blanco and Gonzalez compares performance of multiple combinations of feature detectors and descriptors and shows that Harris based features are fastest whereas SIFT based features are the most robust ones. Their experiments' results are shown in Figure 2–9.

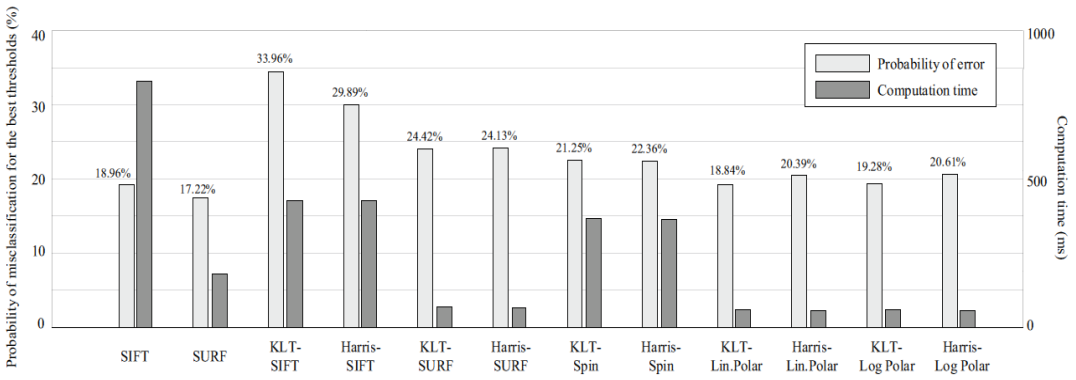


Figure 2–9 Performance comparison of several features, Harris is very fast, SIFT is very robust [22]

However, computational efficiency of proposed moving object detection is very important where Harris is advantageous. Since Harris Laplace detector is computationally costly compared to Harris and primate to SIFT by means of scale independence, in Chapter 3 we only evaluate performance of Harris and SIFT with usual airborne videos.

2.2.4 Registration Algorithms

When key points are extracted and matched between two images, $f_1(x,y)$ and $f_2(x,y)$, a global motion transformation is defined to map $f_1(x,y)$ to base of $f_2(x,y)$. There are several approaches to find transformation model in literature which are analyzed in the following subsections.

2.2.4.1 Homography Transformation

A point (x,y) in 2D can be represented in 3D by mapping to (x,y,1) and this is called as homogenous representation. If there is an invertible matrix H which maps every point between two images by Equation (2–32), a homography can be defined. Hartley and Zisserman states that a homography can be defined between two images

if and only if 3 collinear points in one image are also collinear. This hold for images where scene is planar [19].

$$c \begin{matrix} x' \\ y' \\ 1 \end{matrix} = H \begin{matrix} x \\ y \\ 1 \end{matrix} \quad (2\text{–}32)$$

2.2.4.1.1 Estimation of Homography Matrix

H is 3-3 homography matrix which has 9 unknowns. Since c is arbitrary constant, H has 8 degrees of freedom. Therefore H can be found exactly using 4 matched points if matched points are not collinear. If there are more than 4 matched points as in most of the cases, a homography that minimize predefined an error function is estimated for noisy scenarios.

2.2.4.1.1.1 Normalized DLT

DLT (Direct Linear Transform) is a linear solution of Equation (2–32) that minimizes ||Ah||=0. It starts with expressing Equation (2–32) in the form of Ah=0:

$$H = \begin{matrix} h_1 & h_2 & h_3 \\ h_4 & h_5 & h_6 \\ h_7 & h_8 & h_9 \end{matrix} \quad (2\text{–}33)$$

Then Equation (2–32) is equivalent to Equation (2–34).

$$-h_1 * x - h_2 * y - h_3 + h_7 * x + h_8 * y + h_9 * x' = 0$$

$$-h_4 * x - h_5 * y - h_6 + h_7 * x + h_8 * y + h_9 * y' = 0 \quad (2\text{–}34)$$

$$A = \begin{matrix} -x & -y & -1 & 0 & 0 & 0 & xx' & xx' & x' \\ 0 & 0 & 0 & -x & -y & -1 & xy' & yy' & y' \end{matrix}$$

$$h = \begin{matrix} h_1 & h_2 & h_3 & h_4 & h_5 & h_6 & h_7 & h_8 & h_9 \end{matrix}^T$$

$$Ah = 0 \tag{2–35}$$

If h = 0, Equation (2–35) is directly satisfied. Therefore a constraint ||h|| = 1 is predefined. Magnitude is insignificant because constant c in Equation (2–32) normalizes homography matrix H.

If there are four matched points which are not linear, an exact solution can be found. If there are more than four matched points, as in real applications, locations of pairs will be noisy and only an approximate solution can be found. Using Singular Value Decomposition (SVD), best approximation to h is found. Specifically if A= UDVT where D is diagonal with positive and ascending values, h is the last column of V [19].

Homogenous representation may result in divergence from correct result in noisy matched features sets. Typically pixel values in order of 100 however third element of a feature in 3D is set as one. Therefore small noise may result in high corruption at third element. This exaggerated noise effect is eliminated by normalization. Before finding H matrix, a transformation matrix is used for mapping each point to another domain where centroid of all points is carried to origin and divergence brought to $\overline{2}$. At normalized new domain, $\hat{H}$ is calculated with SVD then retransformed into original domain by Equation (2–36).

$$H = (T')^{-1}\hat{H}T \tag{2–36}$$

T' and T are the transformation matrixes for points in second image and first image, respectively.

Actually, normalization step provides non-invariance to similarity transformations [19]. Non-invariance means that for a solution of error minimization under noisy scenarios, if resulting homography matrix is H for feature correspondences $x_i \rightarrow x_i'$ between two images, when similarity transformations are applied to original images,

resultant homography matrix of error minimization algorithm can be found by Equation (2–37).

$$x_i = T_S x_i$$

$$x_i' = T_S' x_i'$$

And combining with Equation (2–32),we left with:

$$x_i' = T_S' H T_S^{-1} x_i$$

Then,

$$H = T_S' H T_S^{-1} \quad (2–37)$$

In other words, solution of error minimization algorithm should converge to H. Unfortunately, DLT is not invariant. Therefore, normalization step is necessary. As will be discussed in following sections, geometric error minimizations algorithms are also non-invariant to similarity transformations [19].

2.2.4.1.1.2 Geometric Error Minimization

Normalized DLT algorithm minimizes ||Ah|| which is not so meaningful analytically. Instead of it, total distance between projected points and correspondences can be minimized. For N correspondences $x_i \rightarrow x_i'$ between two images, an error function can be defined with Equation (2–38).

$$\epsilon = \sum_{i=1}^{N} (d(x_i', H x_i)^2 + d(x_i, H^{-1} x_i')^2) \quad (2–38)$$

First term is the 2D forward transfer error and second term is backward transfer error. Geometric error function is intuitively meaningful and is minimized by iterative algorithms such as Newton's method or gradient descent.

2.2.4.1.1.3 Geometric Reprojection Error Minimization

Geometric reprojection error minimization algorithm uses similar error function with geometric error minimization. In this case, perfect homography matrix H is found which correlates points close to original pairs $x_i \rightarrow x_i'$. Error function is defined with Equation (2–39)

$$\epsilon = \sum_{i=1}^{N} (d(x_i', x_i')^2 + d(x_i, x_i)^2) \quad (2–39)$$

where

$$x_i' = H x_i \quad (2–40)$$

It is analogous to locate a real feature in 3D then project into first image and second image. However, computational cost of minimizing geometric reprojection error is higher compared to geometric error because other than unknown 8 parameters of $\hat{H}$, 2D coordinates of features x_i have to be found. Therefore, we do not use geometric reprojection error minimization in our implementations.

Geometric reprojection error minimization algorithm gives better results compared to normalized DLT. However, since it is an iterative algorithm, computational complexity is higher compared to linear approach. It needs an initial homography solution which is generally chosen as solution of normalized DLT algorithm. Furthermore, it may stack on local minima and not converge to correct solution. Also ending of iterations should be provided by well-defined and analyzed algorithms [19]. Therefore we do not propose to use iterative error minimization algorithm but compare and discuss its performance with usual thermal airborne videos in Chapter 3.

2.2.4.1.1.4 Homography Estimation by RANSAC

Random Sample Consensus is used to fit an input data to a hypothesis where some outliers exist and used frequently in literature in many fields. Error minimization

algorithms mentioned above may work gently if the only error source is measurement noise. However, if outliers which can be considered as complete wrong data exist in input set, minimized solutions start to diverge from correct result. Homography estimation by RANSAC first eliminates outliers then applies aforesaid error minimization algorithms.

For robust homography estimation, input set is all matched features where some mismatched pairs (outliers) exist. For homography estimation, hypothesis is that for a single homograph matrix H, all pairs are related with Equation (2–32).

RANSAC is an iterative method, tries to find best homography from all iterations. It starts with random selection of 4 matched pairs. This is the minimal set that defines a homography. From initial set, a homography matrix is calculated and matched points obeying hypothesis of H are counted as inliers. Inliers check is done by Equation (2–41).

$$MFS = \{(X_1', X_1), (X_2', X_2), (X_3', X_3), \dots (X_P', X_P)\}$$

$$MS_i = \{\{(X_{R1}', X_{R1}), (X_{R2}', X_{R2}), (X_{R3}', X_{R3}), (X_{R4}', X_{R4})\}, \quad 1 \le R_1, R_2, R_3, R_4 \le P$$

$$|X_j' - H_i * X_j| < T_D \tag{2–41}$$

Here (X_j', X_j) is a matched pair in MFS (Matched Features Set) defined in homogenous coordinates, P is number of matched pairs, T_D is the distance threshold and H_i is calculated homography matrix at i^{th} iteration from MS_i (Minimal Set at i^{th} iteration). After checking each matched pairs by Equation (2–41), number of inliers is count and a new iteration is performed by random selection of a new minimal set. At the end, homography matrix is re-estimated with highest inliers set using an error minimization algorithm.

Selection of T_D is very critical. It defines the maximum error for being inlier. If it is set as too low, inliers may be interpreted as outliers. And if it is set as too high, some outliers may be included in inliers set which would result in bad homography estimation. In general, T_D is set for a specific implementation, empirically [19].

Since RANSAC is an iterative method, it may not end with correct homography. If the probability of that the minimal set includes an outlier is high, more iterations shall be performed. Hartley and Zisserman state that necessary number of iterations, N to get outliers free minimal set at a probability of p can be found by Equation (2–42).

$$N = log(1-p)/log(1-(1-\varepsilon)^{s})$$

$$\varepsilon = NumberOfOutliers/P \quad (2\text{–}42)$$

where s is the size of MS which is 4 in this case.

Table 2–1 Necessary iterations to get outliers free MS with p equals to %99, obtained by Equation (2–42) [19]

Sample size	Proportion of outliers ε						
s	5%	10%	20%	25%	30%	40%	50%
2	2	3	5	6	7	11	17
3	3	4	7	9	11	19	35
4	3	5	9	13	17	34	72
5	4	6	12	17	26	57	146
6	4	7	16	24	37	97	293
7	4	8	20	33	54	163	588
8	5	9	26	44	78	272	1177

Hartley and Zisserman state that MS for homography estimation may be constructed from larger sets than conventional four members sets [19]. This would increase quality of registration due to the fact that constructed hypothesis from initial MS get close to correct global motion. However, it decreases the probability of outlier free MS. Therefore, more iteration is performed for the same p given by Equation (2–42). Since as the iteration number increases, computational complexity of algorithm increases as well, Hartley and Zisserman do not propose to use larger MS [19].

2.2.4.1.1.5 LMedS

Least Median Squares another robust estimation method which is used to find a model from an input data set where measurement noise and outliers may exist and quite similar to RANSAC. Intuitively, median of a noisy set with outliers closer to correct result compared to mean. LMedS exploits the assumption that most of the inputs in data set are inliers and median of errors are resultant from inliers. Therefore, if the outlier fraction exceeds %50, LMedS fails.

LMedS starts with random selection of four matched pairs. From initial set, homography matrix is calculated and error for each matched feature is sorted. Then median of errors for feature are stored. Random selection of initial set continues for predetermined iteration number. With the lowest median of errors, inliers are classified and final homography matrix is calculated.

2.2.4.2 Affine Transformation

Rather than eight-parameter modeling of transformation, transformation can be simplified to six parameters given in Equation (2–16) with nadir view assumption of the camera. Then, similar registration method described for homography transformation matrix is performed. Reduction in number of unknowns decreases computational complexity. However, as in most of the airborne thermal cameras, forward looking is possible. Therefore, homography solution with eight degrees of

freedom is more feasible for airborne videos [1]. We also evaluate affine modeling performance in Chapter 3.

CHAPTER 3

PROPOSED ALGORTIHM

In the proposed algorithm, we use a feature-based technique as given in Figure 3–1. The basic steps of the algorithm are composed of preprocessing, feature detection, feature matching, homography estimation and motion identification on difference image. Each step of the algorithm and the intermediate results will be presented on the example consecutive frames given in Figure 3–2 in the following subsections.

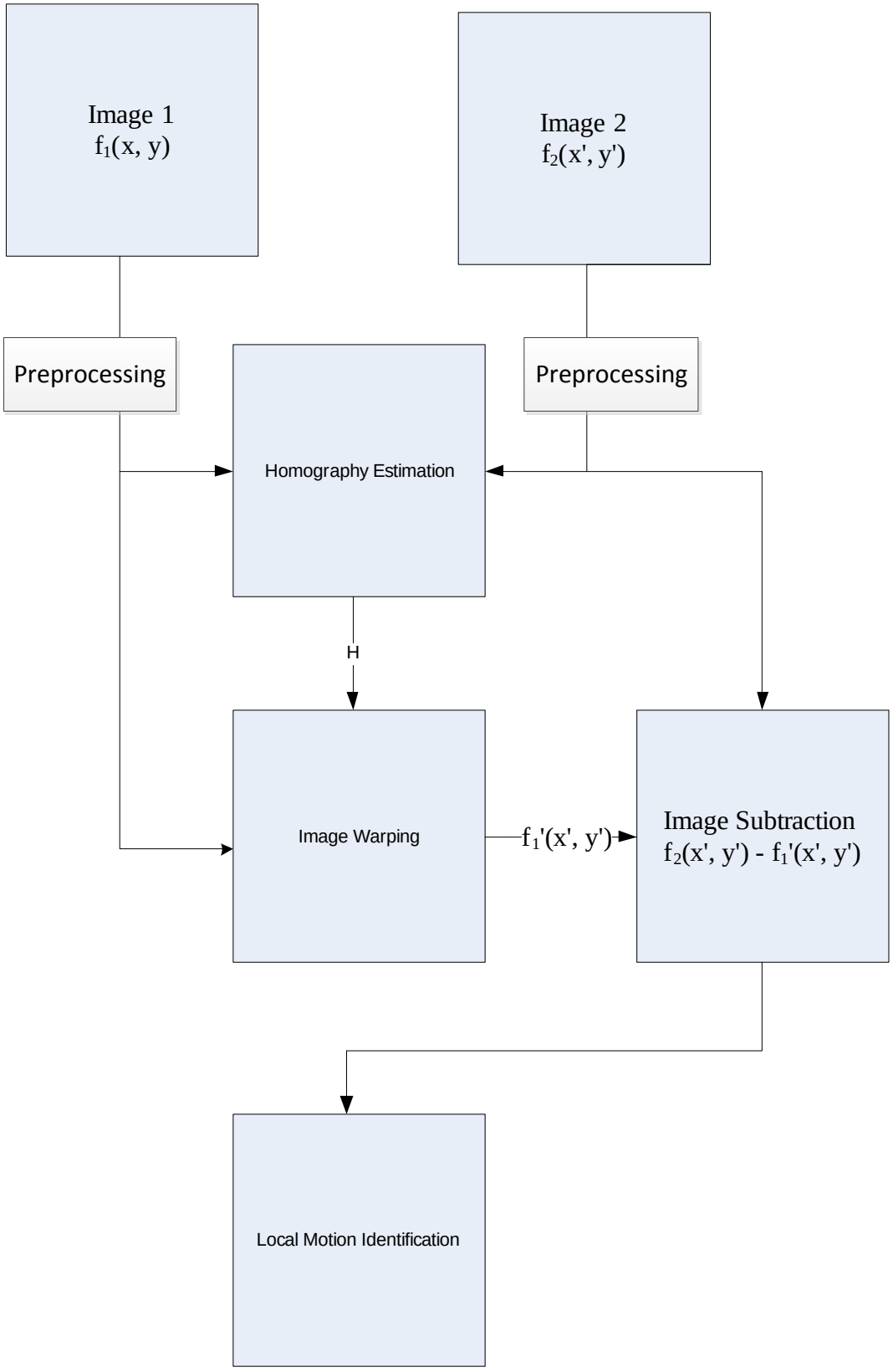


Figure 3–1 Proposed Algorithm Steps

a)

b)

Figure 3–2 Original two frames of a thermal airborne video, a car is seen as moving at the middle of the road, a) first frame $f_1'(x,y)$, b) second frame $f_2(x',y')$

3.1 Preprocessing

Processing of thermal videos has many difficulties compared to day-tv. One of the main difficulties arises from noise characteristics. Especially when derivative operations performed on a thermal image, noise corrupts the results significantly. Therefore, before processing the video, two step filtering is performed. To decrease salt-and pepper like noise, median filtering in 3-3 neighborhood is used. This filter assigns the value of a pixel to median value in defined window. Median filtering is good by means of preserving edges and corners compared to convolution with a smoothing mask. However, a Gaussian filter is still performed which has a variance of 0.5 and mask size of 7-7. Low variance and high mask size is chosen to ensure that corners which are used as features preserved at the end of the operation. Moreover, for interlaced videos, half sizing is performed as a first step.

(a)

(b)

Figure 3–3 Resulting images after preprocessing (a) first image, (b) second image

Effect of filtering variables (mask sizes of filter masks) on Harris performance is analyzed in [22]. They compare the repeatability of features on a publicly available data set and their experiments show that Harris corners' performance is higher for smaller masks (so less filtering) as given in Figure 3–4. Since thermal videos are generally noisier than day videos; we propose a little larger Gaussian mask.

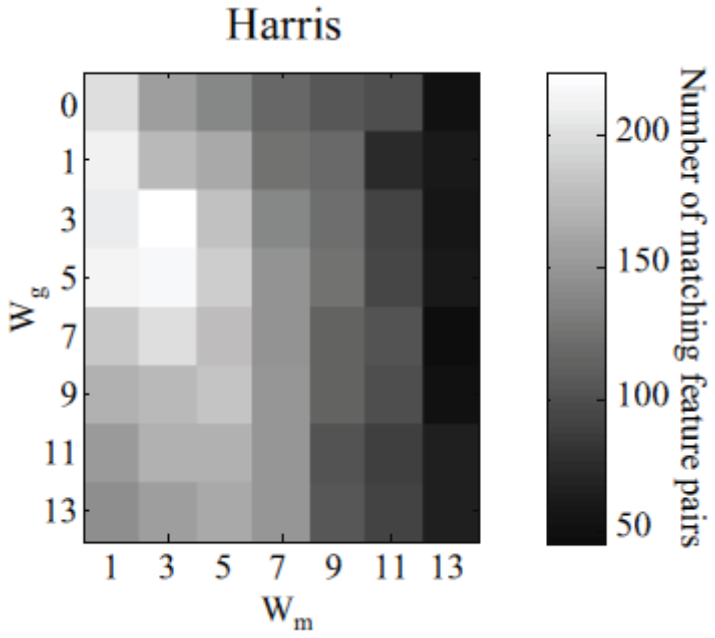


Figure 3–4 Repeatability of Harris corners, as the smoothing increases, repeatability decreases, W_g and W_m are dimensions of Gaussian and median masks, respectively [22].

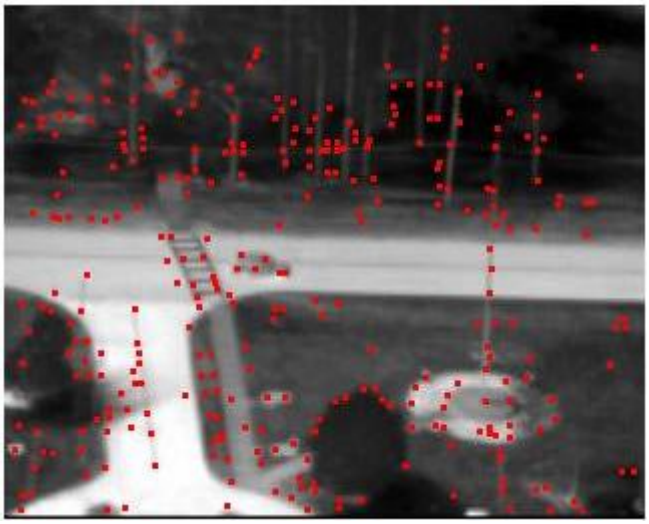
Since we use Harris corners, corner preserving filters may be used for performance improvements. Bilateral filtering [27] is frequently used as edge preserving smoothing filter. However aforesaid low size filters are shown to be successful on noise removal and feature preservation for IR airborne videos. Therefore, we only use median and Gaussian filtering.

3.2 Feature Detection

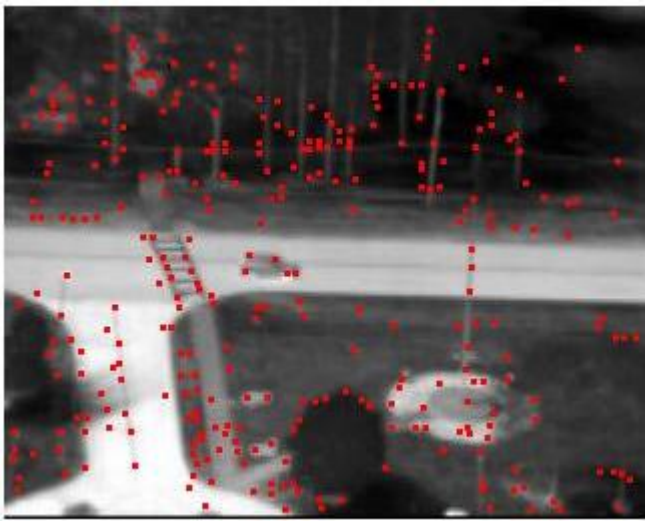
Scene motion in airborne videos necessitates global motion suppression which is done with registering consecutive frames. In order to warp two images, some features are identified. In literature, many kinds of features are selected for different purposes such as Harris, SIFT and SURF. For global motion suppression, a feature detected in first frame should also be detected in next frame at the same location. Moreover, many features should be detected easily by means of computational complexity. Since airborne videos are generally taken from very far distances, between two consecutive frames, rotational and scaling effects are not significant for

feature matching. The most important property that selected features should satisfy is localization and Harris corners have the best localization. Moreover, computational cost for detecting features is less for Harris corners. Therefore Harris corners are chosen as the key points to describe global motion. Performance evaluations between possible features are shown in Chapter 2 and Chapter 3 in detail.

For each image, M_C values are calculated by Equation (2–22) within a window of 7x7 where non-maximum suppression is performed in 5x5 neighborhoods. Best N points are selected as feature set after non-maximum suppression.



a)



(b)

Figure 3–5 Detected corners, a) first frame, b) second frame

Detected features are selected to extract complete scene motion. Therefore, distribution of selected features over frame is important. In some cases, local objects have many corners and most of the selected features are gathered on them. In order to ensure that feature set is uniform over image, Kirchof and Stilla divides each frame into M sub-regions and selects K features from all of them [1]. However, this forces bad feature collection and wrong matches especially at uniform regions. Instead of such an approach, constant number of features (N) selected after ordering

all corners from highest to lowest M_c values. As explained in part 3.4, some of the features are later eliminated. Empirically it is found that for frames where corners are rich, number of inliers do not exceed 200. Therefore N is selected as 320 which guarantee that good features are selected. In literature, it is generally assumed that 300 features are enough for most of the applications which is compatible with our selection [21].

3.3 Feature Matching Between Consecutive Frames

Since difference between consecutive frames can be due to both global motion and local motion, detected features are matched in order to fit global motion to a model. Since for airborne videos, rotational and scale changes are negligible for feature matching, normalized cross correlation method is used for match criteria [1]. Mean and variance of corners are measured in a 9-9 window and correlation value is estimated for each candidate pair. In Equation (2–31), summation is performed over 9-9 window same as given in ref. [1].

Finding correspondences for a particular feature in first frame can be done by calculating correlation coefficient for all features in second frame then choosing the highest value. For this application, since there are N features in each frame, there would be N comparisons for each feature in first frame. In total, N*N*2 comparisons are done and it is quite costly. However limiting the search area is possible if characteristics of usual airborne videos are considered. Since global motion is limited in most of the cases, for a particular feature in first frame at location (x_0, y_0), correlation coefficient is estimated with features in second frame which are located in a circle with a diameter of %10 diagonal of frame and centered at (x_0, y_0) [1]. Then, correlation coefficients are stored in a matrix and matching of two features is accepted if mutual consistency exists. That is if feature1 prefers feature2 and feature2 prefers feature1, a pair is found. Unmatched features are eliminated.

Since scene motion does not have to be at pixel resolution, sub-pixel localization of correspondences is important. Therefore, after finding rough locations of matches,

one more correlation is performed in one-tenth of a pixel resolution by using linear interpolation of neighboring pixels and correlating the blocks where features are at the centers. Blocks have size of 11x11 and sub-pixel locations are found at the points where correlation coefficient is the maximum.

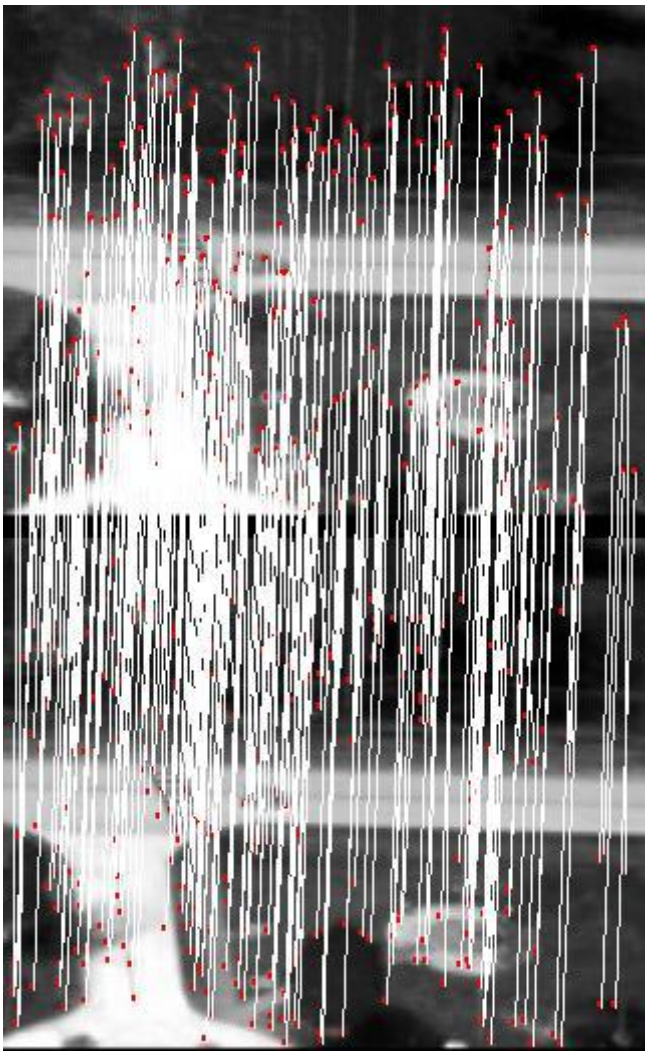


Figure 3–6 Matched features between images

3.4 Elimination of Outliers and Homography Estimation

In literature, there are several methods to extract global motion as explained in Chapter 3. Normalized DLT, geometric error minimization or geometric reprojection minimization work gently, if only error source for feature matching is measurement

noise. However, although most of the detected features are located on background, some of them are on moving targets. Moreover, due to thermal image properties and movement of camera, locating same feature at a wrong location in the next frame is possible. These matched feature pairs are called as outliers and eliminated before normalized DLT or aforesaid error minimization algorithms. Elimination of these outliers can be done by RANSAC. As explained in Chapter 3, RANSAC starts with random selection of minimal set members from matched pairs. Since four correspondences are enough for homography matrix calculation; minimal set is constructed from four pairs. Using this set, H is calculated with over determined solution and number of inliers obeying hypothesis of H is count.

Zisserman states that minimal set may have higher number of initial pairs. This would increase quality of registration due to the fact that constructed hypothesis from this minimal set get close to correct global motion. However, since there are wrong matches in matched pairs, probability of having wrong pairs in minimal set increases if the member of minimal set increases. Therefore, more iteration is performed according to Equation (2–42) which increases computational complexity significantly so it is not proposed to use larger minimal set [19]. Moreover, according to Equation (2–42), as the outliers increase, necessary iterations for good outlier elimination also increase which results in higher computational complexity. Therefore if we can eliminate some of the outliers prior to RANSAC, it would be possible to increase quality of registration as well as decrease computational complexity using larger minimal set and less iterations. In order to increase registration quality and decrease complexity, “2 Step RANSAC” is proposed. First step of this new recommended algorithm simply filters some bad matches by translational movement of pixels. Second step is conventional RANSAC with larger minimal set.

Although camera motion can be in any direction or rotational, translational movement of features are correlated significantly especially in sub-regions. Implemented algorithm takes advantage of this correlation. After N matched features are found, each image is divided into M sub-regions. In each region, translational

movements of features are measured and variance of feature movement σ= [ΔX_1 ΔY_1] is calculated. Then, “K-medians - RANSAC like algorithm” is performed.

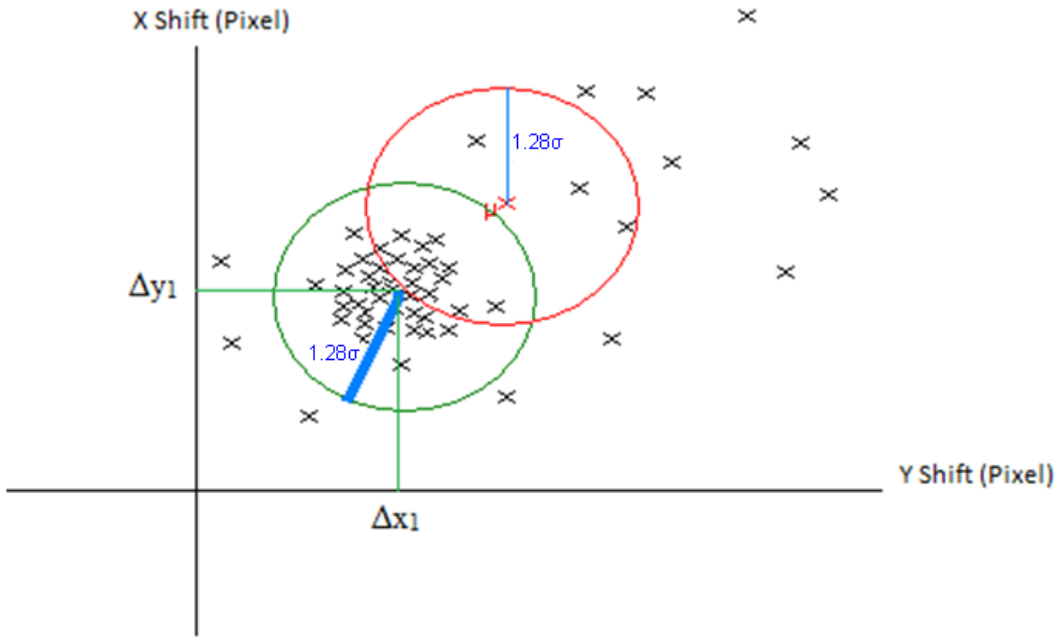


Figure 3–7 An example of translational movement of corners in a sub-region, the most dense region is intended to select as inliers

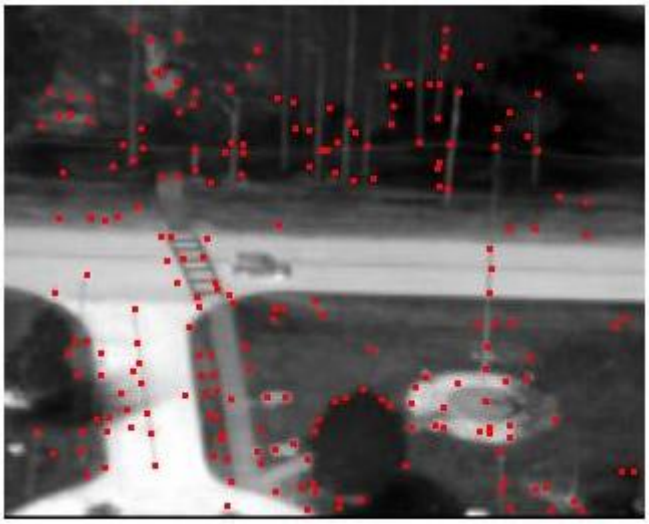
Assume that translational shift between matched pairs in a sub-region is mapped as in Figure 3–7. Translational movement for this sub-region is gathered around (Δx1, Δy1). However, mean may not be close to (Δx1, Δy1). Therefore, center of the densest region shall be found. This could be achieved by simple RANSAC algorithm assuming point hypothesis with a threshold of R. Point hypothesis can be defined with one-member initial set. Although computational complexity is quite low compared to conventional single step RANSAC, iteration number can be further decreased by one step median calculation at each iteration for only inliers. Here this approach is called as “K-medians - RANSAC like algorithm”.

“K-medians - RANSAC like algorithm” considers translational movement of two pixels as a point in 2-D. An initial point is selected as the center of a circle with radius R. Median of points inside circle is calculated and new circle is drawn around median. Then, number of inliers is count and first iteration of RANSAC is completed. With other random selections of initial point, RANSAC is performed for predefined iteration number. At the end, outliers are eliminated for the highest number of inliers set.

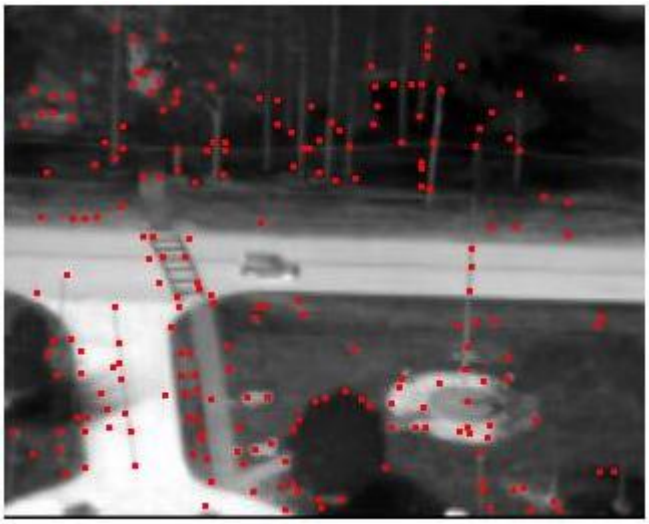
R defines filter characteristics. As the R increases, number of filtered bad matches decreases. As the R decreases, good matches may be filtered. Rather than constant setting of R, we set as 1.2σ empirically where σ is the variance of translation in corresponding sub-region. For the cases where rotational global motion is dominant, variance of the translations and so R is high which does not let elimination of most of the features. For Gaussian distribution of translation points, 20% of matches are eliminated.

From remaining inliers set, second step is performed. An initial set is chosen randomly with I members. In literature I is selected as four but since matched pair set is filtered, initial set can be enlarged. I is selected as eight at this algorithm. Then, homography matrix for eight pairs is calculated using normalized DLT. Number of inliers obeying initial hypothesis is count and more iteration is performed with other random selections of initial set. Finally, the lowest numbered outlier set is eliminated.

Number of members in initial set is defined experimentally for the best performance and its effect will be analyzed in Chapter 3. Moreover, normalized DLT can be replaced with other error minimization algorithms explained in Chapter 2. Reason of normalized DLT selection is discussed in Chapter 3 as well.



(a)



(b)

Figure 3–8 Remaining matched features, (a) first image, (b) second image

3.5 Analysis of Difference Image

Using all of the remaining inliers, final H homography matrix is calculated with normalized DLT algorithm. Then, first frame is warped into basis of second frame. Image warping is performed pixel by pixel using Equation (2–32). However, location

of a pixel at warped image may not be integer valued. Therefore, reverse mapping is preferred. Gray value of a pixel at warped image is found using linear interpolation of 4 neighboring pixels at first image by H^{-1}.

To find moving objects, difference operation is performed between warped image and second image. Then, binarization is performed on difference image with a threshold 0.11. Threshold setting is determined empirically and its effect on system performance is mentioned in Chapter 3.

$$difImage\ x',y'\ = f_1'\ x',y'\ - f_2\ x',y'$$

$$binImage = \begin{matrix} 1,\ if\ \ 0 \le difImage\ x',y'\ < T_B \\ -1,\ if\ \ -T_B \le difImage\ x',y'\ < 0 \\ 0,\ else \end{matrix} \qquad (3\text{–}1)$$

Since moving objects are assumed to be hot compared to background, two blobs are seen around moving objects, one with negative valued and positive valued. In the direction of movement, positive blob and at the back, negative blob appears. Distance between two blobs can be at maximum as much as velocity of the object or length of the object. Therefore, blobs that do not have a correspondence close to itself can be eliminated. To eliminate those false alarms, two images are formed, one with negative values and one with positive values. After binarization, positive and negative images are dilated with a circular morphological operator. Then, dilated positive difference image is multiplied by negative difference image pixel-by-pixel. Similarly, dilated negative difference image is multiplied by positive difference image. As in Figure 3–14, blobs which do not have a correspondence are eliminated.

The radius of mask determines the maximum reasonable speed or maximum vehicle length and it is obviously dependent on range of target and focus mode of camera. Therefore, radius of mask is updated in real-time from a predetermined table.

Figure 3–9 Binary positive difference image

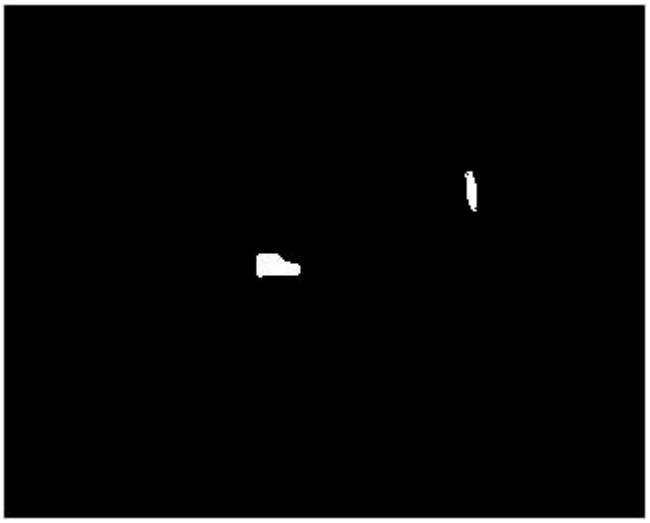


Figure 3–10 Binary negative difference image

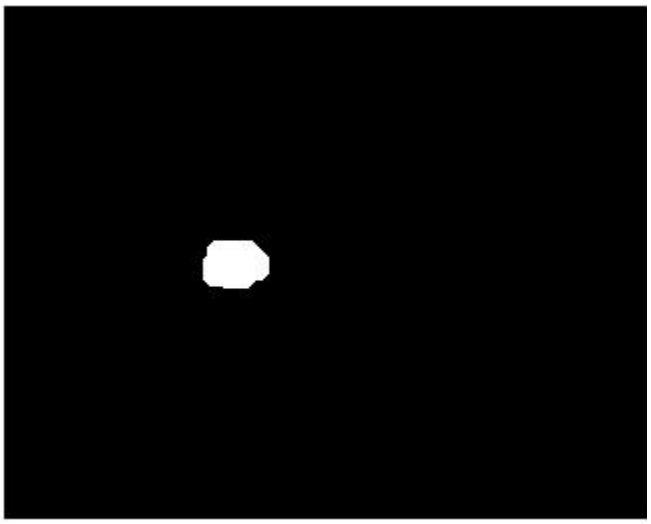


Figure 3–11 Dilated positive difference image

Figure 3–12 Dilated negative difference image

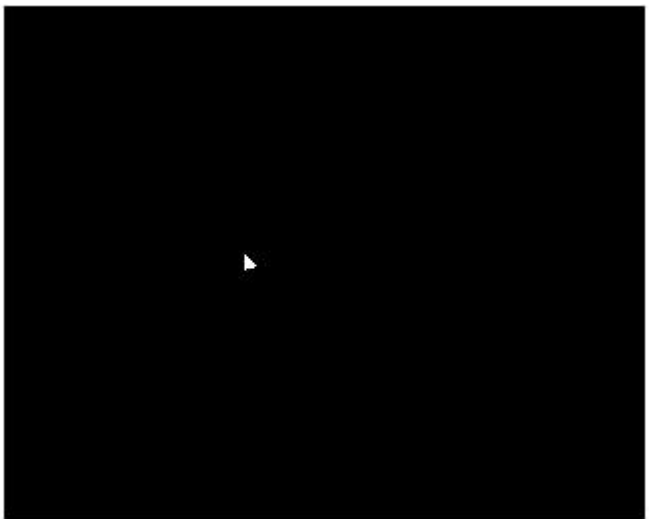


Figure 3–13 Result of a.*d where the false alarms are eliminated

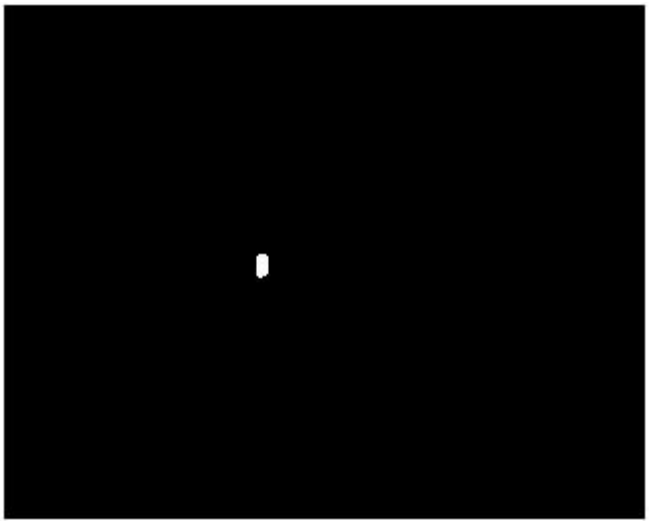


Figure 3–14 Result of b.*c where the false alarms are eliminated

At the end, final morphological operations, image opening and closing is applied to get rid of remaining false alarms resultant from sharp edges. Circular morphological operators are chosen to preserve blobs and eliminate edges. First, opening is performed with a circular operator of radius one then closing is performed with an operator of radius five. Finally, unified non-zero regions in final negative difference

image are marked as moving objects. Moreover, regions close to boundaries are also ignored due to bad registration quality.

Figure 3–15 Visual detection result

3.6 Experimental Results

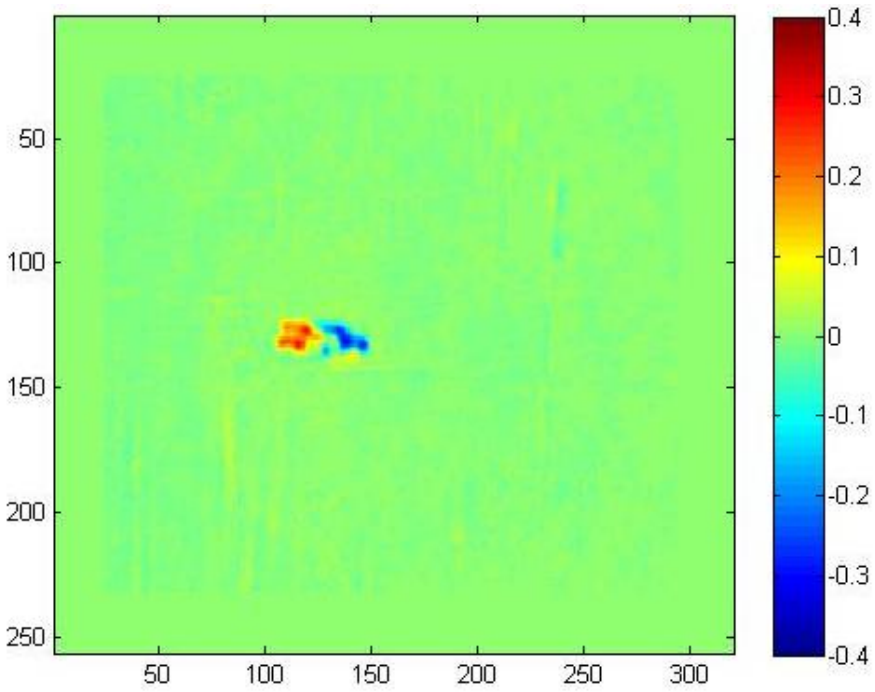
Proposed algorithm has many alternatives as explained in Chapter 2. We experiment the proposed algorithm for many different parameters and alternatives to demonstrate its performance. We have used IR video data set PkTest01 and PkTest02 in ref [30] for our experiments. Images are represented by only their intensities taking values from 0 to 1.

3.6.1 Comparison of Proposed Algorithm with Feature-Based Algorithms in Literature

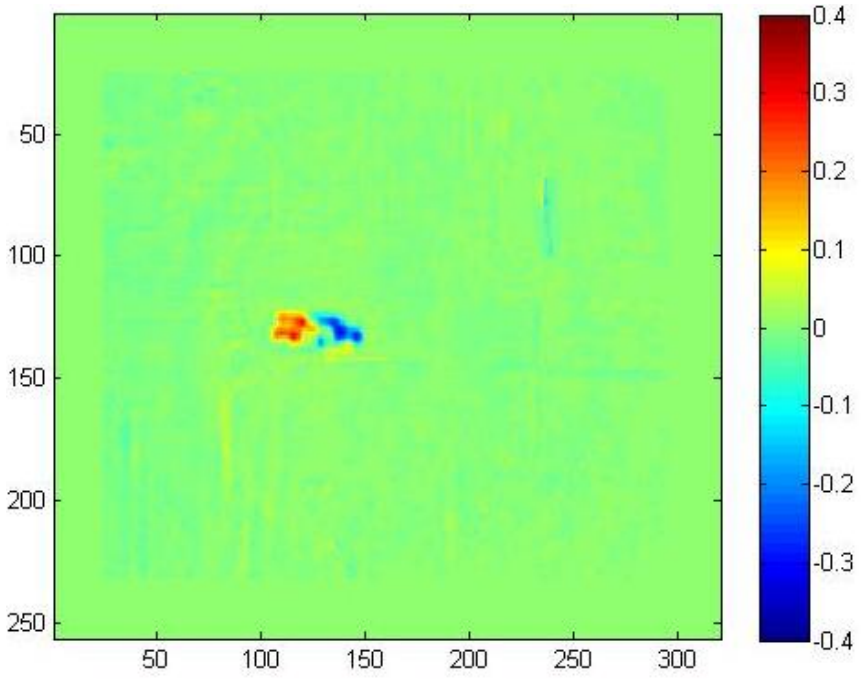
In feature-based algorithms, global motion model or transformation matrix between two images, $f_1(x, y)$ and $f_2(x, y)$ are extracted from detected and matched feature

points. In general, matched features set includes not only measurement noise, but also complete erroneous matches, called as outliers. Therefore, robust registration algorithms, LMedS or RANSAC is used in literature and proven to be successful [19]. LMedS and RANSAC are quite similar methods and result in almost same registration performance where outlier ratio is smaller than %50. If outliers have a higher portion with respect to all matched features, LMedS fails to fit a global motion model. Therefore, RANSAC has higher interest for global motion modeling.

As defined in Chapter 3, we propose a new approach to RANSAC, where some initial erroneous matches are filtered and conventional RANSAC is performed as a second step with larger minimal set. Here, we compare registration quality for the given two images in Figure 3–2 for proposed algorithm and conventional RANSAC. Iteration number for RANSAC is taken as 10. Blind regions close to boundaries and real vehicle motion region are excluded.



(a)



(b)

Figure 3–16 Difference images after registration, (a) result of proposed algorithm with 2-Step RANSAC (b) with conventional RANSAC

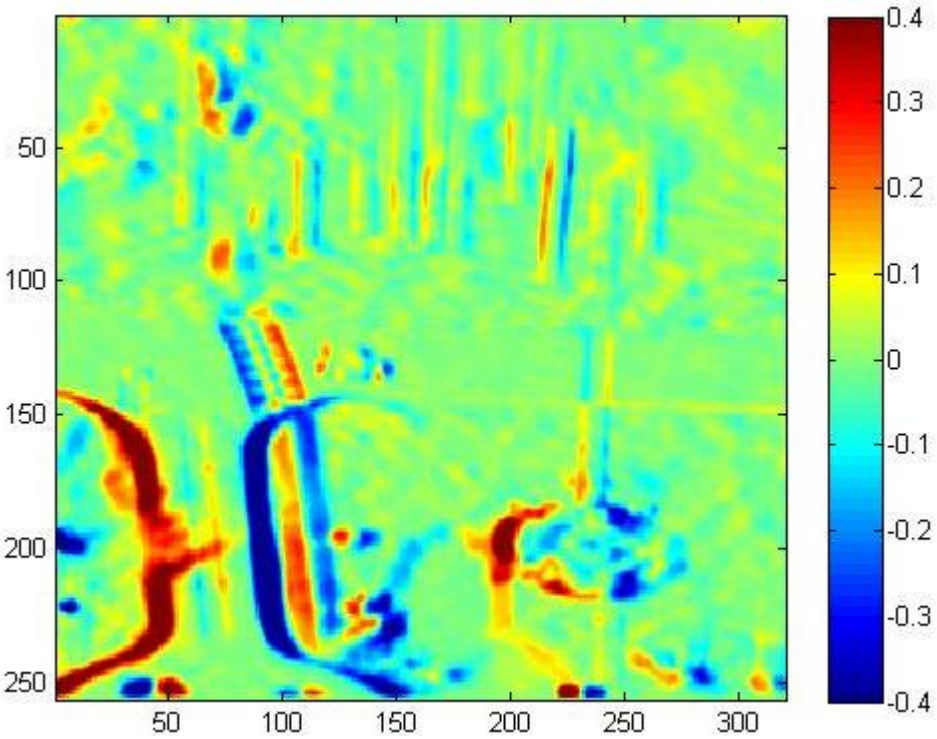


Figure 3–17 Difference of original images without registration.

$$f_1(x,y) \overset{Transformed}{} f'_1(x,y) \tag{3–2}$$

$$error\ x,y\ = abs\ \ f_2\ x,y\ -\ f'_1\ x,y\ \ \ .*\ mask(x,y) \tag{3–3}$$

$$errorRate = \frac{\sum_{x,y} error\ x,y\ *255}{pixelCount} \tag{3–4}$$

$$precisionRate = \frac{t_p}{t_p+f_p} \tag{3–5}$$

$$recallRate = \frac{t_p}{t_p+f_n} \tag{3–6}$$

Error rates to be compared algorithms are estimated by equations above where mask(x,y) composed of ones except in the region of moving vehicle and pixelCount is the total number of pixels in each image. For demonstration purposes, we show only one difference image resultant from registration of consecutive frames but error rates are given as the average of equally sampled 20 difference images in proceeding 400 frames. On the other hand, precision and recall rates are given for 400 frames where t_p, f_p and f_n stand for total number of true positives, false positives and false negatives, respectively. Local motion is detected between consecutive frames separated by four frames. Moreover we show computation time estimated during experiments. However, since we use MATLAB for all experiments, timing results may not be the same on real time systems. Using these criteria, results are found to be:

Table 3–1 Comparison of 2-Step RANSAC and conventional RANSAC

Error rate for proposed 2-Step RANSAC	0.9595
Error rate for conventional RANSAC [19]	1.1225
Precision rate for proposed 2-Step RANSAC	1
Precision rate for conventional RANSAC [19]	0.795
Recall rate for proposed 2-Step RANSAC	0.774
Recall rate for conventional RANSAC [19]	0.764

Therefore, in order to achieve desired registration performance, more iteration must be performed in conventional RANSAC. However, it results in increase in computational complexity. Computation times for different iteration numbers estimated in MATLAB environment are given Table 3–2.

Table 3–2 Analysis of RANSAC iteration number for computation time

Computation time for 10 iterations	176 seconds
Computation time for 30 iterations	180 seconds
Computation time for 50 iterations	185 seconds
Computation time for 100 iterations	193 seconds

These results prove that early filtering before RANSAC improves quality. There are two reasons. Firstly, iteration number kept as very low, namely 10. If the outlier ratio is high, necessary number of iterations to reach outlier-free minimal set increases by Equation (2–42). Secondly, proposed algorithm uses larger minimal set rather than four matched pairs which results in a better approximation of correct model at initial stage. However, increasing the number of initial set necessitates increase in iteration number. If algorithm is to run on a speed critical real time system, only filtering may be performed and using four matched pairs for initial set and less iteration (since outlier ratio is decreased), system load can be decreased.

Here, we also show the effect of selecting different initial sets after filtering the outliers at the first step and running 10 iterations. Corresponding error rates for difference images are given in Table 3–3.

Table 3–3 Error rates for different minimal sets

RANSAC with four members minimal set	1.0246
RANSAC with five members minimal set	1.0157
RANSAC with six members minimal set	1.0149
RANSAC with seven members minimal set	0.9945
RANSAC with eight members minimal set	0.9595
Larger minimal sets	slowly changes

As the number of minimal set increases, if minimal set does not include an outlier, quality of registration increases as well. However, for the given two images, minimal set with eight members are seem to be best. After eight, although quality of registration does not decrease, not significant improvements are obtained. One would expect decrease in quality, if the minimal set continued to enlarge because probability of outlier free minimal selection decreases. However, first step of filtering eliminates most of the worst outliers and remains with less outlier which are not as strong as previously filtered outliers. Therefore, still for larger minimal set, probability of selection of outlier-free minimal set is high. Even when a minimal set includes some outliers, they will not be dominant (inliers ratio is high and outliers are not very strong).

However, if the rotational movement is dominant, filtering of first part is expected to be poor. An example of such case is investigated below. Error rates are calculated from 10 difference images equally sampled in 100 frames.

(a)

(b)

Figure 3–18 Example of two consecutive images where rotation is dominant and a false alarm is seen, (a) first image, (b) second image

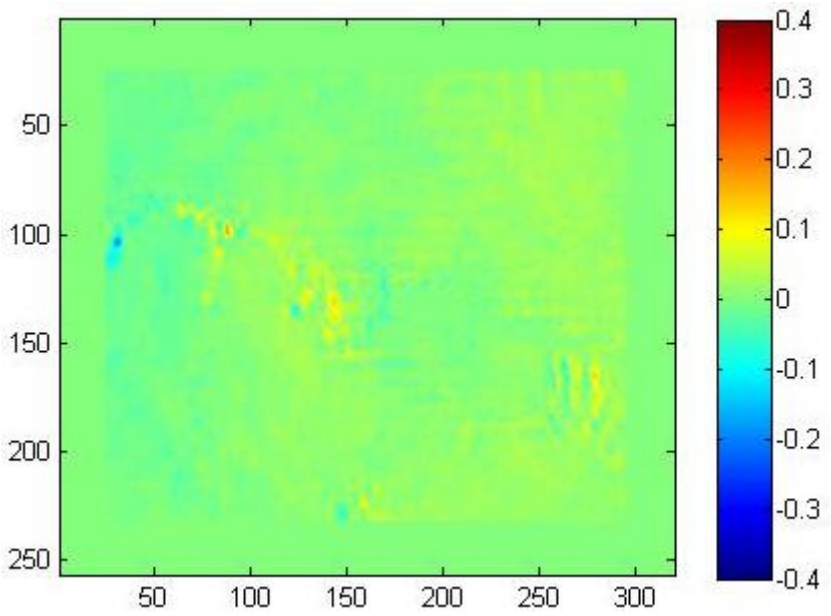


Figure 3–19 Difference image obtained by proposed algorithm, sharp edges give false alarm due to low binarization threshold, errorRate = 2.4737

In order to detect moving objects with low contrast difference compared to background as given in Figure 3–2, we set binary threshold for difference image to 0.11, which is very low. For the given threshold if rotational movement exits, registration may fail for very sharp edges.

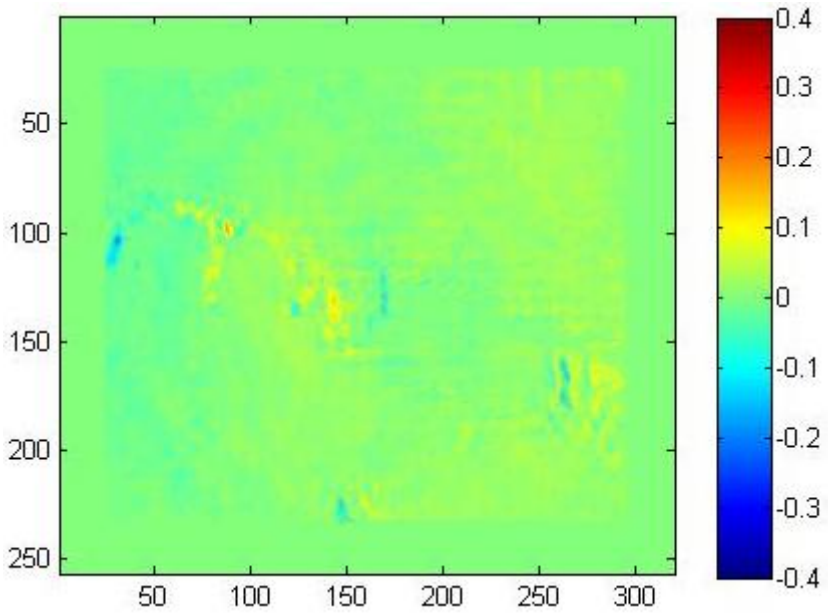


Figure 3–20 Difference image obtained by conventional RANSAC, sharp edges give false alarm due to low binarization threshold, errorRate = 2.5245

Figure 3–20 shows the result of conventional RANSAC algorithm which gives very similar results with our proposed algorithm. The reason that minor improvements are obtained is dominant camera motion is rotational. However, it still gives better result compared to conventional RANSAC in literature.

3.6.2 Comparison of Different Registration Algorithms

As discussed in Chapter 3, match features set includes outliers which shall be eliminated before error minimization algorithms and after elimination of outliers, one of previously described mean like optimization algorithms normalized DLT, geometric reprojection or geometric error minimization algorithms is selected. However, as stated in Chapter 2, other than normalized DLT, they are iterative algorithms and computationally complex. Therefore, we propose normalized DLT for final homography model fitting to global motion. However, we still perform

some experiments with geometric reprojection error minimization algorithm to compare registration quality of linear and iterative methods.

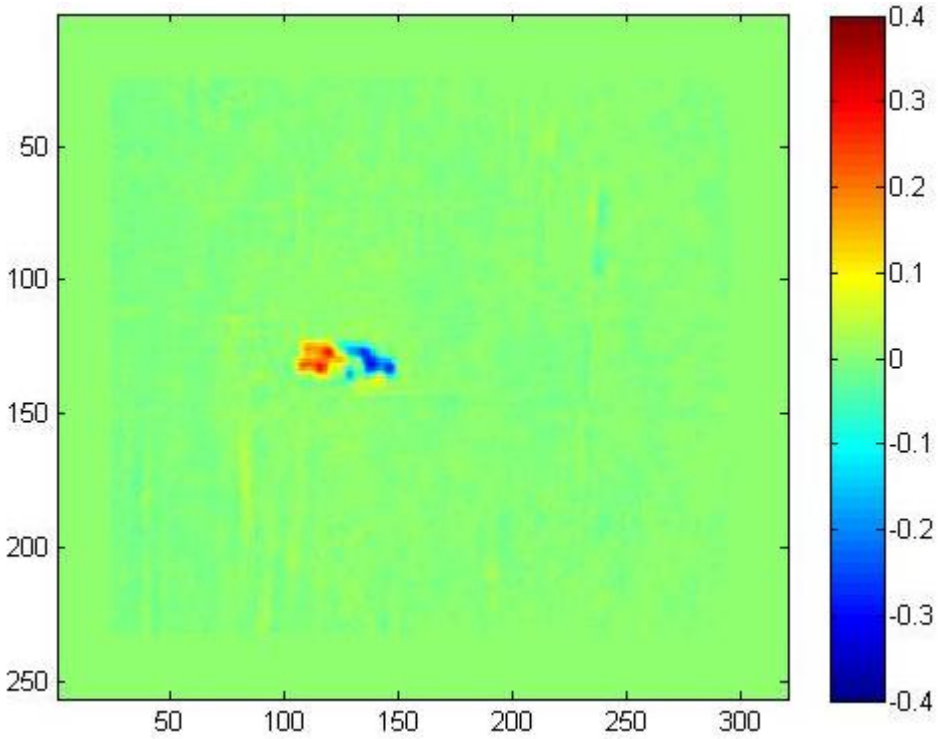


Figure 3–21 Geometric reprojection error minimization result, error rate = 0.9487

Figure 3–21 shows the difference image after warping with geometric reprojection error minimization with same inputs in Figure 3–2 and other parameters. Here error rate is found to be 0.9487 which is low compared to 0.9595 of normalized DLT and precision and recall rates are found to be same with normalized DLT. However, we lose more from computational efficiency to gain little performance improvement. Therefore, we do not propose iterative algorithms for error minimization.

We also test with examples given in Figure 3–18 and the resulting difference image is shown in Figure 3–22 where performance of registration is similar with normalized DLT.

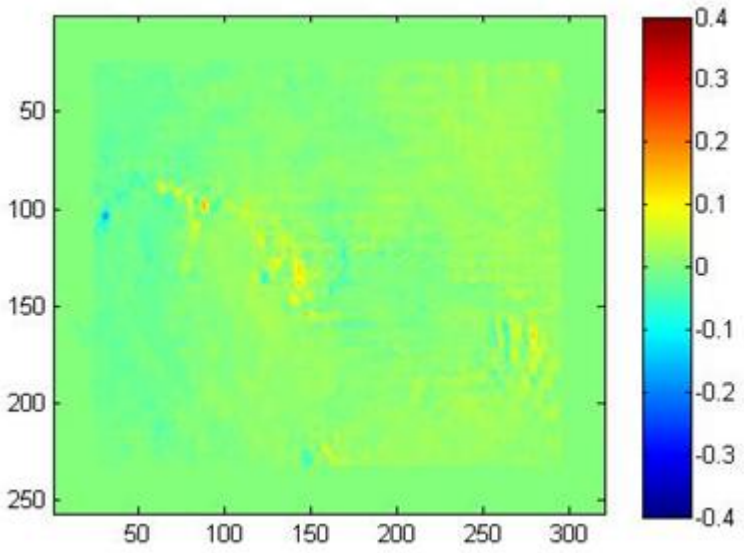


Figure 3–22 Geometric reprojection error minimization result, error rate = 2.4416 similar to normalized DLT

3.6.3 Comparison of Harris and SIFT for Registration

SIFT is proven to be very powerful and robust for feature detection and matching even when scaling, rotational and affine changes exist between images. However, it is computationally complex compared to Harris. We know that, between two frames of airborne videos, aforesaid corruptions are not as significant as translational changes. Therefore, we test these two features on usual thermal airborne videos to see decrease in quality of registration due to Harris selection. We used consecutive images given in Figure 3–18 to show performance of SIFT on registration quality.

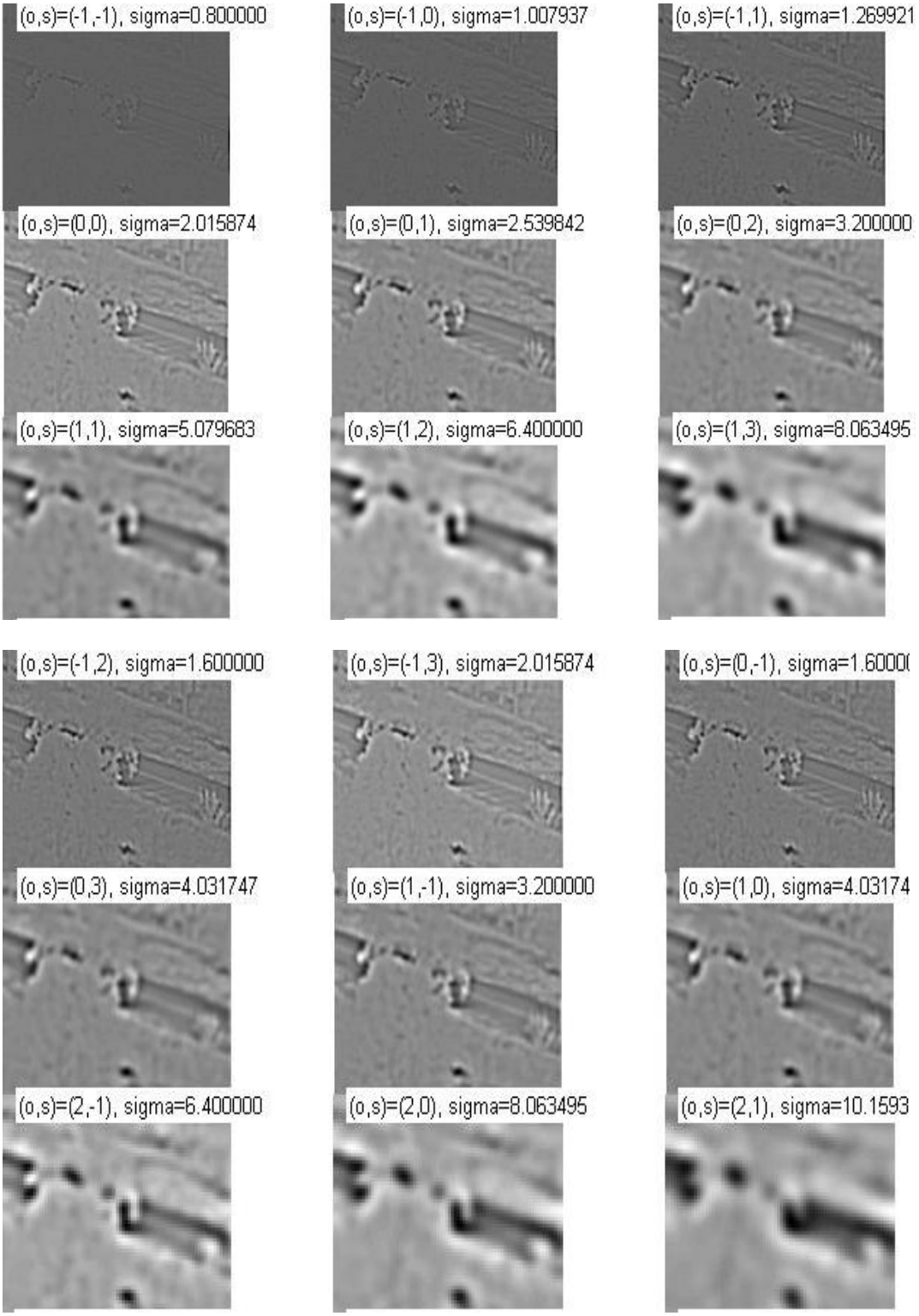


Figure 3–23 First image represented at different scales

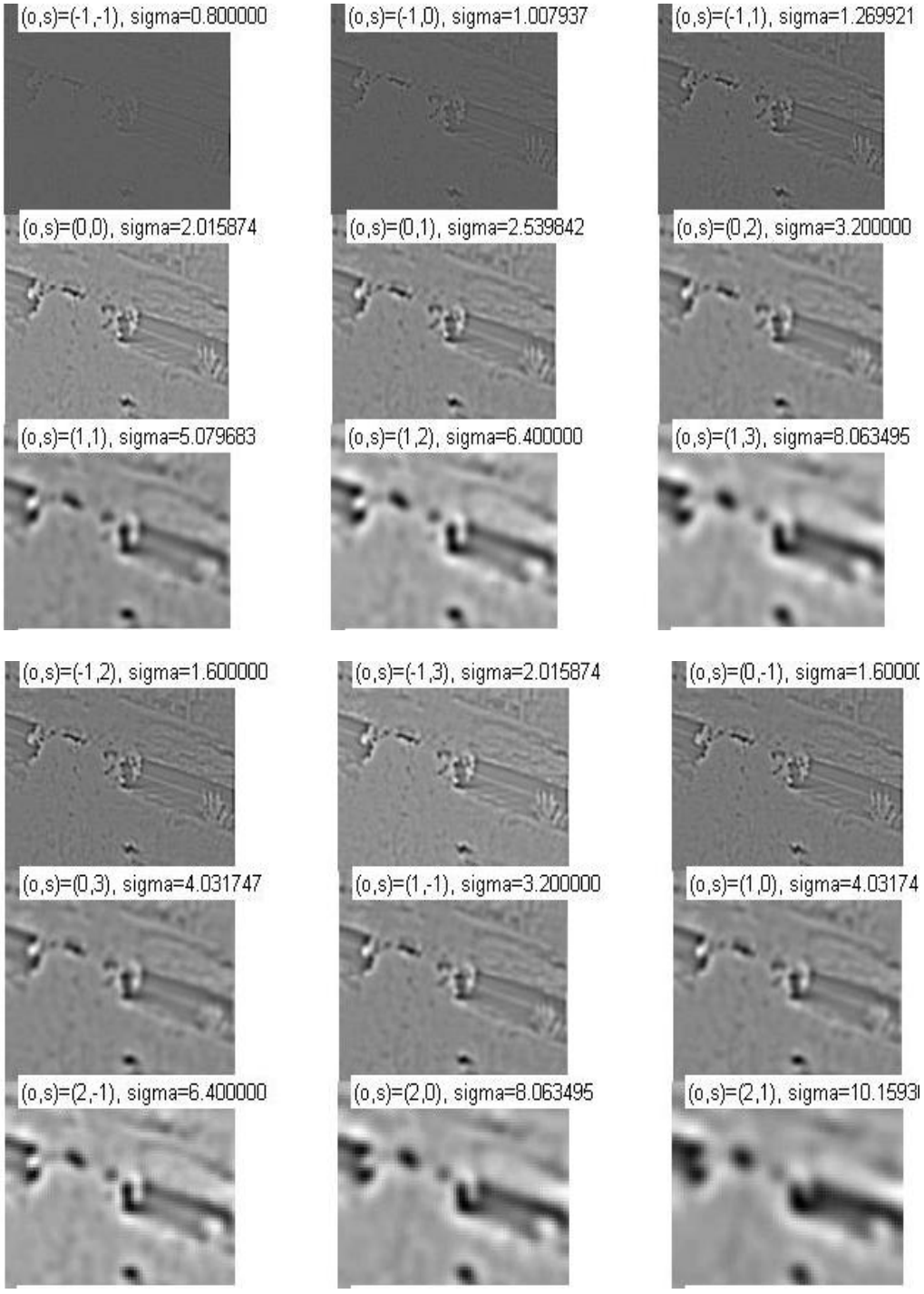


Figure 3–24 Second image represented at different scales

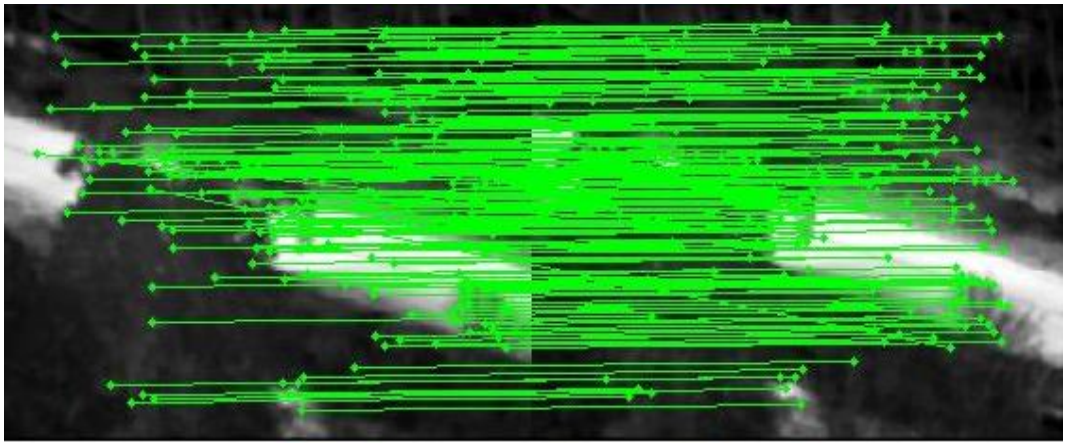


Figure 3–25 Matched features

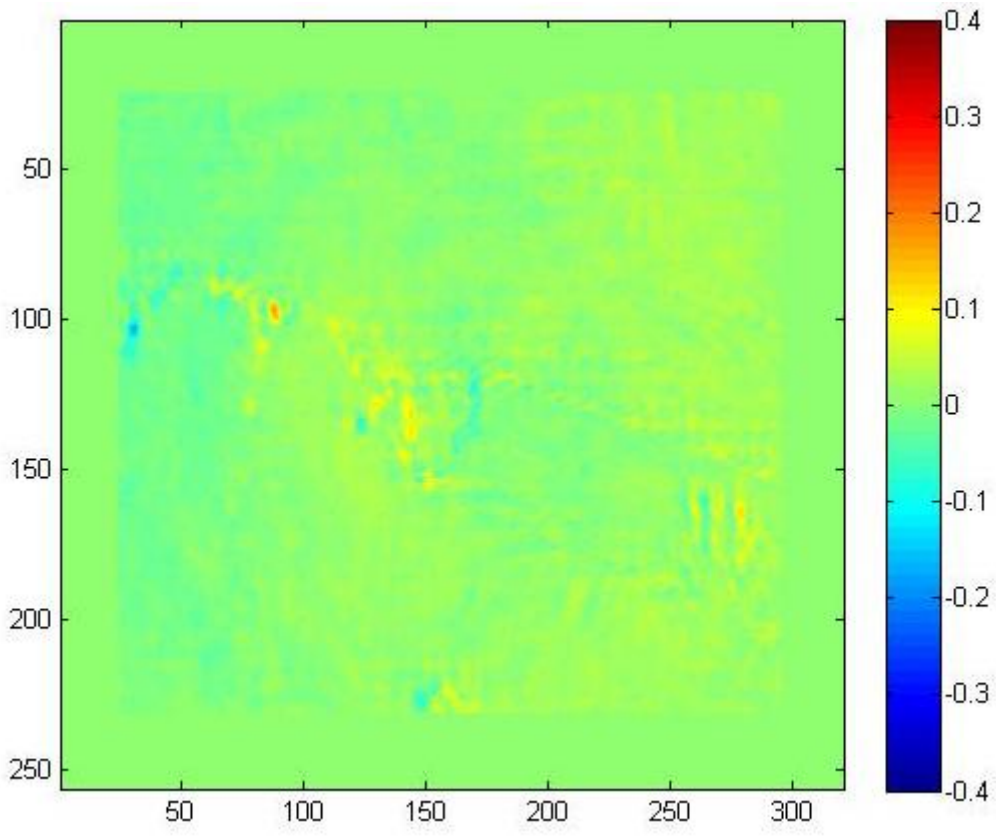


Figure 3–26 Difference image, error rate equals to 2.4035

Table 3–4 Comparison of Harris and SIFT

Error rate for Harris	2.4737
Error rate for SIFT	2.4035
Precision rate for Harris	1
Precision rate for SIFT	1
Recall rate for Harris	0.774
Recall rate for SIFT	0.784

Table 3–5 Comparison of Harris and SIFT for computation time

Computation time for Harris	176 seconds
Computation time for SIFT	805 seconds

SIFT extraction and matching is performed as explained in Chapter 2. SIFT detects 349 features on first image and 390 on second image. From those detected features, 261 are matched by Low's matching method. First step of outlier elimination of proposed algorithm (K-medians RANSAC like algorithm) eliminated five bad matches. Then, 256 features are left as inliers which are used for final homography estimation. Resultant difference image is given in Figure 3–26. Comparing visually with difference image result of Harris implementation given in Figure 3–19, there is little improvement in quality. Moreover, error, precision and recall rates are very

similar with Harris. However, Harris features selection and matching is faster than SIFT in several orders. Therefore, for time critical operations, Harris features are proposed.

3.6.4 Comparison of Proposed Algorithm and Intensity-Based Methods

In Chapter 2, we analyzed three intensity-based methods namely phase correlation, optical flow and iterative affine registrations. Phase correlation is fast but reliable only translational global motion exits. Therefore it is not considered in our evaluations. Optical flow estimation method, on the other hand, extracts optical flow vector for each pixel and decide on local motion regions which is quite costly. In conclusion, only affine global motion estimation is evaluated.

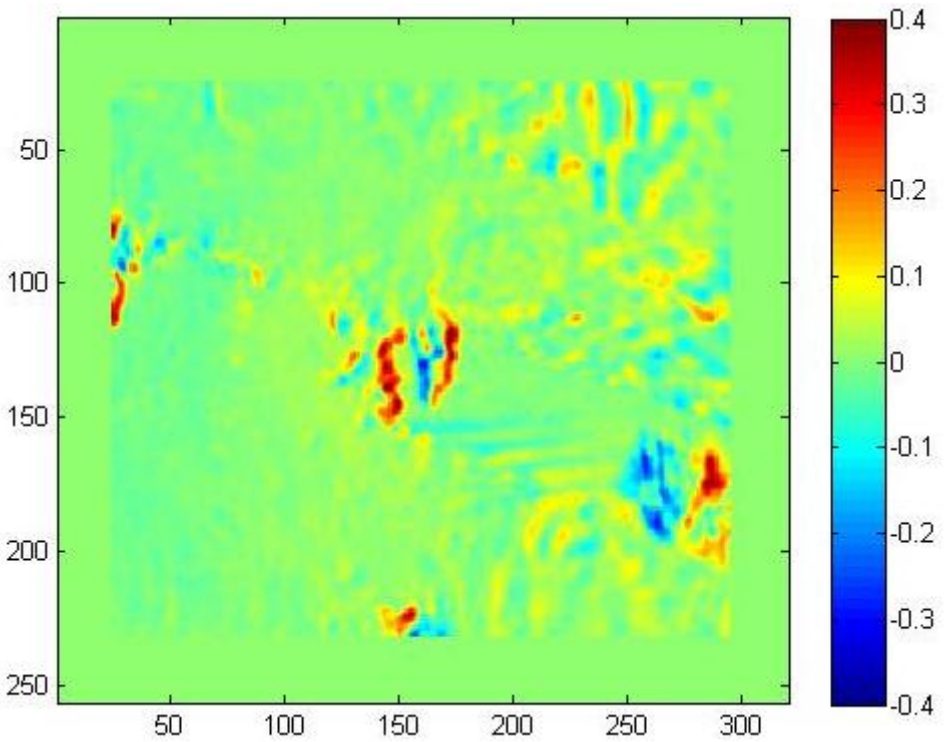


Figure 3–27 Difference image after registration, edges give false responses, error rate equals to 4.9215

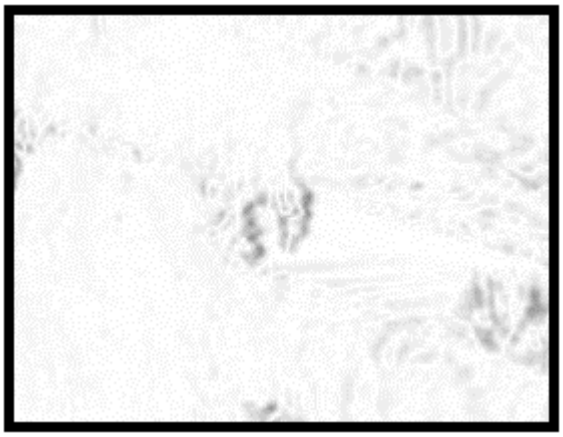


Figure 3–28 Absolute valued difference image, edges are not registered well.

Affine registration is performed as described in Chapter 2 in a pyramidal structure. Iterations are performed until θ in Equation (2–17) is slowly changing for all parameters (smaller than %1). Resulting registration quality was poor compared to feature-based proposed algorithm as expected. Moreover, complexity of iterative registration is far away to be implemented in a real time system.

3.6.5 Performance Evaluation of Proposed Algorithm for Different Image Sets

We test our proposed algorithm with different videos. Results of the experiments are given in the following equations.

(a)

(b)

Figure 3–29 Moving car is detected

(c)

(d)

Figure 3–29 Moving car is detected (cont'd)

(a)

(b)

Figure 3–30 Stationary vehicle is not detected

(c)

(d)

Figure 3–30 Stationary vehicle is not detected (cont'd)

(e)

Figure 3–30 Stationary vehicle is not detected (cont'd)

In Figure 3–30 stationary vehicle waiting in front of the traffic lights is not detected but moving vehicle is detected successfully.

(a)

(b)

Figure 3–31 An example of false alarm

Figure 3–31 shows an example of false alarm of the proposed algorithm. A stone is seen with very high contrast difference where registration of consecutive frames fails due to significant camera motion and very low threshold of binarization step. If we increase threshold to decrease false alarms, we would end up with reduction in true detection ratio.

Also example videos where features are poor worth to be experimented for proposed algorithm's performance evaluation. Here is two consecutive sample frames:

(a)

(b)

Figure 3–32 Examples of image pair where features are poor in the background, (a) first image, (b) second image

In Figure 3–32, a grass mover is seen on a green field. Without registration, direct differencing gives nice results because background is not textured and movement of camera does not result in significant background change. Moving vehicle can be detected by a simple difference operation on original frames.

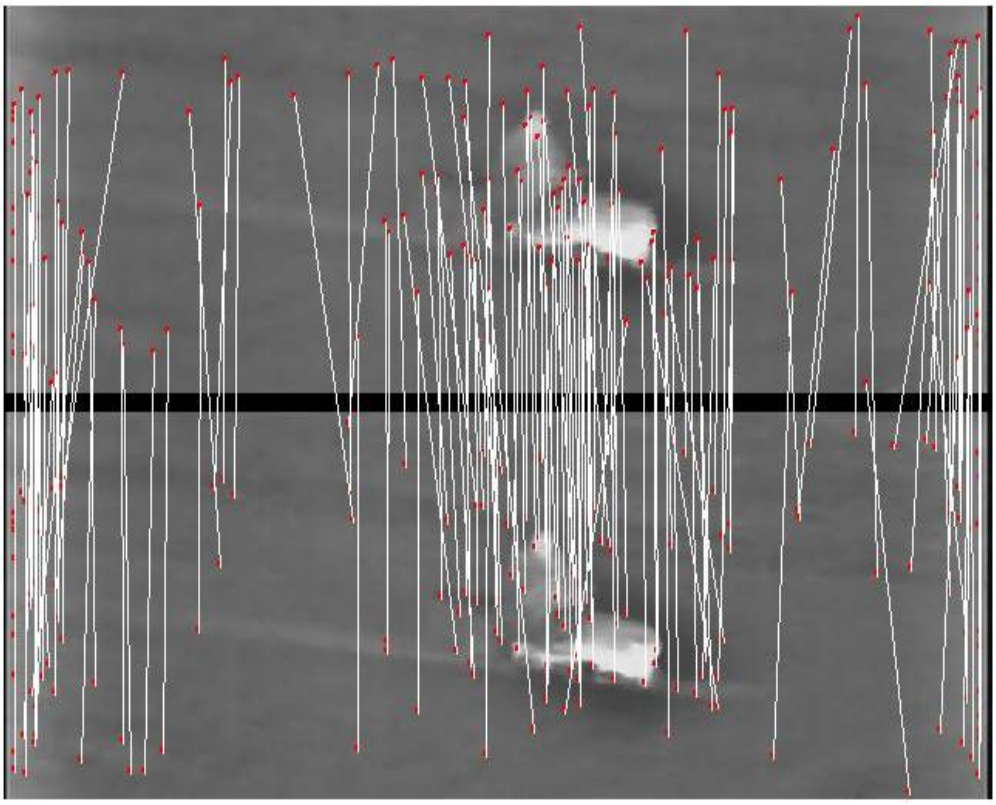


Figure 3–33 Matched features between frames, half of the features are matched

(a)

(b)

Figure 3–34 Inliers of images, (a) on first image, (b) on second image

As it is seen in Figure 3–34, very few inliers remain for image warping. The reason is that the background is almost uniform.

(a)

(b)

Figure 3–35 Absolute difference images, (a) on warped images, (b) on original images

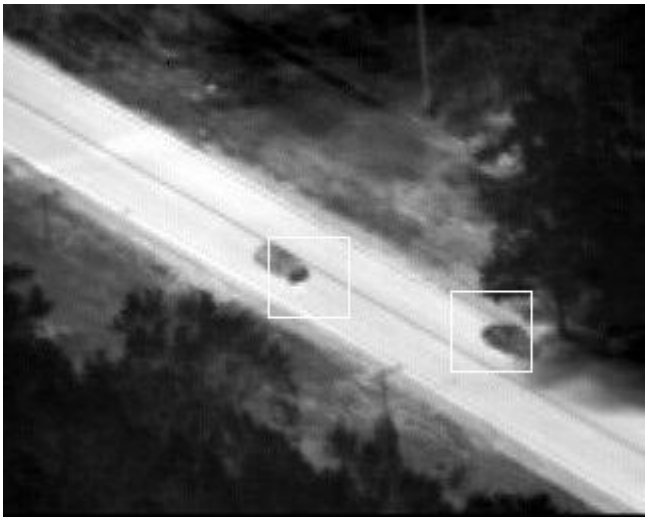
Although very few inliers (32) are used for registration, proposed algorithm increases the registration quality.

Results of the proposed algorithm for multiple moving objects scenarios are also shown below. We see that, multiple moving objects can be detected and tracked by our proposed algorithm as well.

(a)

(b)

Figure 3–36 Multiple moving vehicles example



(c)

Figure 3–36 Multiple moving vehicles example (cont'd)

(a)

(b)

Figure 3–37 One moving vehicle is detected as two vehicles

In Figure 3–37, each vehicles are considered as multiple moving objects. Main reason of that the distance between front and back wheels are very high and dominant contrast difference is resultant from wheel heat. A larger closing operator may be used when focus of the camera is high.

(a)

(b)

Figure 3–38 Four moving vehicles example

(c)

(d)

Figure 3–38 Four moving vehicles example (cont'd)

(e)

(f)

Figure 3–38 Four moving vehicles example (cont'd)

(g)

Figure 3–38 Four moving vehicles example (cont'd)

(a)

(b)

Figure 3–39 Detection result under occlusion

(c)

(d)

Figure 3–39 Detection result under occlusion (cont'd)

(e)

(f)

Figure 3–39 Detection result under occlusion (cont'd)

CHAPTER 4

CONCLUSION

4.1 Summary and Conclusion

The primary difficulty in moving object detection by moving IR cameras is the existence of global motion resultant from sensor motion. In this paper, we analyze several approaches to model global motion and propose a complete local motion detection algorithm with some improvements.

We group global motion modeling algorithms in two parts: 1) Intensity-based and 2) Feature-based. Other than phase correlation, intensity-based algorithms are computationally quite costly for real time systems. Phase correlation algorithm; on the other hand, gives promising results only if global and local motion are translational, which is not the case for most of the airborne thermal videos. Therefore, we focus on feature-based methods in our works.

Feature-based algorithms extract some key points over frames and match with a similarity measure. Then motion of features is exploited to extract global motion model. There are many features and matching algorithms. SIFT, SURF and ORB are very popular because of their invariance to scale, rotation, translation, affine changes and noise. Researches over state-of-the-art techniques show that SIFT is the most powerful by means of robustness and correctness. However, computational complexity of SIFT is high which makes it difficult to be implemented on a fast response, real time systems. On the other hand, Harris corner extraction and matching is very fast compared to SIFT. We analyze and show that for usual thermal

airborne videos, performance of Harris corners are slightly less than SIFT. Therefore Harris corners are selected as features.

We model global motion with planar homography which has 8 degrees of freedom. Optimum global motion parameters are found by over determined solution of matched features. However, since outliers bias the result wrongly, we first eliminate outliers than then use normalized DLT. Elimination of outliers is done by 2-Step RANSAC where first step considers dominant motion as translational in sub-regions. Second step is conventional RANSAC with larger minimal set. First step of the outlier elimination is very fast compared to second part and eliminates strongest outliers. This results in possibility of decreasing iteration number of second step of RANSAC which decreases computational complexity. Moreover, larger minimal set selection results in quality increase of image warping which is possible with less outlier.

Finally, homography matrix is used to warp first image into second (base) image and difference between base image and warped image is obtained. Since moving objects are hot compared to their environments, they are detected as two blobs, one positive and one negative. In order to eliminate false matches, blobs which do not have an anti-blob close to itself are eliminated. After that, morphological operations are performed to decrease point errors and at the end non-zero regions are identified as moving objects.

We show at Chapter 3 and 4 that our proposed algorithm successfully detects motion even when contrast difference of moving vehicle with respect to its surrounding is not very high. This is provided by low threshold selection of binarization of difference image. However, under strong affine camera motions, sharp edges may give false alarms due to low threshold. Moreover, our proposed algorithm relies on the planar surface assumption which is feasible where camera is located at far distances from scene. At close viewing situations, false alarms may be seen.

To conclude, we propose a moving hot object detection algorithm for airborne thermal videos considering the limited computation capacity of real time systems which successfully work for most of the experimented scenarios.

4.2 Future Work

Image warping quality between consecutive frames significantly affects the output of proposed algorithm. Therefore, improvements in image warping will be further investigated. Rather than homography estimation between only two frames, more frames can be used for error minimization. Moreover, most airborne platforms have systems to measure movements of camera and platform that can be used as input to system for increasing image warping quality.

Moving vehicles generally moves in some routines therefore tracking of motion and comparing the route changes can be further used for false alarm elimination. Motion identified regions which abruptly change in direction and speed can be ignored.

Size filters can be also used for false alarm reduction [29]. However, size of usual moving objects obviously depend on camera focus, target range and object type. Therefore, adaptive size filtering will be further investigated after enough data is collected from real IR videos.

Detected moving objects are generally vehicles or humans and at the local motion detected regions, some features can be extracted and tested by a classifier. Aggarwal uses Bayesian classifier where he compares four features of mean, variance, skewness and kurtosis [2].

Calculating the optical flow for every pixel is quite costly for global motion suppression. However, after local motion identification by our proposed algorithm, optical flow for local motion detected regions can be created and used for false alarm reduction. Optical flow in the area of local motion should be similar in magnitude and direction and regions that show incoherent optical flow may be eliminated.

Moreover, different ways to decrease computational complexity will be analyzed. At the first step of RANSAC, angular and translational velocity and acceleration data of the camera and platform, obtained by sensors of the airborne systems may be used for a better filtering. This will decrease outliers' ratio and necessary iterations. Furthermore, different features and matching algorithms are worth to investigate to decrease system load.

REFERENCES

[1] Kirchhof M., Stilla U., “Detection of moving objects in airborne thermal videos”, ISPRS Journal of Photogrammetry and Remote Sensing, pp. 187-196, 2006.

[2] Strehl, A., Aggarwal, J.K., “A motion-based object detection and pose estimation method or airborne FLIR sequences”, Machine Vision and Applications, pp. 267–276, 2000.

[3] C. Harris and M.J. Stephens, “A combined corner and edge detector”, Alvey Vision Conference, pp. 147–152, 1988.

[4] E. Rublee, V. Rabaud, K. Konolige, G. Bradski, “ORB: An Efficient Alternative to SIFT or SURF”, IEEE In International Conference on Computer Vision, Barcelona, Spain, 2011.

[5] M.A. Fishler and R.C. Bolles,“Randon sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography”, Comm. ACM, vol. 24, pp. 381-395, 1981.

[6] C. Tomasi and T. Kanade, “Detection and tracking of point features technical report”, Carnegie. Mellon University Technical Report CMU-CS-91-132, Apr. 1991.

[7] Davis, J., “Mosaics of scenes with moving objects”, Proceedings IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR98), pp. 354–360., 1998.

[8] Woelk, F., Koch, R., “Fast monocular Bayesian detection of independently moving objects by a moving observer”, Proceedings of DAGM Symposium. Lecture Notes in Computer Science, pp. 27–35, 2004.

[9] J. Shi and C. Tomasi, “Good features to track", Proc. IEEE Comput. Soc. Conf. Comput. Vision and Pattern Recogn., pp. 593-600, 1994.

[10] D.G. Lowe, “Distinctive image features from scale invariant key points”, International Journal of Computer Vision, vol. 33, no. 2, pp. 91-110, 2004.

[11] H. Bay, T. Tuytelaars, and L. Van Gool. “Surf: Speeded up robust features”, In European Conference on Computer Vision, 2006.

[12] Yilmaz, A., Javed, O., and Shah, M. 2006. “Object tracking: A survey”, ACM Comput. Surv., pp. 1-10, 2006.

[13] Mikolajczyk, K., Schmid, C., “An affine invariant interest point detector”, In European Conference on Computer Vision (ECCV), vol. 1, pp. 128–142., 2002.

[14] E. Rosten and T. Drummond, Machine learning for high-speed corner detection, In European Conference on Computer Vision, vol 1, 2006.

[15] P. L. Rosin. “Measuring corner properties”, Computer Vision and Image Understanding, pp. 291 – 307, 1999.

[16] M. Calonder, V. Lepetit, C. Strecha, P. Fua., “Brief: Binary robust independent elementary features”, In European Conference on Computer Vision, 2010.

[17] J. M. Geusebroel, G. J. Burghouts, A. W. M., “The Amsterdam Library of Object Images”, International Journal of Computer Vision, pp. 103– 112, 2005.

[18] C. Bulla., “Local Features for Object Recognition, Instutue of Comminication Engineering”, Aachen University, 2012.

[19] Hartley, R., Zisserman, A., “Multiple view geometry”, Cambridge University Press, Cambridge, 2nd edition, pp. 87-132 UK., 2003.

[20] Quénot G.M., Pakleza J., Kowalewski T.A., “Particle image velocimetry with optical flow, experiments in fluids”, vol. 25, no. 3, pp. 177-189, 1998.

[21] Bauer, J., Sunderhauf, N., & Protzel, P., “Comparing several implementations of two recently published feature detectors”, In Proc. of the International Conference on Intelligent and Autonomous Systems, IAV, Toulouse, France., 2007.

[22] Blanco J. L.,Gonzalez J., “An experimental comparison of image feature detectors and descriptors applied to grid map matching”, University of Malaga, Spain, 2010.

[23] Lazebnic, S., "Local, semi-local and global models for texture, object and scene recognition", PhD Thesis, University of Illinois, pp. 145-151, 2006.

[24] H. Moravec, "Obstacle avoidance and navigation in the real world by a seeing robot rover", CarnegieAMellon University Robotics Institute Technical Report,, Sep. 1980.

[25] Irani, M. , Anandan, P. , "A Unified Approach to Moving Object Detection in 2D and 3D Scenes", IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 20, No. 6, June 1998.

[26] Sırtkaya, S., "Moving Object Detection In 2D and 3D Scenes", MS Thesis, Middle East Technical University, 2004.

[27] Qian, G., Chellappa, R. , "Moving Targets Detection using Sequential Importance Sampling", Proc. of Int'l Conf. on Computer Vision, Vol 2, pp. 614-621, July 2001.

[28] C. Tomasi and R. Manduchi, "Bilateral filtering for gray and color images," IEEE International Conference on Computer Vision, pp. 839–846, 1998.

[29] Jain, R. , Kasturi, R. , G.Schunk, B. , "Machine Vision", McGRAW-HILL, International Editions, 1995.

[30] Collins, R. T, Zhou X., Teh S. K., "An Open Source Tracking Testbed and Evaluation Web Site", IEEE International Workshop on Performance Evaluation of Tracking and Surveillance, January, 2005.

[31] Yu, Q., Medioni G., "A GPU-based implementation of Motion Detection from a Moving Platform", IEEE Workshop on Computer Vision on GPU, 2008.