DYNAMIC-BASED VIDEO DATA REGISTRATION

A Thesis Presented
By
Bo Li
to
The Department of Electrical and Computer Engineering
in partial fulfillment of the requirements
for the degree of
Master of Science
in
Electrical and Computer Engineering
in the field of
Control Systems and Signal Processing

Northeastern University
Boston, Massachusetts
April 2010
NORTHEASTERN UNIVERSITY
GRADUATE SCHOOL OF ENGINEERING

Date: April 2010

Author: Bo Li
Title: Dynamic-based Video Data Registration
Department: Electrical and Computer Engineering
Degree: M.Sc. Convocation: April Year: 2010

Permission is herewith granted to Northeastern University Graduate School of Engineering to circulate and to have copied for non-commercial purposes, at its discretion, the above title upon the request of individuals or institutions.

Signature of Author

THE AUTHOR RESERVES OTHER PUBLICATION RIGHTS, AND NEITHER THE THESIS NOR EXTENSIVE EXTRACTS FROM IT MAY BE PRINTED OR OTHERWISE REPRODUCED WITHOUT THE AUTHOR’S WRITTEN PERMISSION.

THE AUTHOR ATTESTS THAT PERMISSION HAS BEEN OBTAINED FOR THE USE OF ANY COPYRIGHTED MATERIAL APPEARING IN THIS THESIS (OTHER THAN BRIEF EXCERPTS REQUIRING ONLY PROPER ACKNOWLEDGEMENT IN SCHOLARLY WRITING) AND THAT ALL SUCH USE IS CLEARLY ACKNOWLEDGED.
# Table of Contents

Table of Contents ................................................................. iii

List of Tables ................................................................. v

List of Figures ................................................................. vi

Abstract ................................................................................ viii

Acknowledgements ............................................................... ix

1 Introduction ........................................................................ 1
   1.1 Background ..................................................................... 1
   1.2 System Overview ....................................................... 3
   1.3 Thesis Organization ................................................... 5

2 Scale Invariant Feature Transform ........................................... 8
   2.1 SIFT Algorithm ........................................................ 8
      2.1.1 Scale-space Peak Selection ...................................... 9
      2.1.2 Keypoint Localization .......................................... 10
      2.1.3 Orientation Assignment ....................................... 11
      2.1.4 Keypoint Descriptor ........................................... 12
   2.2 Harris-SIFT ............................................................... 13

3 Plane Induced Homography Given Fundamental Matrix .......... 15
   3.1 Projective Geometry ................................................... 16
   3.2 Epipolar Geometry ..................................................... 17
   3.3 Plane Induced Homography Given Fundamental Matrix And Line Correspondences .................................................... 19
List of Tables

<table>
<thead>
<tr>
<th>Table</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>5.1</td>
<td>Algorithm 1: register to map image (synthetic data)</td>
<td>29</td>
</tr>
<tr>
<td>5.2</td>
<td>Quantitative inspection of Hankel RM</td>
<td>31</td>
</tr>
<tr>
<td>5.3</td>
<td>Algorithm 2: register to map image (real data) (Harris-SIFT)</td>
<td>35</td>
</tr>
<tr>
<td>5.4</td>
<td>Algorithm 3: plane induced homography</td>
<td>43</td>
</tr>
</tbody>
</table>
## List of Figures

1.1 The overview of system equipments .......................... 6
1.2 Flow chart of the algorithm used in this thesis .............. 7

2.1 The construction of Difference of Gaussian (DOG) .......... 10
2.2 Maxima and minima of the difference-of-Gaussian images detection . 11
2.3 The SIFT keypoint descriptor ................................. 13
2.4 An illustration of Harris-SIFT algorithm ....................... 14

3.1 An illustration of the projective transformation ............... 17
3.2 Point correspondence geometry. ............................... 18
3.3 The pencil of planes defined by a line correspondence .......... 21

5.1 Synthetic data in the lab .................................. 28
5.2 Correspondences found by Affine-SIFT ....................... 30
5.3 Overlay the transformed vehicle image on the satellite image. .... 30
5.4 The performance of Hankel rank minimization .................. 32
5.5 Compare mapping result of with and without Hankel ............ 32
5.6 Roughly rectification and registration with and without RANSAC .. 33
5.7 Real road data ............................................ 38
5.8 SIFT is stuck in the real world data ........................... 39
5.9 The plots of the SIFT feature descriptors ....................... 39
5.10 Harris corner detection on the satellite and pseudo-satellite image . . 40
5.11 The illustration of SIFT descriptors . . . . . . . . . . . . . . . . . . . . . 40
5.12 Examples of Harris-SIFT results . . . . . . . . . . . . . . . . . . . . . . 41
5.13 The result of Harris-SIFT . . . . . . . . . . . . . . . . . . . . . . . . . . . 42
5.14 The estimation of fundamental matrix by RANSAC . . . . . . . . . . 43
5.15 Line detection by Hough transform . . . . . . . . . . . . . . . . . . . . . 45
5.16 The illustration of Hough transform . . . . . . . . . . . . . . . . . . . . . 46
5.17 The mapping result of the plane induced homography algorithm . . 47
5.18 Compute the distance between the camera and features . . . . . . . . 49
Abstract

Knowing the precise location where data is collected is a key feature for automated road inspection, including pavement surface and subsurface condition evaluation. The accuracy of commercially available GPS systems (5 to 10 meters) is inadequate because data for road inspection is collected at 2.5cm or smaller intervals with sensors mounted on vehicles moving at 30 mph or faster. Video data recorded from a camera mounted on the vehicle can provide additional data registration to landmarks in the scene and previously recorded data. However, using video data poses additional challenges including the collection, processing and visualization of vast amounts of data, temporal and spatial registration among different cameras used at different times, natural clutter from unstructured environments, noise, and missing key data due to occlusion or dropped frames.

In this thesis, an image registration system composed of GPS, video tracking and tools from system theory is presented to register video sequences taken from moving vehicles achieving an accuracy of less than 15 cm. After narrowing down the searching scope with GPS, Scale-Invariant Feature Transform (SIFT) features are detected in the current and reference frames. Then, corresponding features are matched to obtain the projective transformation between each pair of frames where outliers are filtered by RANSAC. Finally, the current frame is registered to the reference frame whose geoinformation has been precisely known. In order to improve robustness and tracking accuracy, a dynamic system minimizing the rank of the Hankel matrix built by homography matrices in previous frames is introduced to predict the one in the following frame.
Acknowledgements

I would like to express my gratitude to my advisor, Prof. Octavia Camps for her
tireless efforts in supervising of my thesis.

Prof. Octavia Camps’s encouragement, patience, and ideas have been invaluable in
my research. Her extensive wealth of knowledge and enthusiasm for research problems
inspired me and taught me the right way to work.

I am also deeply indebted to Prof. Camps for her guidance throughout the course of
this research. She introduced me to Computer Vision area and has been a continual
source of help, support, and advice for the research in this thesis. Her dedication to
students is of great importance to me.

I would also like to acknowledge Prof. Gilead Tadmor and Prof. Mario Sznaier who
read my thesis and taught me linear system and control theory knowledge in their
courses. Without their help, this work would be in a much different state.

Finally, I would like to thank students in the Robust Control Lab. Mustafa Hacette-
peli and Necmiye Ozay provided me a number of valuable suggestions when I was
working on my thesis. I am amazed and inspired by their selflessness in guiding and
teaching younger students such as myself. Binlong Li and Fei Xiong also did me a
great favor for collecting data and discussing the problems I met in the thesis.

Boston, Massachusetts                      Bo Li
April 15, 2010
Chapter 1

Introduction

1.1 Background

Video registration has been the subject of substantial research in the last decade [19]. In general, video registration is required whenever we need to know what part of an object or scene a video frame depicts or where an object in that frame is located relative to a fixed coordinate system.

Consider, for example, the motivating geo-registration problem in this research. Geo-registration deals with the registration of a sensed image to a geodetically accurate reference image, such that the geodetic accuracy of the reference image can be transferred to the sensed image [1]. An intuitive idea is to register frames in the video to a reference map, like Google satellite imagery. However, the forward-view and top-view images only share features on the ground which is way insufficient to match one to the other. Therefore, the registration from video to video is taken into account. Still,
the reference and mission videos may be taken years apart, consequently may significantly differ due to changes in weather, environment, illumination, or viewpoint. Thus, a robust video registration technology invariant to these changes is required to accurately locate vehicles.

In general, video registration consists of two steps: image registration and feature tracking. Image registration is the process of geometrically aligning two images of the same scene taken from different viewpoints, at different times, or by different sensors. Image registration is a crucial step in computer vision tasks in which the final information is obtained from the combination of various data sources like image fusion, change detection, and multichannel image restoration. Due to the diversity of images to be registered and various types of degradations it is impossible to design a universal method applicable to all registration tasks. Nevertheless, the majority of the registration methods consists of three steps: feature detection, feature matching, and transform estimation [23]. Features are expected to be distinct, spread all over the image and efficiently detectable in both images. Feature-based detection methods have made impressive progress in recent years, such as region features [6], line features [2], and point features [10]. The detected features in the reference and sensed images can be matched by means of the image intensity values in their close neighborhoods, or the feature spatial distribution. In particular, correlation-like methods [7], Mutual
information methods [13], and Methods using invariant descriptors [9] have been proposed and proven to be able to perform effectively. After the feature correspondence has been established the mapping function is constructed which should transform the sensed image to the reference one where R. Hartley and A. Zisserman developed an automatic homography estimation algorithm in [11].

Video tracking is the process of locating moving objects in a sequence of frames. Due to temporary occlusion or dropped frames, key features are probably lost sometimes, so a consecutive tracking of features would play an important role. The two most common tracking techniques used in video tracking systems are Kalman filtering [12] and particle filtering [3, 18]. But the rule of the dynamic system has to be given before making predictions. In order to further improve robustness, the Hankel Rank Minimization estimation is adopted in this thesis, where the assumption of the dynamic law is unnecessary.

1.2 System Overview

The VOTERS (Versatile On-board Traffic Embedded Roaming Sensors) project, funding the research in this thesis, is developing new multi-sensor technology systems for cars and trucks that will allow for real-time assessment of road and bridge infrastructure across the country. The sensors will utilize acoustics and radar to look for potholes and cracks in the concrete and other abnormalities where a vehicle localization system is required to pinpoint collected data to very precise location. However,
the accuracy of commercially available GPS systems (5 to 10 meters) is inadequate because data for road inspection is collected at 2.5cm or smaller intervals with sensors mounted on vehicles moving at 30 mph or faster; and the cost (USD90,000) of advanced GPS systems (with accuracy of 10 centimeters) is not acceptable for mass application. Taking into account the trade-off between the accuracy and cost, a video registration system composed of GPS, on-board video camera and tools from system theory is presented to register video sequences taken from moving vehicles to the references, as shown in Figure 1.1. The experiment results show that it could achieve an accuracy of less than 15 centimeters.

Two image registration approaches are presented in this thesis: one is to register the sensed video to a reference map and the other is to a reference video. A flow diagram of the system for processing each frame is shown in Figure 1.2. As shown in the figure, the first step of this system is to obtain the reference, a reference map like Google Satellite Imagery, or a reference video where every frame is labeled with the ground-truth geographic information (longitude, latitude, and altitude) from the High-accuracy GPS or other references such as the Google Map Street View. When the vehicle running on the road, the video is taken from a forward looking regular camera mounted on the roof of the car and location information is collected from the normal GPS. After narrowing down the search scope by GPS information at the
moment, the map or the frame from the reference video closest to the current location is selected as the candidate corresponding image. Then, the image registration algorithm is implemented between the current and the candidate reference image. In the sequel, the features in the view of the camera have been exactly located and it is straightforward to get the accurate location of the vehicle after computing the distance between the camera and features.

At last but not least, in order to deal with the problem of missing data and dropped frames, a dynamic system minimizing the rank of the Hankel matrix built by homography matrices in previous frames is applied to predict the one in following frames.

1.3 Thesis Organization

This thesis will focus on developing algorithms of registering videos with dynamic methods. In Chapter 2, the SIFT feature descriptor is introduced for the detection and matching of features in images. In Chapter 3, geometry concepts of image mapping applied in the step of image registration are described and a plane induced homography method is proposed for registering to videos. Chapter 4 of this work presents the Hankel matrix based rank minimization dynamic system which is helpful to improve algorithm robustness. The performance of the image registration algorithms and the dynamic system is characterized in Chapter 5. Finally, a summary of the approaches and results as well as future areas of research relevant to this topic
are given in Chapter 6.

Figure 1.1: The overview of system equipments: (a) the vehicle-mounted camera (Panasonic PV-GS500); (b) the GPS receiver (Garmin 76CSx).
Figure 1.2: Flow chart of the algorithm used in this thesis
Chapter 2

Scale Invariant Feature Transform

A detailed explanation of the Scale Invariant Feature Transform descriptor (SIFT) is given in [15]. However, for completeness, a brief overview is discussed in this chapter. Furthermore, a new approach, Harris-SIFT, is proposed to reduce the probability of disruption by illumination change or different sensors in this chapter.

2.1 SIFT Algorithm

SIFT algorithm is designed to extract distinctive invariant features from images for matching between different views of an object or scene. The features are invariant to image scale and rotation, and partially robust across a substantial range of affine distortion, change in 3D viewpoint, and addition of noise. Due to the unpredictable difference of view points between the sensored and reference videos, the SIFT algorithm is applied to detect and describe features for the image matching. The SIFT algorithm consists of four major stages: (1) scale-space peak selection; (2) keypoint localization; (3) orientation assignment; (4) keypoint descriptor.
2.1.1 Scale-space Peak Selection

Detecting locations that are invariant to scale change of the image can be accomplished by searching for stable features across all possible scales, using a continuous function of scale known as scale space [22]. The scale space of an image is defined as a function, \( L(x, y, \sigma) \), that is produced from the convolution of a variable-scale Gaussian, \( G(x, y, \sigma) \), with an input image, \( I(x, y) \):

\[
L(x, y, \sigma) = G(x, y, \sigma) \ast I(x, y)
\]  

(2.1.1)

where \( \ast \) is the convolution operation in \( x \) and \( y \), and

\[
G(x, y, \sigma) = \frac{1}{2\pi\sigma^2}e^{-(x^2+y^2)/2\sigma^2}
\]  

(2.1.2)

To efficiently detect stable keypoint locations in scale space, [14] proposed using scale-space extrema in the difference-of-Gaussian function convolved with the image, \( D(x, y, \sigma) \), which can be computed from the difference of two nearby scales separated by a constant multiplicative factor \( k \):

\[
D(x, y, \sigma) = (G(x, y, k\sigma) - G(x, y, \sigma)) \ast I(x, y) = L(x, y, k\sigma) - L(x, y, \sigma)
\]  

(2.1.3)

An efficient approach to construction of \( D(x, y, \sigma) \) is shown in Figure 2.1. For each octave of scale space, the initial image is repeatedly convolved with Gaussians to produce the set of scale space images shown on the left. Adjacent Gaussian images are subtracted to produce the difference-of-Gaussian images on the right. After each
octave, the Gaussian image is down-sampled by a factor of 2, and the process repeated.

![Figure 2.1: The construction of Difference of Gaussian (DOG).[15]](image)

### 2.1.2 Keypoint Localization

In order to detect the local maxima and minima of $D(x, y, \sigma)$, each sample point is compared to its eight neighbors in the current image and nine neighbors in the scale above and below as shown in Figure 2.2. It is selected only if it is larger than all of these neighbors or smaller than all of them. After a keypoint candidate has been found by comparing a pixel to its neighbors, a detailed fit to the nearby data for location, scale, and ratio of principal curvatures is performed to reject points that
have low contrast or are poorly localized along an edge.

![Figure 2.2: Maxima and minima of the difference-of-Gaussian images detection.][15]

### 2.1.3 Orientation Assignment

By assigning a consistent orientation to each keypoint based on local image properties, the keypoint descriptor can be represented relative to this orientation and therefore achieve invariance to image rotation. For each image sample, \( L(x, y, \sigma) \), at a certain scale, the gradient magnitude, \( m(x, y, \sigma) \), and orientation, \( \theta(x, y, \sigma) \), is precomputed using pixel differences:

\[
m(x, y) = \sqrt{(L(x + 1, y) - L(x - 1, y))^2 + (L(x, y + 1) - L(x, y - 1))^2} \quad (2.1.4)
\]

\[
\theta(x, y) = \tan^{-1} \left( \frac{(L(x, y + 1) - L(x, y - 1))}{(L(x + 1, y) - L(x - 1, y))} \right) \quad (2.1.5)
\]
An orientation histogram is formed from the gradient orientations of sample points within a region around the keypoint. The orientation histogram has 36 bins covering the 360 degree range of orientations. Each sample added to the histogram is weighted by its gradient magnitude and by a Gaussian-weighted circular window with a $\sigma$ that is 1.5 times that of the scale of the keypoint. Peaks in the orientation histogram correspond to dominant directions of local gradients.

### 2.1.4 Keypoint Descriptor

A keypoint descriptor is created by first computing the gradient magnitude and orientation at each image sample point in a region around the keypoint location, as shown on the left of Figure 2.3. These are weighted by a Gaussian window, indicated by the overlaid circle. The purpose of this Gaussian window is to avoid sudden changes in the descriptor with small changes in the position of the window, and to give less emphasis to gradients that are far from the center of the descriptor, as these are most affected by misregistration errors. These samples are then accumulated into orientation histograms summarizing the contents over 4x4 subregions which allows for significant shift in gradient positions, as shown on the right of Figure 2.3, with the length of each arrow corresponding to the sum of the gradient magnitudes near that direction within the region. The descriptor is formed from a vector containing the values of all the orientation histogram entries, corresponding to the lengths of the arrows on the right side of Figure 2.3. This figure shows a 2x2 descriptor array.
computed from an 8x8 set of samples, whereas the experiments in this research use 4x4 descriptors computed from a 16x16 sample array. Therefore, there are 4x4x8 = 128 elements in the feature vector of each keypoint.

2.2 Harris-SIFT

Because the videos might be taken by cameras with different configurations, or at different time, there could be great difference in illumination, or brightness between the sensored frame and the reference images, which eventually stops the SIFT algorithm from finding correspondences. But the SIFT descriptor does perform nicely to describe the distinctive features. Therefore, to location keypoints, instead of selecting scale-space peaks as presented in the first step of the SIFT algorithm, Harris corner detection [10] is implemented which is barely affected by illumination or brightness.
In the very beginning, the frame collected from the camera mounted on the vehicle is roughly aligned to the reference image which could be from the reference video or satellite image. This step allows the sensed frame to have similar viewpoint as the reference image, which is processed off-line. For each corner in the roughly aligned image (P), find it’s SIFT descriptor, then search the SIFT descriptors in a window in the reference image (S) centered at the location of the feature in P. In the matching measurement section, a correspondence is qualified when the ratio of vector angles from the nearest to second nearest neighbor is less than a threshold. The algorithm is summarized in Table 5.3 of Chapter 5. Figure 2.4 gives an example of the Harris-SIFT algorithm.
Chapter 3

Plane Induced Homography Given Fundamental Matrix

After finding the corresponding features between two images, the next step of image registration is image transformation mapping the road plane of the sensor frame to that of the reference one, which is accomplished by applying the homography matrix computed by at least four correspondence points to the sensor frame. However, there are not always enough features on the road to make the projective transformation directly. Therefore, the off-road features are taken into account, such as traffic signs, buildings, and even trees. Given the fundamental matrix obtained from at least eight correspondences everywhere in a pair of frames and two pairs of corresponding lines on the road planes which could be easily detected at most time, the homography matrix between road planes in the two frames can be induced by inexpensive computation. In this chapter, the concepts of homography and fundamental matrix are presented, and then the plane induced homography approach is introduced.
3.1 Projective Geometry

Defined by [11], a planar projective transformation is a linear transformation on homogeneous 3-vectors represented by a non-singular 3x3 matrix:

\[
\begin{bmatrix}
x' \\
y' \\
1
\end{bmatrix}
= \begin{bmatrix}
h_{11} & h_{12} & h_{13} \\
h_{21} & h_{22} & h_{23} \\
h_{31} & h_{32} & h_{33}
\end{bmatrix}
\begin{bmatrix}
x \\
y \\
1
\end{bmatrix}
\tag{3.1.1}
\]

\[
x' = \frac{h_{11}x + h_{12}y + h_{13}}{h_{31}x + h_{32}y + h_{33}}
\]

\[
y' = \frac{h_{21}x + h_{22}y + h_{23}}{h_{31}x + h_{32}y + h_{33}}
\tag{3.1.2}
\]

An example of projective transformation is given in Figure 3.1. Only the ratio of the matrix elements is significant and it follows that a projective transformation has eight degrees of freedom. Each point correspondence gives rise to two independent equations in the entries of \( H \), so four such point correspondences provide a set of equations \( AH = b \). Then the homography matrix \( H \) can be estimated by

\[
H = (A^T A)^{-1} (A^T b)
\tag{3.1.3}
\]

Because the measurement of corresponding points might be inexact, more than four points are used to estimate the homography matrix to improve the accuracy. So it becomes an over-determined problem and the popular RANSAC robust estimation is applied to remove the outliers. The detail description of the homography estimation algorithm is introduced in [11] (Algorithm 4.6).
Figure 3.1: An illustration of the projective transformation

3.2 Epipolar Geometry

The epipolar geometry is the intrinsic projective geometry between two views. It is independent of scene structure, and only depends on the cameras' internal parameters and relative pose. The fundamental matrix $F$ encapsulates this intrinsic geometry. It is a 3x3 matrix of rank 2. If a point in world is imaged as $x$ in the first view, and $x'$ in the second, then the image points satisfy the relation $x'^T F x = 0$. Figure 3.2 shows that to each point $x$ in one image, there exists a corresponding epipolar line $l'$ in the
Figure 3.2: Point correspondence geometry. (a) The two cameras are indicated by their centers C and C' and image planes. (b) An image point x back-projects to a ray in 3-space defined by the first camera center C and x. This ray is imaged as a line l' in the second view. The 3-space point X which projects to x must lie on this ray, so the image of X in the second view must lie on l'.[11]

other image. Given at least eight pairs of point matches, the above equation can be used to compute the unknown matrix $F$. In particular, the equation corresponding to a pair of points $(x, y, 1)$ and $(x', y', 1)$ is

$$x'xf_{11} + x'yf_{12} + x'f_{13} + y'xf_{21} + y'yf_{22} + y'f_{23} + xf_{31} + yf_{32} + f_{33} = 0 \quad (3.2.1)$$

From a set of $n$ point matches, a set of linear equations of the form is obtained

$$AF = \begin{bmatrix} x'_{1}x_{1} & x'_{1}y_{1} & x'_{1} & y'_{1}x_{1} & y'_{1} & y'_{1} & x_{1} & y_{1} & 1 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ x'_{n}x_{n} & x'_{n}y_{n} & x'_{n} & y'_{n}x_{n} & y'_{n} & y'_{n} & x_{n} & y_{n} & 1 \end{bmatrix} F = 0 \quad (3.2.2)$$
The solution is the generator of the right null-space of $A$. Similar to the estimation of the homography matrix, more than the minimum number of correspondences are considered, so the RANSAC is applied as well. The automatic estimation of fundamental matrix is detailed in [11] (Algorithm 11.4).

### 3.3 Plane Induced Homography Given Fundamental Matrix And Line Correspondences

Images of points on a plane are related to corresponding image points in a second view by a planar homography as shown in section 3.1. This is a projective relation since it depends only on the intersections of planes with lines. It is said that the plane induces a homography between the views. There are then two relations between the two views: first, through the epipolar geometry a point in one view determines a line in the other which is the image of the ray through that point; and second, through the homography a point in one view determines a point in the other which is the image of the intersection of the ray with a plane.

In particular, homography matrices induced by planes have to be compatible with epipolar geometry. Suppose four points $X_i$ are chosen on a scene plane. Then the correspondence $x_i \leftrightarrow x'_i$ of their images between two views defines a homography $H$, which is the homography induced by the plane. These image correspondences also obey the epipolar constraint, $x'^T F x = 0$, since they arise from images of scene points. Thus, the homography $H$ is said to be compatible with $F$. Formally, a homography
$H$ is compatible with a fundamental matrix $F$ if and only if the matrix $H^TF$ is skew-symmetric:

$$H^TF + F^TH = 0$$

(3.3.1)

Therefore, given the fundamental matrix $F$ between two views, the homography induced by a world plane is

$$H = A - e'v^T$$

(3.3.2)

where $[e']_xA = F$ is any decomposition of the fundamental matrix, and $v$ is the world plane. However, it is hard to get the coordinates of the road plane in the real world. A plane can be defined by three points, one line and one point, or two lines, and it is obvious that there are at least two lines on the road plane in the driving direction at most time, that is, the lane lines or road edges. So, given the fundamental matrix, the problem is reduced to that of solving for $v$ from the two line correspondences. According to [11], one pair of corresponding image lines determines a line in 3-space, and a line in 3-space lies on a pencil of planes as shown in Figure 3.3. The homography for the pencil of planes defined by a line correspondence $l\leftrightarrow l'$ is given by

$$H(\mu) = [l']_xF + \mu e'l^T$$

(3.3.3)

provided $l'^Te'\neq 0$, where $\mu$ is a projective parameter.

Next, $\mu$ is solved with the two line correspondences. The pencil of planes decided by
Figure 3.3: The pencil of planes defined by a line correspondence. (a) Image lines $l$ and $l'$ determine planes $\pi$ and $\pi'$ respectively. The intersection of these planes defines the line $L$ in 3-space. (b) The line $L$ in 3-space is contained in a one parameter family of planes $\pi(\mu)$. This family of planes induces a one parameter family of homographies between the images.\[11\]

$l\leftrightarrow l'$ is

$$\pi(\mu) = \mu \begin{bmatrix} l \\ 0 \end{bmatrix} + \begin{bmatrix} A^Tl' \\ e^Tl' \end{bmatrix}$$ \hspace{1cm} (3.3.4)$$

Similarly, the pencil of planes decided by $m\leftrightarrow m'$ is

$$\pi(\lambda) = \lambda \begin{bmatrix} m \\ 0 \end{bmatrix} + \begin{bmatrix} A^Tm' \\ e^Tm' \end{bmatrix}$$ \hspace{1cm} (3.3.5)$$

Because the two pairs of lines are on the same plane, there should be an intersection between the two pencils of planes, which will be the road plane. Thus,

$$\pi(\mu) = k\pi(\lambda) \hspace{1cm} (3.3.6)$$

$$\mu \begin{bmatrix} l \\ 0 \end{bmatrix} + \begin{bmatrix} A^Tl' \\ e^Tl' \end{bmatrix} = k(\lambda \begin{bmatrix} m \\ 0 \end{bmatrix} + \begin{bmatrix} A^Tm' \\ e^Tm' \end{bmatrix}) \hspace{1cm} (3.3.7)$$
\[
\begin{bmatrix}
    l & m & A^T m' \\
    0 & 0 & e^T m'
\end{bmatrix}
\begin{bmatrix}
    \mu \\
    -k\lambda \\
    -k
\end{bmatrix} = -\begin{bmatrix}
    A^T l' \\
    e^T l'
\end{bmatrix}
\] (3.3.8)

It is an over-determined problem but the last equation is always true, so the first three equations are used to solve the three unknowns, \( \mu \), \( \lambda \), and \( k \):

\[
\begin{bmatrix}
    \mu \\
    -k\lambda \\
    -k
\end{bmatrix} = -\begin{bmatrix}
    l & m & A^T m'
\end{bmatrix}^{-1} A^T l' 
\] (3.3.9)

Finally, \( \mu \) is inserted to the equation 3.3.3 and the homography matrix \( H \) is achieved.
Chapter 4

Hakel Based Rank Minimization

Ideally, all of the feature trajectories over all frames are expected to be complete. But it is not always the case in the real world applications; features might be occluded or out of view, frames could be dropped, and tough scenarios exist where there is a abrupt change, such as the sharp turn, bumpy surface and so on.

In this chapter, a rank minimization approach is introduced to achieve the recovery of missing data and the prediction of future data. First, the rank minimization problem is defined. Then, the Hankel based rank minimization approach is presented.

4.1 Rank Minimization Problem

Rank Minimization Problem (RMP) arises in diverse areas such as minimum order controller design [16], factor analysis in statistics [20], and Euclidean distance matrix
problems [21]. The general matrix rank minimization problem can be expressed as

\[
\begin{align*}
\text{minimize} & \quad \text{Rank}(H) \\
\text{subject to} & \quad H \in C
\end{align*}
\] (4.1.1)

where \(H \in \mathbb{R}^{m \times n}\) is the optimization variable and \(C\) is a convex set.

Rank minimization problems are NP-hard so some relaxations are necessary before attempting to solve the problem. One of the best efforts to solve the RMP was made by Fazel in her PhD thesis [5], which is named Trace Heuristic. After relaxing the problem, any RMP in equation 4.1.1 can be converted to:

\[
\begin{align*}
\text{minimize} & \quad \text{trace}(X) + \text{trace}(Y) \\
\text{subject to} & \quad \begin{bmatrix} X & H \\ H^T & Y \end{bmatrix} \succeq 0 \\
& \quad H \in C
\end{align*}
\] (4.1.2)

where \(X\) and \(Y\) are both symmetric matrices and \(X \in \mathbb{R}^{m \times n}\) and \(Y \in \mathbb{R}^{m \times n}\) are new variables, which converts the rank minimization problem into a positive semidefinite rank minimization problem. The necessity of the matrix to be rank minimized being positive semidefinite is, if the matrix \(H\) is positive semidefinite in equation 4.1.1 then we can relax problem by changing rank function with trace function. Minimization of trace instead of rank was proven to minimize the convex envelope of rank in [5]. Then, it is an Linear Matrix Inequality (LMI) can be solved by using the existing tools. In this research, the cvx software for MATLAB [8] is used to solve the LMI.
4.2 Hankel Based Rank Minimization Approach For Tracking

The main idea behind the proposed approach is to find the homography matrix, mapping the road plane of the next frame to the one of the reference, that maximizes the smoothness of the resulting sequence of homography matrices. Equivalently, denoting by $H_{1\sim n-1}$ the observed descriptors and by $H_n$ the missing ones, the idea is to find the values of $H_n$ that are maximally consistent with $H_{1\sim n-1}$, in the sense that the predicted homography matrix follows the trend of the homography matrices in previous frames. Specifically, the simplest dynamic system that explains the measurement data corresponds to the values of the missing pixels that minimize this rank. Since it is well known that rank minimization problems are NP hard, as described above, rather than minimizing rank, this problem is relaxed to a convex semi-definite programming problem, with the additional advantage of improving robustness against noise. These ideas lead to the following Algorithm:

1. Given the homography matrices from previous frames, form the Hankel matrix:

$$\text{Hankel} = \begin{bmatrix}
\alpha_1 H_1 & \alpha_2 H_2 & \ldots & \alpha_{(n+1)/2} H_{(n+1)/2} \\
\alpha_2 H_2 & \alpha_3 H_3 & \ldots & \alpha_{(n+1)/2+1} H_{(n+1)/2+1} \\
\vdots & \vdots & \ddots & \vdots \\
\alpha_{(n+1)/2} H_{(n+1)/2} & \alpha_{(n+1)/2+1} H_{(n+1)/2+1} & \ldots & H_n
\end{bmatrix} \quad (4.2.1)$$

Here, the 3x3 matrix $H_i$ is rearranged to 9x1 vector before plugged into the Hankel matrix, and the homography matrix of the $nth$ frame $H_n$ is predicted. Since the
homography matrix only has eight degrees of freedom while there are nine elements in the matrix, it has to be normalized before building up the Hankel matrix. However, normalization of homography matrices will generate different dynamic systems while there is supposed to be only one dynamic system. Furthermore, because of noises, the rank of the system (or the Hankel matrix) will be greater than what it is supposed to be. (The rank of Hankel matrix is three in this research as the rank of homography matrix is three.) To solve these problems, an additional set of parameters $\alpha_i$ are introduced when the rank minimization algorithm is implemented, so the homography matrices are automatically normalized in the rank minimization process and the rank of the system could be as low as the one it should be.

2. Estimate $H_n$, as well as $\alpha_{1\sim n-1}$, which is maximally consistent with $H_{1\sim n-1}$ by solving the following Linear Matrix Inequality (LMI) optimization problem:

$$
\begin{align*}
\text{minimize} & \quad \text{trace}(X) + \text{trace}(Y) \\
\text{subject to} & \quad \begin{bmatrix}
X & \text{Hankel} \\
\text{Hankel}^T & Y
\end{bmatrix} \succeq 0, \\
& \quad \Sigma \alpha_i = \text{constant}, \quad H \in C
\end{align*}
$$

(4.2.2)
Chapter 5
Performance Characterization

The following chapter provides an evaluation of the performance of two video registration algorithms proposed in this thesis. Firstly, the sequence of frames is registered to a reference image, such as a high resolution map image; Secondly, the collected video is registered to a reference video whose frames have been accurately located. Besides, the performance of the Hankel based rank minimization prediction algorithm is characterized.

5.1 Register To Reference Image

As to geo-registration problem, it is intuitive to register the sensed data to a map image that includes the geographic information of every point. In this section, the algorithm is performed on the synthetic data collected in the lab first, then it is moved outside to test on the real road data.
5.1.1 Synthetic Data

It is intuitive to start solving problems with the simplest case, so in the very beginning, the algorithm is developed on the synthetic road data collected by a regular camera (resolution 480x720) in the lab, as illustrated in Figure 5.1.

![Figure 5.1: Synthetic data in the lab (a) Satellite map image (b) Vehicle image](image)

In order to prepare enough features to be detected, five kinds of standard arrows are aligned to simulate the road intersection. The scene is taken by two cameras, one from the top view (like the satellite map image) and the other from the forward view (like the image taken from vehicle). The approach outline is given in Algorithm 1.

It is obvious that there is a great change between these two viewpoints, at least 70 degrees. However, the matching accuracy of the SIFT algorithm could only achieve 50% out to a 50 degree change in viewpoint [15]. Therefore the images have to be preprocessed before running SIFT on them. Instead of finding the initial homography matrices by manually picking correspondences in two images, Affine-SIFT [17] is
Table 5.1: Algorithm 1: register to map image (synthetic data)

<table>
<thead>
<tr>
<th>Algorithm 1: register to map image (synthetic data)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Step 1</strong> Find the initial homography matrices $H_1 \sim H_{16}$ between the first 16 video frames $f_1 \sim f_{16}$ and the satellite image $S$ by Affine-SIFT and RANSAC.</td>
</tr>
<tr>
<td><strong>Step 2</strong> Build Hankel matrix with these homography matrices, and minimize its rank on it to predict the homography matrix $H_{17}^-$ for the next frame.</td>
</tr>
<tr>
<td><strong>Step 3</strong> Roughly rectify $f_{17}$ by applying $H_{17}^-$ on it, and get $f_{17}'$.</td>
</tr>
<tr>
<td><strong>Step 4</strong> Apply SIFT and RANSAC between $S$ and $f_{17}'$ to find correspondences and get $H_{17}^+$.</td>
</tr>
<tr>
<td><strong>Step 5</strong> $H_{17} = H_{17}^- H_{17}^+$.</td>
</tr>
<tr>
<td><strong>Step 6</strong> Return to Step 3 to predict $H_{18}$ with $H_2 \sim H_{17}$ ...</td>
</tr>
</tbody>
</table>

applied to automatically find corresponding points between the satellite image and the initial video frames, as shown in Figure 5.2. (the number of initial frames is determined by experiments to optimize the Hankel rank minimization prediction. In this thesis, it is set to be 16.) Affine-SIFT has been proved to be able to tolerate up to almost a 90 degree viewpoint change, but its computation is so expensive that is only applied in the initial frames.

Then RANSAC is implemented on the correspondences to remove outliers and the homography matrix is estimated by the remaining inliers. The transformed vehicle image is overlaid on the satellite image as shown in Figure 5.3. In this figure, the two images are overlaid accurately, so the Affine-SIFT could perform perfectly to find correspondences and introduce little error in this step.
Figure 5.2: Correspondences found by Affine-SIFT between satellite image and the first video frame.

Figure 5.3: Overlay the transformed vehicle image on the satellite image.
Table 5.2: Quantitative inspection of Hankel RM (in pixel)

<table>
<thead>
<tr>
<th>Frames</th>
<th>155</th>
<th>160</th>
<th>165</th>
<th>170</th>
<th>175</th>
</tr>
</thead>
<tbody>
<tr>
<td>With Hankel RM</td>
<td>15.170</td>
<td>5.594</td>
<td>18.953</td>
<td>13.254</td>
<td>7.900</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Frames</th>
<th>180</th>
<th>185</th>
<th>190</th>
<th>195</th>
<th>200</th>
</tr>
</thead>
</table>

Frames $f_2 \sim f_{16}$ are processed in the same way and 16 homography matrices $H_1 \sim H_{16}$ are obtained. As described in chapter 4, all homography matrices as well as the unknown matrix $H_{17}$ are rearranged as 9x1 vectors and used to build the Hankel matrix (81x9). After minimizing the rank of the Hankel matrix, the predicted homography matrix $H_{17}$ is outputted. The performance of Hankel rank minimization (Hankel RM) is characterized in Figure 5.4, Figure 5.5, and Table 5.2. Figure 5.4 shows that most of the entries of the homography matrices predicted by Hankel RM (red) is very closed to those of the ground truth (blue). Figure 5.5 compares the mapping results of homography matrices predicted by Hankel RM and the one borrowed from the previous frame, and consequently indicates the one with Hankel RM could provide more accurate result. In Table 5.2, they are quantitatively compared by computing the average distance (in pixel) between matching points and the one with Hankel RM does make less average error.
Figure 5.4: The performance of Hankel rank minimization: each subfigure represents each entry of the 3x3 homography matrix, blue lines are the ground truth generated by manually selection and red lines are the results predicted by Hankel rank minimization.

Figure 5.5: Compare mapping result of with and without Hankel (a) Without Hankel: just apply the homography matrix of the last frame (b) With Hankel: apply the homography matrix predicted by previous 16 frames.
By applying $H_{17}$, $f_{17}$ is roughly rectified to $f'_{17}$, and then the SIFT algorithm is implemented between $S$ and $f'_{17}$ to find correspondences between them, as shown in Figure 5.6(a). After removing outliers by RANSAC as shown in Figure 5.6(b), $H_{17}^+$ is computed and the homography matrix of the current frame is finally achieved $H_{17} = H_{17}^- H_{17}^+$. 

![Figure 5.6: The current frame is roughly rectified (right) and registered to the satellite image (left). a. Before RANSAC; b. After RANSAC.](image)

According to the experimental results above, the SIFT algorithm performs effectively to detect distinctive correspondences for image registration, so it is time to
move it outside to check if it is also available for the road data in real world.

5.1.2 Real Data

Since the SIFT algorithm has been proven to be trustable for image registration in the friendly environment, it is time to test it on actual roads in the real world. The satellite image is provided by Google Map and the vehicle video frames are collected from a regular camera mounted on the hood of a car, as shown in Figure 5.7. (Here, features on the pavement are the only interesting information, so the camera is mounted downward; While in the next section, the off-road information is also important for video registration, so the camera is mounted on the roof facing forward.)

In order to remove the huge viewpoint difference between the two images to implement the SIFT algorithm on them, correspondences are manually picked to compute the homography matrix and the pseudo-satellite image is obtained, as shown on the right of Figure 5.8. This figure shows that no one correspondence is detected by SIFT between the two images even though they look similar to each other. This could be explained by the plot of the SIFT feature descriptors in Figure 5.9: at each corresponding locations in this figure, it is hard to find similar SIFT features between them. This problem might be caused by the great illumination and brightness differences between the two images collected by different sensors or at different time.
Table 5.3: Algorithm 2: register to map image (real data) (Harris-SIFT)

<table>
<thead>
<tr>
<th>Algorithm 2: register to map image (real data) (Harris-SIFT)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Step 1 Apply Harris corner detection on the satellite image (S) and the pseudo-satellite image (P).</td>
</tr>
<tr>
<td>Step 2 For each point in P, compute its SIFT descriptor.</td>
</tr>
<tr>
<td>Step 3 Find the SIFT descriptors in a window in S centered at the location of the feature in P.</td>
</tr>
<tr>
<td>Step 4 Compute the Euclidean distances between the descriptors of the point to be matched and those of the neighbor corners.</td>
</tr>
<tr>
<td>Step 5 If the ratio of vector angles from the nearest to second nearest neighbor is less than a 0.6, the nearest neighbor is kept as the matching point.</td>
</tr>
</tbody>
</table>

It is true that the SIFT algorithm does not work in this scenario, but the SIFT descriptor is still an effective approach to describe distinctive features. The only problem is that the SIFT features could not be located at the spots where they are expected to be, that is, interesting features. Therefore, instead of selecting keypoints from scale-space peaks as Lowe did [15], Harris corner detection is employed which is free from illumination or brightness changes. In this section, a new approach called Harris-SIFT is proposed, which is summarized in Table 5.3.

Since the resolution of the satellite image is much lower than the one of the vehicle image and Harris corner detection is not scale invariant, the satellite image is enlarged by 4 times to achieve a similar scale as the vehicle image. Then the blurred satellite image is sharpened to facilitate Harris corner detection. The preprocessed satellite image and pseudo-satellite image are illustrated in Figure 5.10. And Figure 5.11 gives an example of the SIFT descriptors of a pair of corresponding points. Then, for
each point in P, the SIFT descriptors are computed in a window in S centered at the location of the feature in P and search for the nearest neighbors. Only if the ratio of the Euclidean distance to the nearest neighbor and the second nearest is less than 0.6, it is considered as a matching point. Figure 5.12 gives a couple of examples. While the algorithm works for a couple of individual points, it could not give a reliable result when it is implemented in the entire frame. As shown in Figure 5.13, there are only 8~10 inliers out of 35 correspondences after RANSAC, and consequently the homography matrix would not be trustable.

Therefore, according to the results above, the Harris-SIFT algorithm seems unsuitable for solving this problem. After comprehensive analysis of possible reasons, the dominant one is still the great disparity in pixel information between the satellite image and the vehicle image, which eventually stops SIFT from finding the correspondences. Harris-SIFT does provide correct correspondences when the features are enough similar to each other and there are not many noisy features closed to them. However, the differences in illumination or brightness might be too great to make the same feature dissimilar between the two images taken by different sensors or at different times. What is more important, a critical limitation exists if the satellite image is chosen as the reference to be registered: there are ample landmarks only at intersections while just lanes and road edges lie on the road at most time, but the SIFT algorithm will fail when the object has a self similar or periodic structure.
In summary, there are three major troubles in this research until now: 1. Great differences in pixels between the satellite and vehicle images; 2. The resolution of the satellite image is relatively lower than the vehicle image. 3. There are not enough features on the road in both images. In order to overcome these difficulties, a new system is proposed in next section where the sensed video is registered to a reference video which is taken by the same camera and accurately located beforehand. Therefore, not only are the illumination differences between images eliminated, but also the plenty of off-road features are available for detection and matching.
Figure 5.7: Real road data (a)Satellite image (b)Vehicle image.
Figure 5.8: SIFT is stuck in the real world data: no correspondences are detected. left:satellite image; right:pseudo-satellite image.

Figure 5.9: the plot of the SIFT feature descriptors: left:satellite image; right:pseudo-satellite image.
Figure 5.10: Harris corner detection on the satellite and pseudo-satellite image: left: preprocessed satellite image; right: pseudo-satellite image (for easy testing, only the corners around the white arrows are kept)

Figure 5.11: The illustration of SIFT descriptors of a pair of corresponding points in the satellite and pseudo-satellite image.
Figure 5.12: Two examples of Harris-SIFT results: Blue circles are the selected point in P, green circles are the detected matching point. The size of searching window is 15x15 and the size of images is 512x512.
5.2 Register To Reference Video

The reference video has three advantages over the reference map: 1. The off-road features are available for detection and matching. 2. It has similar viewpoint as the sensed video; 3. The two videos could be taken by the same type of cameras. The reference video is obtained also by a regular camera mounted on the roof of the car, but an advanced GPS is equipped. In order to use the off-road features, the fundamental matrix and line correspondences on the road are combined to get the homography matrix between road planes in the two frames. The steps of the algorithm are summarized in algorithm 3.

The same as registering to reference image, the SIFT algorithm is applied to detect
Table 5.4: Algorithm 3: plane induced homography

<table>
<thead>
<tr>
<th>Algorithm 3: Plane induced homography</th>
</tr>
</thead>
<tbody>
<tr>
<td>Step 1</td>
</tr>
<tr>
<td>Step 2</td>
</tr>
<tr>
<td>Step 3</td>
</tr>
<tr>
<td>Step 4</td>
</tr>
<tr>
<td>Step 5</td>
</tr>
</tbody>
</table>

features and find matching points, but the scan region is not only restricted to road plane. Then, with the algorithm 11.4 in [11], the fundamental matrix between the two frames is automatically estimated by using RANSAC, as shown in Figure 5.14.

![Figure 5.14: The estimation of fundamental matrix by RANSAC: 27 inliers are kept from 28 correspondences provided by SIFT](image)

Meanwhile, the Hough transform [4] is applied on the road plane to detect lines
in the driving direction, as shown in Figure 5.15. The Hough transform is a method
of detecting straight lines in images by a voting procedure. Given a set of collinear
edge points, each of them have associated a line in parameter space. These lines
intersect at the point corresponding to the parameters of the line in the image space.
At each point of the parameter space, count how many lines pass through it within
an array of counters. The higher the count, the more edges are collinear in the image
space. The peaks in the counter array are considered as the parameters of lines in
the image space by setting a threshold. In practice, the straight line \( y = mx + b \)
can be represented as a point \((b, m)\) in the parameter space. However, there is the
problem that vertical lines give rise to unbounded values of the parameters \(m\) and
\(b\). For computational reasons, it is therefore better to parameterize the lines in the
Hough transform with two other parameters, commonly referred to as \(\rho\) and \(\theta\). So
the line function will be \(\rho = x \cos \theta + y \sin \theta\), as illustrated in Figure 5.16. In this
example, four lines are detected and the most left and the second most right lines are
selected as the two lines to define the road plane. By tracking the lines with their
directions, their corresponding lines in the reference frame could be easily found.

Finally, the fundamental matrix and the two pairs of lines are given to equation
3.3.9 to get the parameter \(\mu\) used for solving equation 3.3.3, and then the homography
matrix between the road planes of the two frames is achieved. In Figure 5.17, the
mapping result of the sensored frame is overlapped on the reference one where the
Figure 5.15: Line detection by Hough transform. Left: the reference frame; Right: the sensed frame.

Matching error (in pixel) is $[\delta x, \delta y] = [11.23, 4.99]$.

According to the decent mapping result in Figure 5.17, the plane induced homography algorithm is an effective approach of video registration for the research in this thesis. Nevertheless, the algorithm might be so complicated that is sensitive to noises introduced in the step of fundamental matrix computation or line detection, which would eventually generate a inappropriate homography matrix.
Figure 5.16: The illustration of Hough transform. (a): the line is parameterized by $(\rho, \theta)$. (b): the counter array in the parameter space where the "bright" points are the detected lines in the image space.
Figure 5.17: The mapping result of the plane induced homography algorithm. The right frame in Figure 5.14 is transformed to the left one by applying the homography matrix and then overlapped on the left frame.
After successfully registering the sensored frame to the reference frame, the features in the view of camera have been accurately located. Then, the vehicle would be subsequently located by computing the ground distance between the features and the camera with the triangulation method, as illustrated in Figure 5.18. The height of the camera $h$ and the pitch angle $a_2$ could be easily measured and the field of view $b_2$ can also be known by referring the camera manual. In Figure 5.18(a), the vertical distance $y$ is computed by:

$$\arctan a_1 = \frac{PA}{PO} \arctan a_2$$

$$y = h \arctan a_1$$

And the lateral distance $x$ is computed by:

$$\tan b_1 = \frac{OA}{OB} \tan b_2$$

$$x = y \tan b_1$$

where $PA, PO, OA, OB$ are measured with pixels in the image; $O$ is the center of the image, $B$ is the right boundary, $A$ is the feature point on the road and $P$ is the projection of the camera $C$ on the ground.

At last, with the relative ground distance between the camera and features, the location of the vehicle could be given by:

$$P = (x_A - x, y_A - y)$$
where \((x_A, y_A)\) is the coordinate of the feature point A.

Figure 5.18: Compute the distance between the camera and features. (a) compute the vertical distance \(y\). (b) compute the lateral distance \(x\).
Chapter 6

Conclusion And Future Work

In this thesis, two methods of performing video geo-registration are presented to improve the accuracy of vehicle localization. At first, the sensored video is registered to a reference image, the Google satellite map, by finding the homography matrix obtained from the corresponding features that are detected by the SIFT algorithm and filtered by the RANSAC algorithm between the road planes in the sensored frame and the reference image. However, due to the limitations of inadequate landmarks on roads and illumination differences between the two images, this approach is not always appropriate for the universal application.

In the sequel, another method is proposed to overcome these difficulties. Instead of registering the sensored video to the top-view satellite image, the video is registered to a reference video whose frames have been exactly located in advance. In this way, besides landmarks on the pavement, the off-road features could be considered as well to estimate the fundamental matrix between the two frames by using RANSAC. In the
meantime, Hough transform is applied on the road to detect pairs of corresponding lines, two of which are selected to define the road plane. Finally, the road plane induced homography matrix is computed given the fundamental matrix and the two pairs of correspondences.

In addition, a dynamic system is introduced to improve the robustness of the algorithm. The rank of the Hankel matrix built by homography matrices in previous frames is minimized to predict the one in following frames. The experimental results indicate that the Hankel Rank Minimization approach is able to effectively improve the registration accuracy and the system robustness.

Future work should be focused on improving the robustness of the plane induced homography method. First, multiple lines on the road could be taken into account to define the road plane rather than only two lines which might be detected inaccurately by the Hough transform, and then least square can be used to average the result. Then, before estimating the fundamental matrix between the two frames, the corresponding points could be optimized to minimize the geometric error subject to the epipolar constraint. In addition, the Hankel Rank Minimization approach should be implemented on the real road data to check if it is able to make correct prediction as well as the synthetic data.
Bibliography


