Haaga-Helia

Panorama image stitching techniques

Du Nguyen

Haaga-Helia University of Applied Sciences

Bachelor's Thesis

2022

Bachelor of Business Information Technology

Abstract

Author(s)
Du Nguyen

Degree
Bachelor of Business Information Technology

Report/thesis title
Panorama image stitching techniques

Number of pages and appendix pages
37 + 2

Images has been around for as long as almost 200 years and we have came far since then. The first photograph ever produced was in 1826 by Joseph Nicéphore Niépce and the first digital image was produced in 1920. In short, digital imaging is a representation of characteristics of an object visually.

Besides communication through languages, the information in the form of images plays a very important role in exchanging information. In the information technology field, image processing and graphics have occupied a very important place because its unique properties set it apart from other fields.

Today, most conventional cameras despite high resolution but can only record a portion of objects of size as large as a park or a city. So, the question is how to combine all those small pictures into one large picture? The large image that fully displays the objects of that size. This is also the reason why the author chose the topic of panorama photo stitching based on matching feature as the project of his thesis. The thesis will also be covering the topic of processing, compression, storage, printing and display of the aforementioned images.

Keywords
Image, panorama, process, techniques

Table of contents

1 Introduction

Digital image processing has many practical applications. One of the earliest applications was processing images from the Ranger 7 mission at the Jet lab Propulsion in the early 1960s. Spacecraft-mounted imaging systems had some size and weight restrictions, so the image received has reduced quality such as blur, geometric distortion, and background noise. Those images were successfully processed thanks to computers. Images of the Moon and Mars that we see in journals were all processed by computers.

Besides communication through languages, the information in the form of images plays a very important role in exchanging information. In the information technology field, image processing and graphics have occupied a very important place because its unique properties set it apart from other fields. We know that most of the information that humans gather through sight is from images. Therefore, processing images and graphics is an important part important in the exchange of information between humans and machines.

In today's modern life, robots play an increasingly important role in industrial field and as well as at home. They can be great at doing boring or dangerous work, and jobs where speed and precision are beyond human capacity. As robots become more sophisticated, computer vision will play an increasingly important role. People will ask for computers not only detect and identify industrial parts, but also understand what they see and take appropriate action. Image processing will major impact on computer vision.

Other applications of image processing are extremely diverse. In addition to these applications discussed above, also included other fields such as household electronics, astronomy, biology, physics, agriculture, anthropology, etc…

Image processing is also used in photo collages to create images that has a width and depth that the camera does not usually allow for angles to be so wide.

Today, most conventional cameras despite high resolution but can only record a portion of objects of size as large as a park or a city. So, the question is how to combine all those small pictures into one large picture? The large image that fully displays the objects of that size. This is also the reason why the author chose the topic of panorama photo stitching based on matching feature as the project of his thesis.

.

2 Overview of digital images

2.1 The concept of digital images

A digital image is a finite set of pixels with a suitable grey level used to describe closest to the real photo. The number of pixels determines the resolution of the image The higher the resolution, the more clearly the features of the image are displayed and make the image more realistic and sharper.
Images can be represented in one of two models: the Vector model or Raster model.
Vector model: In addition to saving storage space, easy to display and printing, vector images also have advantages of easy selection, copy, move, and search... in this model, the vector direction of the neighbouring pixels is used to encode and reconstruct the original image. Vector images are acquired directly from digital devices such as Digitalize or converted from Raster images through vectorization programs.
Raster model: is the most used model nowadays. The image is represented as a matrix of pixels. Depends on need the fact that each pixel can be represented by one or more bits. The Raster model is convenient for acquisition, display and printing. Images used in the scope of this thesis are based on the Raster model.

2.1.1 Pixels

A pixel is an element of a digital image at coordinates (x, y) with grayscale or a certain colour. The size and distance between those pixels are appropriate selection so that the human eye perceives spatial continuum and the grey level (or colour) of the digital image is the closest to the real image. Each element in matrix is called an image element.

2.1.2 Grayscale

Is the result of a corresponding transformation of a point's luminance value image with a positive integer value. Usually it is defined in the range from 0 to 255 depending on the value at which each pixel is represented.

2.1.3 Histogram

Histogram, also known as the grey histogram of an image is a function that provides the frequency of occurrence of each grey level.
The grey histogram of an image with grey levels in the range [0, L-1] is a discrete function $p(r_k)=n_k/n$. Where n_k is the number of r_k grey level pixels, n is the sum number of pixels of the image and k = 1, 2, 3, …, L-1. Plot this function with all values of k will generally represent the occurrence of grey levels of an image. The grayscale diagram of an image can be represented by the frequency with which each grey level occurs on the coordinate

system Oxy. In which, the horizontal axis represents the number of grey levels from 0 to N (number of bits of the grey image), the vertical axis represents the number of pixels of each grey level.

Looking at the histogram, it is possible to know the intensity distribution of an image, or images where the histogram distribution is skewed to the right, that means that the image has good brightness, otherwise the image is a dark image.

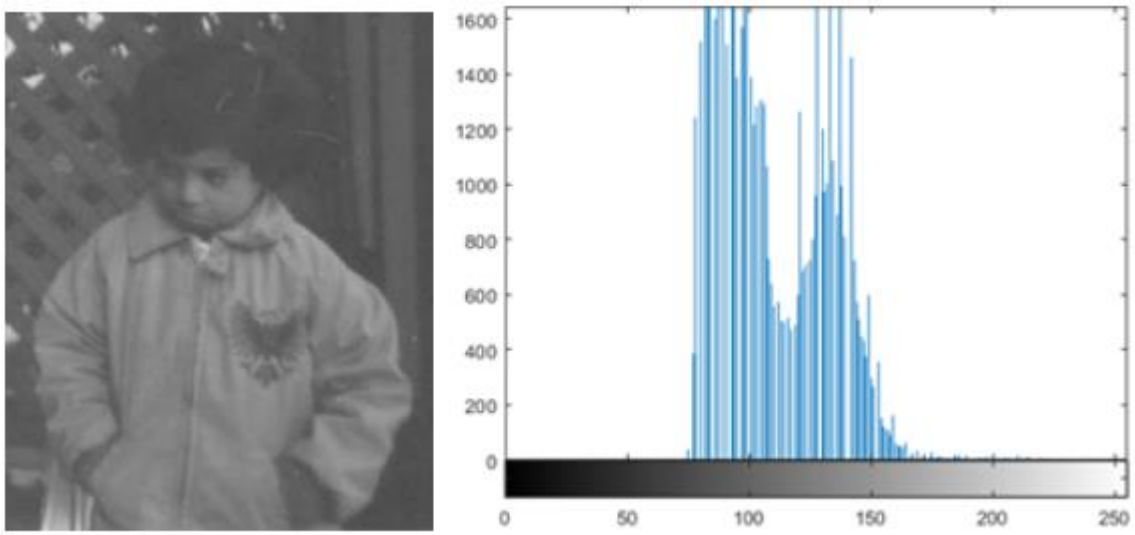

Figure 2.1: The input image is shown on the histogram

2.1.4 Resolution

The resolution of an image is the assigned pixel density on a digital image that is displayed. As shown above, the distance between pixels must be selected so that the human eye can still see the continuum of the photo. The selection of the appropriate distance creates a distribution density, which is the resolution and is distributed along the x and y axes in two dimension.

2.2 Image processing issues

2.2.1 Transform images

The term image transform is often used to refer to a matrix unit and techniques used to transform images. As well as one way signal is represented by a series of basic functions, the image can also be represented under some discrete series of basis matrices called the base image. The basic image equation has the form:

$$A^*_{k,1} = a_k a_l^{*T}$$

Where a_k is the kth column of the matrix A. A is the units matrix. That means $AA^{*T}=1$. The $A^*_{k,l}$ are defined above with k, l = 0, 1, 2, …, N-1 as the base image. There are many types of transformations used such as:

- Fourier Transform
- Kronecker Product
- Karhunen-Loeve

Due to the processing of a lot of information, the multiplication and addition operations in the development is too large, so the above transformations are intended to reduce the dimensions of the image, so image processing is more efficient.

2.2.2 Image display

In image display, it is common to use the feature elements of the image, which are pixels. Image representation models provide a logical or quantitative description of the properties of this function. In image display, attention should be paid to image fidelity or standards for measuring image quality or the effectiveness of the processing techniques.
Digital image processing requires the image to be sampled and quantized. Image quantization is the conversion of an analog signal to a digital signal of an image sampled to a finite number of grey levels.
Some models are often used in image display: Mathematical models, statistical models. In mathematical modelling, two-dimensional images are represented by functions of two orthogonal variables called basis functions. With statistical modelling, an image is considered as an element of a set characterized by quantities such as: mathematics, covariance, variance.

2.2.3 Image analysis

Image analysis involves the identification of quantitative measures of an image to give a complete description of the photo. The most used techniques are image edge detection techniques, such as differential filtering or normal detection plan. In addition, one can also use techniques to image area. From the obtained image, the technique of splitting or fusion is performed. based on evaluation criteria such as: colour, intensity, etc... methods known as Quad-Tree, edge thinning, boundary binary. Finally, there are structure-based classification techniques.

2.2.4 Image recognition

Image recognition is a process involving descriptions of the object that one wishes to specify. The identification process usually follows the extraction of the main features of the object:

- Identification by parameter
- Identification by structure

In fact, humans have applied identification techniques quite successfully with many different objects such as fingerprint image recognition, word recognition (letters, digits, accented letters).

2.2.5 Image compression

Image data as well as other data that need to be stored or transmitted online. As mentioned above, the amount of information to represent an image is a lot. Therefore, reducing the amount of information or compressing the data is a necessity.

Classification of compression methods include:

- Based on the principle of compression:

  Lossless compression: After decompression we get exactly original data.

  Lossy compression: After compression we do not get data like the original.

- Based on how the compression is performed:

  Spatial compression: Direct impact on sampling of the image in the spatial domain.

  Temporal compression: Affect the transformation of the image origin.

- Based on the philosophy of encryption:

  First-generation compression methods: Includes methods that calculation is simple.

  Second generation compression methods: Based on the saturation of the compression ratio.

2.3 Digital images features

In the area of image processing, digital image feature is a piece of digital image information suitable for computational tasks related to a given application.
These features can be special textures in the digital image such as points, edges of an object or an object in the image. On the other hand, the features of a digital image can also be the result of a comprehensive transformation or feature point detection methods applied over the entire image.
A feature in an image is a pixel that contains more information than its neighbours. Feature-based image representation will be more condensed, reducing the search space in applied problems.

2.3.1 Colour features

It is a prominent feature and is most commonly used in image processing applications. Each pixel (colour information) can be represented in 3D colour space. Commonly used colour spaces are: RGB, CIE, HSV ...

Currently, search engines such as Google, Bing ... are all based on colour features to find related images combined with texture and shape features.

2.3.2 Structural features

Texture provides information about the spatial arrangement of colours and intensity in an image. Texture is characterized by the spatial distribution of intensity levels in an area adjacent to each other. Textures consisting of parent textures or aggregates are sometimes called texels.
Texture feature is widely used and very intuitive but not precisely defined because of its wide variability. There are many ways to describe textures: Statistical methods often use space frequencies, event matrices, boundary frequencies, etc.

2.3.3 Shape features

The shape of an image or region is an important feature in identifying and distinguishing images in pattern recognition. Defining the shape of an object is often difficult. Shapes are often represented verbally or graphically, and people often use terms such as round, distorted. Computer-based shape processing is very complex, and while many methods of describing actual shapes exist, there is no universal method for describing shapes. There are two main types of shape features commonly used:

- Boundary-based features: use only the outline of the shape.
- Area features: use the entire area of the shape.

The main goal of shape in pattern recognition is to measure the geometric properties of an object used in object classification, comparison, and recognition.
Shape measures are numerous in the domain of image processing theory. They range from rudimentary global measures that aid in object recognition, to detailed measures that automatically look for distinctive shapes.

2.3.4 Invariant local features

These are the features that do not change when rotating the image, scaling the image or changing the brightness of the image. SIFT is a widely used invariant feature:

- SIFT: stands for Scale-Invariant Feature Transform, is one of the most famous algorithms today used to detect and describe digital image features. This algorithm was published by David Lowe in 1999.
- SURF: stands for Speeded Up Robust Features, introduced in 2006 by a group of researchers including Herbert Bay, Tinne Tuytelaars and Luc Van Gool. Developed based on SIFT algorithm but improved for faster processing speed than SIFT

In SIFT algorithm, finding scale-space is based on approximating Laplace of Gaussian using Difference of Gaussian, while SURF uses box blur, processing speed will be greatly

improved with the use of integrated image. In the direction determination step, SURF uses wavelet response in both vertical and horizontal direction, then the main trend by summing those responses.

2.4 Image matching

2.4.1 Introduction to image matching

Image matching is a problem that has been attracting the attention of researchers and developers[1]. Whenever this problem is solved, it opens up many useful applications such as: image search, object recognition, tracking and detection, image compositing, etc. Matching two images is to find similar areas on two images.

Usually, to match an image, it is necessary to compare the basic elements that make up it. The simplest is to compare pixels. However, this comparison requires a lot of computation time and often does not achieve the desired accuracy.

The first solution to the image-matching problem was proposed by Hobrough in the late 1950s. The first automatic association search system was introduced by the Wild Heerbrugg company in 1964, but it is not widely used. However, Hobrough's idea of applying cross-correlation was used by many people. Since the 1970s, the focus on developing image matching and correlation matching has achieved great success and is applied in the similarity measurement system for images (Helava, 1978). Today, image matching technology is incorporated in many image processing software used as a computational tool. There is a lot of research done with the desire to find two similarities on two photos. Similarity search algorithm can be performed on 2D images.

The main problem of image matching is choosing a suitable object and how to compare it. Pixel-by-pixel comparisons will not be possible with large images because it will require more computation, take more time, or if you want to shorten the time, more powerful processing hardware is required. Furthermore, it leads to inaccuracies because of the repetition of colours with the same grey level value and noise of the image. To solve that problem, instead of pixel-by-pixel matching, which leads to too large input data, we will reduce the input data by including features of both images and then do the matching on the other images with that characteristic.

2.4.2 Image matching by region

This method is also known as the correlation or sample matching method. This method combines feature matching and component matching. The grey intensity of the image is used as the basis for image matching. Since it is impossible to match each pixel of both images, we will instead match a set of neighbouring pixels to reduce the number of times

calculate. In the first image, a window of size m*n is used (usually m = n to easily find the coordinates of the centre point of the window) compared with a "sample" which is also the window. same size in the second photo. The comparisons are performed on the window. In image measurement, cross-correlation and least squares are the techniques used widely used in region-based image matching.

The larger the sample size, the higher the specificity requirement of the matched entity. On the other hand, the geometric distortion caused by image rotation will also affect the matching results of large samples. The entity specificity requirement is also not satisfied if the region is repetitive or the contrast and structure are low (Example: sand, desert, sea water). Areas obscured by other taller objects should also be removed. To obtain acceptable results, the sample size must be small or the shape must adapt to the geometrical deformation.

To avoid faulty in match results, the position of the search window must be specified in region-based matching. The size of the search window depends on the exact position and on the distortion due to the orientation of the image.

After finding the most suitable position, it is necessary to evaluate the accuracy and reliability of the found comparison results. Setting thresholds for matches is one way to minimize skewed matching. In addition to the threshold method, the geometrical adjustment method can be used to calculate and eliminate false matches.

2.4.3 Feature-based matching

Contrary to region-based matching method, feature-based matching method uses abrupt changes in grey level values corresponding to image features as the basis for matching such as edges, corners, or image features. Feature-based matching is superior to region-based matching. The image feature-based matching technique basically consists of 3 main steps:

- Select points as feature points of the image (edges, corners, points) in each image independently.
- Construct a list of possible pairs of points that are similar.
- Perform the match and return the result set of similarities.

Often people will integrate both region and feature matching into the image compositing software to achieve the most accurate results and faster processing speed, less time. With the current development of technology, performing a match on small images takes little time, but for large images, the optimization of the algorithm, improving the execution speed is also a matter of concern.

2.4.4 Interest points

Image feature-based matching works best on high-contrast image areas. Points that can be described by either high disparity in grey level values or with steep gradient are called points of interest. Points of interest should be distinct, invariant to geometric distortion and image quality, and stable. Finding points of interest in an image is done in two steps:

- Calculate the features in each window of the selected image.
- Compare the found value with a given threshold.

The characteristic is different for each different operator but is basically based on the grey level value inside each sliding window. Only windows whose values are greater or less than the threshold is accepted as points of interest. A list of points of interest for each image is matched against its pixel coordinates (the centre point of each sliding window) and their description as a result of the processing.

2.4.5 Edges and regions

Edge can be described as a sudden change in grey level value in a small area. The edge usually corresponds to the edge of the object in the image. The edge extraction process is very complex and involves 3 steps including:

- Determine the pixels lying on the edge, the value of the interrupted grey level will be determined by the average of the edge operators. Whether that point is determined to be on the edge is based on the result of comparing the grey level value with a given threshold.
- Connect the pixels together and make the border.
- Group edges together, segment.

The edge operator will detect the change of the grey level value in the image, based on the first derivative to find the extreme and locate the edge point. Some edge operators can be used such as Sobel Operator, Prewitt Operator. The Sobel operator will be less affected by noise in the image because neighbouring pixels are included.

The Laplacian operator is based on the second derivative. In order not to be affected by noise, it is combined with the Gaussian operator to smooth the image and remove noise.

3 Panorama matching based on characteristics

3.1 Photo stitching overview

3.1.1 Photo stitching introduction

Photo stitching is the process of combining multiple small images stacked on top of each other to create a larger, higher-resolution photo. Usually, image compositing is done using computer software.

Photo stitching has many different applications. The most traditional application is to create wide-space and satellite images from a set of images, to construct geographic maps, to combine images taken on the surface of a star into a single high-resolution image. larger resolution, etc…

The main problems in image stitching are aligning component images, correcting distortions, changing colours, and blurring borders between images. All these operations are intended to make the images look like a single image, rather than a composite of several small images.

The stitching of the components of the objects together to obtain more complete corresponding images is a very difficult task when done manually, on the other hand, the images acquired for stitching are often skewed and deformed to some extent. The requirement is to determine the error of information between the parts of the image to be merged, then correct the deviation and finally put them together.

3.1.2 Photo mosaic

Photo mosaic is the creation of a new image by merging small images into a large image so that when looking at the whole picture, we can still see the content of the previous large image.

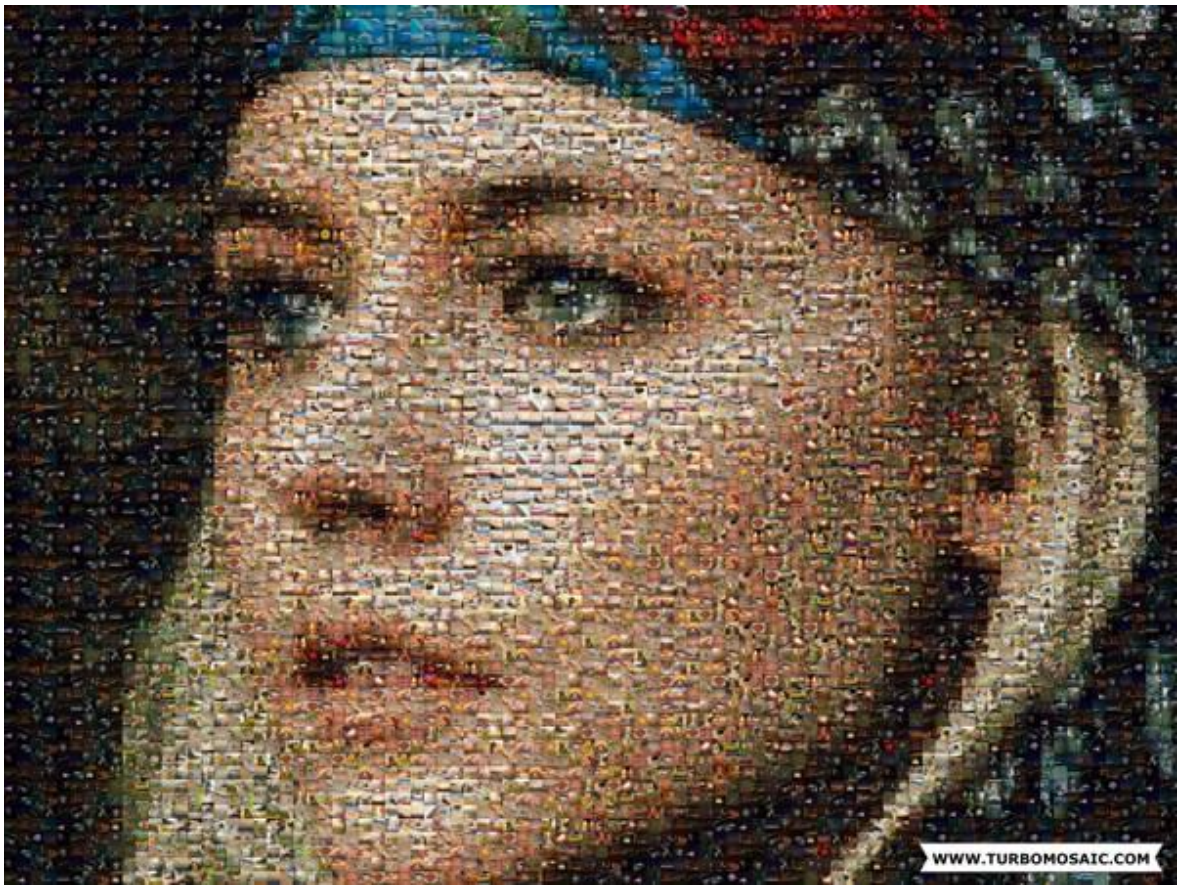

Figure 3.1: Example of photo mosaic

From the original overall picture, by different processing techniques, small images are integrated into it to create a new image. Of course, if viewed as a whole, it is still the same original big picture, but it is slightly different because the details inside have been replaced by single images.

3.1.3 Panorama

A panorama is a wide view of an object in space. It allows the representation of a wide view of paintings, graphic drawings, photographic art, film or video, or 3D models.
The term panorama appeared before we had panorama cameras. The origin of the word panorama was identified by an Irish painter Robert Baker used to describe large-scale paintings in Edinburgh. These panoramas are wrapped in a cylindrical tube and slowly pulled out to reveal the picture.
In 1881, the Dutch painter - Hendrik Willem Mesdag created the Panorama Mesdag school with cylindrical tubes that scrolled panoramas with huge sizes, 14m high and can be from 40 to 120m long. In the 19th century, there were two panorama paintings that are considered the largest of this period, which is a painting depicting the Battle of Atlanta with a height of nearly 13m and a length of 110m. The largest identified painting is in Wroclaw (Poland) with dimensions of 15m x 120m.

Due to human needs and the development of science and technology, people have created panorama cameras. If a normal camera can only take pictures with a 90-degree angle, a panorama camera can take pictures with an angle of 175 degrees, 180 degrees or 360 degrees. In front of a large space, the camera is often powerless to record images at a wide angle, but the panorama camera does its job. Panoramic cameras are usually taken with positive film (also known as film slide). After taking a photo, you can watch the film to know how the image will be printed.

Because the image angle of the panorama is wide, the panorama camera does not have a long lens like a normal camera. The lens of the panorama camera has an arc shape. When shooting, the lens will scan from left to right, so we have to use a tripod when shooting.

Figure 3.2: Example of a panorama camera

Panorama photos simply mean viewing images with a wider angle of view than normal photos, i.e., extremely large picture frames that a frame taken with a camera cannot fully show. The image is stitched together from digital photographs of parts of a landscape (these scenes overlap) into a complete panorama.

We can simply understand that panorama is a wide format photography mode by taking many consecutive photos, with the information of the previous photo being partially shown in the following photo, to assist users. Then with the help of image processing software, we will get a wide-format image.

Figure 3.3: Example of a panorama picture

3.1.4 Image acquisition in Panorama

The first stage of image compositing requires the selection of a suitable shooting position so that the image is least subject to geometrical distortion. It is necessary to clearly determine the type of panorama image to be combined to choose the appropriate shooting method.

Different image acquisition methods can be used so that different input images can be obtained from which different types of panoramas can be generated. There are 3 ways to capture images:

- The camera is placed on a tripod and we rotate the camera while taking pictures to get the input image.
- The camera is mounted on a skateboard, the input image is obtained by both moving the board and shooting. The advantage of this method is to ensure the stability and accuracy of the input image, to ensure that there is no or little geometrical variation of the image, the input images are on the same line.
- The photographer directly holds the camera in his hand and takes the picture by rotating or walking in a direction perpendicular to the direction of the camera. The disadvantage of this shooting method is that the input image may be distorted due to the impact of the photographer such as shaking, the image is tilted, and it is not in a straight line.

All three shooting methods above need to ensure that the following image must contain about 15% or more of the content of the previous image to ensure the determination of the position of the two images and should try to limit as much as possible. The image is transformed, resulting in an inaccurate composite result.

The method of acquiring images uses a camera placed on a skateboard and moved with a direction parallel to the plane to be photographed, the camera is placed in front of the subject to be photographed, and the image is captured by moving the skateboard and shoot to a desired limit.

Figure 3.4: Example of a panorama camera on a slide

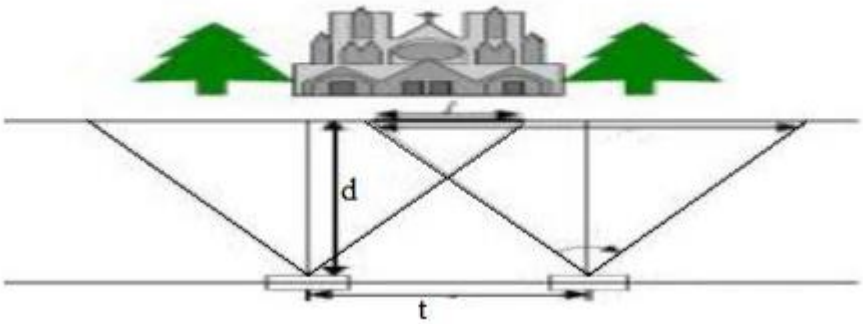

Figure 3.5: Model of a camera with a skateboard

Where t is the camera's slid distance between two shots, d is the distance between the camera and the subject being captured.

Make sure that the direction of the camera slide is parallel to the plane containing the subject to be photographed, otherwise the size of the subject will be changed between the two images.

However, the disadvantage of this method is that the image after being stitched will not give the viewer a real feeling.

The method of using a handheld camera is relatively easy to implement. Users only need to hold the camera and shoot while rotating or moving perpendicular to the shooting direction. However, the resulting image is more difficult to combine due to impacts such as tilting, shaking, etc.

In case the user takes a photo by turning his body, then the photographer acts as a tripod, but there will still be deviation due to unwanted impact.

In case the user takes a photo by moving parallel to the plane containing the object to be photographed, then the photographer acts as a slide. However, then it will be difficult to ensure a stable distance from the camera to the plane containing the subject being photographed.

3.1.5 Image transformation in Panorama

The process of changing the geometry of an image to match the previous adjacent image to form a panorama. The photos should be arranged in the correct order of taking before and after to ensure the highest possible accuracy. Image transformation is the most important process in panorama technique. The process includes three steps:

Step 1: Transform the image to a specified projection space

There are many different projection spaces such as spheres, cylinders. Projecting the image onto a projection screen helps to present the images in a more realistic way. It can also be understood as a board that we will combine the photos into a panorama by pasting these captured photos on the board and using image transformations to change the image so that it fits the previous adjacent image.

Step 2: Align the images

In panorama stitching technique, image alignment is one of the most important parts. To be able to align the images, it is necessary to identify the similarity points between the two images and make the distance between those two points as short as possible, even zero (in case of overlap). To be able to do that we need to go through many different steps. Key point can be considered as an important piece of information extracted from an image and is the most prominent and clearest part of the image. The key points will not be changed by image distortions such as rotation or acceleration … The number of key points must be large enough that a transformation model between the two images can be calculated.

Depending on the type of key points so we can decide to use the appropriate method of determination. Key points can be identified based on techniques such as Harris corner detector, edge detection, etc.

- **SIFT (Scale-invariant feature transform):** An algorithm for describing local features based on points of interest and invariant to image scaling or rotation, unaffected by brightness and noise in the image.
- **Calculation of the identity matrix:** The identity matrix between two images can be calculated using RANSAC (Random sample consensus). An identity matrix is a matrix that represents the transformation of one image relative to another.

In case the input image is not in order or is a component image of many different panoramas, it is important to identify each pair of images to combine. The returned results can be many different panorama images as shown in **Figure 2.7** M. Brown and D. G. Lowe called this technique as recognizing panorama.

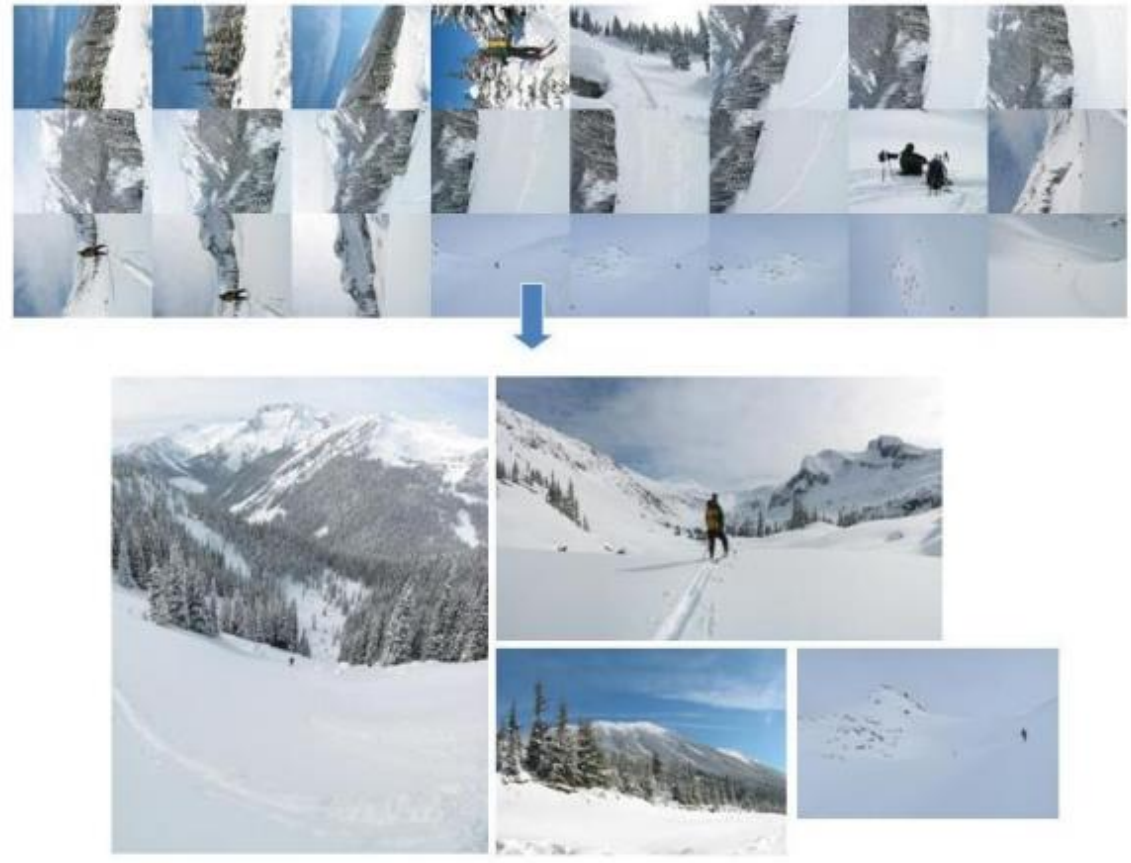

Figure 3.6: Example of recognizing panorama

Step 3: Projecting the photo

The solution is to select an image as the centre and transform the other images according to that image. It is possible to project the images onto a flat surface, then a flat panorama image will be obtained. Alternatively, a cylindrical projection (Szeliski 1994, Chen 1995) or a spherical projection (Szeliski and Shum 1997) can be used.

3.1.6 Mixing images in panorama

After stitching the images, the result is a panorama. However, due to external influences such as light and exposure, when taking an input image, it will lead to a difference in colour brightness between two similar image areas between two images, so when merged, the stitched part will be clearly seen. together as shown in **Figure 2.10**. So, it is necessary to balance the brightness of the merged part of the two images to reduce the clarity of the joined area as shown in **Figure 2.11.**

Figure 3.7: Example of panorama that hasn't been combined

Figure 3.8: Example of panorama that has been combined

3.1.7 Cropping images in panorama

Cropping is a technique used to remove redundant objects or areas from the resulting image.

Figure 3.9: Panorama before cropping

Figure 3.10: Panorama after cropping

3.2 Panorama stitching techniques

3.2.1 Panorama stitching based on colour histogram matching

Transforms an image so that its colour histogram matches a specified histogram.
Given two images, the reference image and the final image. We calculate the histogram for two images, the reference image is $\mathbf{F_1()}$ and the target image is $\mathbf{F_2()}$. Then for each $\mathbf{G_1}$ grey level value between 0 - 255, we find the equivalent grey level value $\mathbf{G_2}$ such that $\mathbf{F_1(G_1) = F_2(G_2)}$ from which we get the result $\mathbf{M(G_1) = G_2}$. Finally, apply the $\mathbf{M()}$ function to each pixel of the reference image.

3.2.2 Panorama stitching based on texture analysis

Image texture is a set of indices computed in image processing designed to determine the amount of perceived texture of an image. Texture images provide information about the spatial arrangement of colours or intensity in an image or selection of an image.

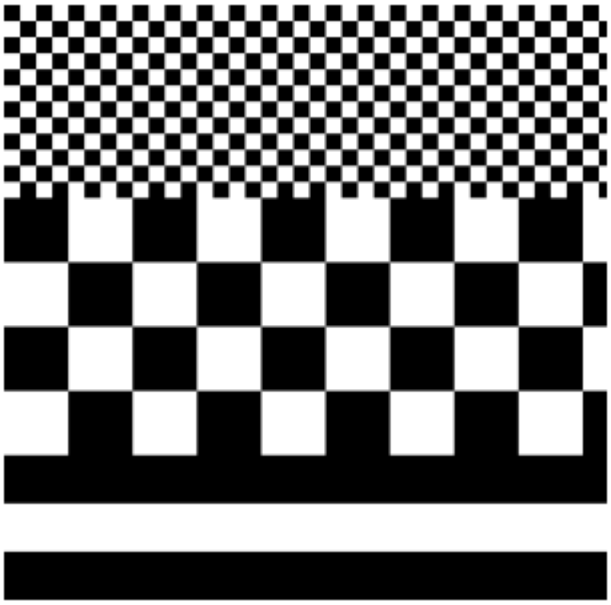

Figure 3.11: Examples of artificial textures

Figure 3.12: Examples of natural textures

3.2.3 Shape analysis

The use of computers to detect objects of similar shape in a database or matching parts. In order for computers to be able to analyse and process geometry, objects must be represented as numbers. Cubic analysis is applied in many fields such as: archaeology to find similar objects or missing parts, architecture to identify objects that spatially conform

to a particular space, medical imaging to understand shape changes related to disease or aid in surgical planning, virtual environments or on 3D models to identify subjects for copyright purposes, security applications such as facial recognition, the entertainment industry (movies, games) to build and process geometric models or animations, design computer-aided and computer-aided manufacturing to process and compare designs of mechanical parts or design objects.

3.2.4 Panorama stitching based on geometric correction

Determine the distortion of the first image compared to the second image, this image can be translated, scaled to a certain ratio. The job to do is to transform and correct this distortion to a minimum.

3.2.5 Panorama stitching based on image feature

The algorithm uses pairs of similarities as a result of the algorithm matching the features of both images, thereby building a similarity matrix to be able to "project" the image onto a plane in space.

3.3 Panorama stitching based on invariant features

3.3.1 Extract invariant features of an image

One of the most basic prominence search methods is the search for boundary floats, points on the curve that have maximum curvature, or corner points.
These points are initially detected by the sharpness of the boundary: the boundary of the object is stored as a chain code; the angle is discovered through finding places on the boundary that bend significantly. This angle detection technique is very complex and multi-step implementation.
Harris corner detector uses a window that can slide in any direction using Gaussian function and Taylor expansion.
Conceptually, the Harris corner detector will search for large changes in grey intensity in different directions by using a small window to do the task of checking and detecting points defined as corners.

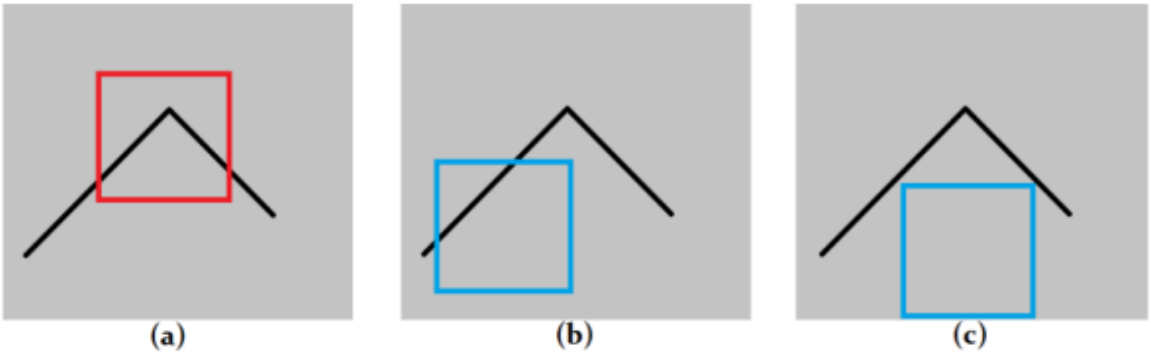

Figure 3.13: Harris corner detector

In Figure 3.13(a) the sliding window is in the area of the image containing the corner, when moved in any direction there is a change in the grey intensity.
In Figure 3.13(b) the sliding window is on the image area containing the edge, when moving the sliding window in either direction of the edge there is no change in grey intensity.
In Figure 3.13(c) the sliding window is on the non-angled image area, after moving the sliding window there will be no change in grey intensity.
Based on this we can detect which point is corner point and which point is not.

3.3.2 Feature extraction algorithm

Suppose for a grey image **(I),** for each point **(u, v)** and displacement **(x, y)** we can calculate the average change in grey intensity by a shift window from **(u , v)** to **(u + x, v+ y)** as follows:

$$S(x,y)=\sum_u\sum_v w(u,v)(I(u+x,v+y)-I(u,v))^2$$

Where S(x, y) is the sum of squares of deviation values, also known as grey intensity changes at (x, y) and W(u, v) is the window at (u, v)
I(u, v) and **I(u + x, v + y)** are the grey intensity values of pixels at positions **(u, v)** and **I(u + x, v + y)**
The value **I(u + x, v + y)** can be expanded according to Taylor's formula as follows:

$$I(u+x,v+y)\approx I(u,v)+I_x(u,v)x+I_y(u,v)y$$

With I_x, I_y is the derivative with respect to components x, y.
From there, (2. 2. 1) can be rewritten as:

$$S(x,y)=\sum_u\sum_v w(u,v)\left(I_x(u,v)x-I_y(u,v)y\right)^2$$

Expressed in matrix form, then **S(x, y)** we have:

$$S(x,y) \approx (x,y)\, A \begin{pmatrix} x \\ y \end{pmatrix}$$

In which, **A** is as follows:

$$S(x,y) = \sum_u \sum_v w(u,v) \begin{bmatrix} I_x^2 & I_x I_y \\ I_x I_y & I_y^2 \end{bmatrix} = \begin{bmatrix} \langle I_x^2 \rangle & \langle I_x I_y \rangle \\ \langle I_x I_y \rangle & \langle I_y^2 \rangle \end{bmatrix}$$

Let λ_1 and λ_2 be eigenvalues of **A**, and **k** is a constant, usually in the range **[0. 04, …, 0.15]**.

Then the following expression will decide whether the window **w** contains an angle or not:

$$M_c = \lambda_1 \lambda_2 - k(\lambda_1 + \lambda_2)^2 = \det(A) - k\left(trace^2(A)\right)$$

If both λ_1 and λ_2 are small. That is, the function **S(x, y)** is almost unchanged in any direction. Then the image area inside the window has almost no change in intensity. That is, in this case the corner point is not found.

If λ_1 is large and λ_2 is small, or vice versa. That is, the function **S(x, y)** has a small change if the window slides in one direction, and a significant change if it moves in the orthogonal direction. This shows that there is an edge.

If both λ_1 and λ_2 are large. It means that the function **S(x, y)** has a significant change in grey intensity when shifting the sliding window in any direction. This shows that a corner point exists.

3.3.3 Determine the direction for the feature

By assigning an orientation to each feature based on local image attributes, the feature descriptor can be represented relative to this orientation and thus achieve invariance to rotation phenomena. The measure of feature is used to find a Gaussian filtered image L with the closest size such that all computations will be performed in the same measure invariant.

3.3.4 Describe the features

The above has performed detection and assigned coordinates, size and direction for each feature point. Those parameters require a reproducible 2D local coordinate system to describe the local image area and thereby induce invariance to those parameters. This step will compute a descriptor for a maximal image region that is highly characteristic (invariant with different changes in brightness, zoom-in, rotate).

One simple approach is to sample the local image densities in the vicinity of the feature at an appropriate measure and match these densities using the standard correlation measure.
A better approach is suggested by Edelman, Intrator and Poggio (1997). This approach is based on a biological vision model, namely a complex neuron model in the brain system. The neurons will correspond to a gradient at a specific spatial frequency and direction, but the position of the gradient on the retina is allowed to slide over a small area of the viewport.

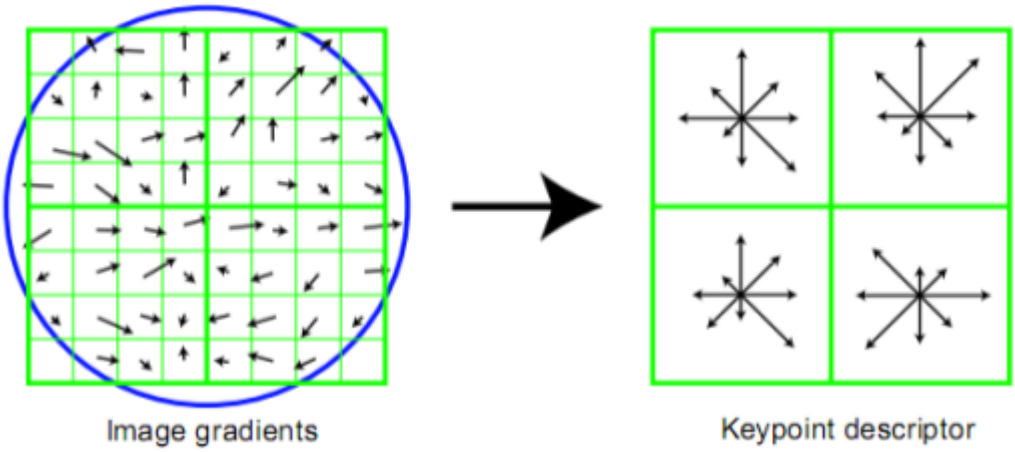

Figure 3.14: Local descriptor

The left image is a simulation of the gradient amplitude and direction at each sample in a neighbourhood with the highlight. Those values are centred in a Gaussian window (located inside the circle). These samples are then aggregated into a direction histogram that briefly describes the content in the 4x4 subregion as depicted on the right with the length of each row corresponding to the sum of the gradient amplitudes near that direction within a region.
After determining the direction, it will be represented as vectors 4x4x8 = 128 dimensions (Number of dimensions = 8 directions x (4x4) = 128 dimensions) by summing the orientation vectors of points in area, these vectors have the following characteristics:

- Common root.
- The length of each vector corresponds to its gradient magnitude m.

3.4 Matching invariant features

3.4.1 Distance measure and similarity measure

Similarity is one of the good methods for computers to distinguish images by their content. Content-based matching will query images by similarity measurement method based on features, its determination can be in many forms such as edge detection, colour, pixel

position..., methods like histogram, colour and analysis use histogram to determine similarity.
Therefore, metric is important in content-based image matching. The measure has the meaning to decide how the match result will be, how accurate it is. Many distance measurements have been exploited in image matching.

3.4.2 Matching local characteristic

The match will be performed on the key point sets found. The main step in the matching technique is to find a subset of key points that match in two images, to do this, to find pairs of matching key points in the two images. The subset of matching key points is the similar image region. The matching of two feature sets refers to the problem of finding the nearest neighbour of each feature point (Figure 2.17).
A method proposed by D. Mount that allows quick searching of the used neighbourhoods, ANN stands for Approximative Nearest Neighbour. It allows to organize data as kd-tree. Specifically, two points in the feature space are the same if the Euclidean distance between the two points is the smallest and the ratio of the nearest distance to the second nearest distance must be less than some threshold.
Assume the key point pair has descriptors of:
$A = (a_1, a_2, a_3, \ldots, a_{128})$ and $B = (b_1, b_2, b_3, \ldots, b_{128})$
Then the Euclidean distance between A and B is calculated by the formula:
$D(A, B) = \Sigma(a_i - b_i)^2$

3.5 Homography matrix

3.5.1 Homography introduction

In mathematics, Homography is a displacement using a geometric projection, or in other words it is a combination of pairs of points in a perspective projection. Real images in three-dimensional space can be transformed into image space by projection through Homography transformation matrix, also known as H-matrix. Transformation projections through Homography matrix are not guaranteed in size and shape. angle of the projected object but ensure the ratio.
In the field of machine vision, Homography is a mapping from the object plane to the image plane. Homography matrices are often related to processing tasks between any two images and have very wide applications in image editing, image merging, motion calculation, rotation or displacement between two images.

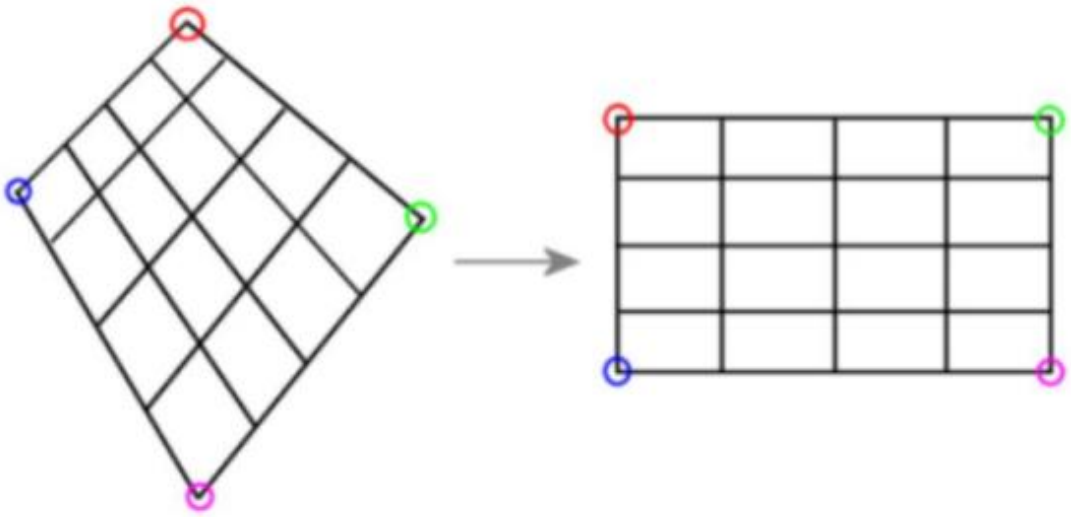

Figure 3.15: Homography projection

3.5.2 Homography calculation

Homography is a mathematical definition. It is the displacement using the geometric dimension, or in other words it is the combination of the pair of points in the perspective projection. Real images in three-dimensional space can be transformed into image space by projection through the Homography transform matrix H. Transformation projections through the Homography matrix, although not guaranteed about the size and angle of the projected object but guaranteed about the ratio.

To calculate the Homography matrix from the corresponding pairs of points, we use Direct Linear Transform which consists of 2 steps: First, from the corresponding pairs of points, we convert to the matrix $\mathbf{A_i h = 0}$. Then, apply the decomposition. decay SVD to calculate the matrix **H**.

The SVD algorithm, also known as the Singular Value Decomposition algorithm, was published by Golub and Kahan in 1965, which is a decomposition technique used to reduce the rank (or dimension) of matrix.

SVD allows the analysis of a complex matrix into three component matrices. The purpose is to provide solutions to problems involving large matrices, complex to smaller problems.

$$A = USV^T$$

Where U is an orthogonal matrix of order m*n (m number of index words) the line vectors of U are the index word vectors. S is a r*r copy matrix with degenerative values σ1≥ σ2≥ …≥ σr with r = rank(A). V is an orthogonal matrix of order r*n – the column vectors of V are the basic vectors. The rank of matrix A is the positive numbers on the diagonal of matrix S.

DLT method with key points found from Harris algorithm:

In non-uniform coordinates, we have the formula:

$$c\begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = \begin{bmatrix} h_1 & h_2 & h_3 \\ h_4 & h_5 & h_6 \\ h_7 & h_8 & h_9 \end{bmatrix} \begin{bmatrix} x \\ y \\ 1 \end{bmatrix}$$

Divide the first line of the above formula by the third line and the second line by the third line, respectively, we have:

$$-h_1x - h_2y - h_3 + \left(h_7x + h_8y + h_9\right)_u = 0$$

$$-h_4x - h_5y - h_6 + \left(h_7x + h_8y + h_9\right)_u = 0$$

Writing in matrix form, we have:

$$A_ih = \begin{bmatrix} -x & -y & -1 & 0 & 0 & 0 & ux & uy & u \\ 0 & 0 & 0 & -x & -y & -1 & vx & vy & v \end{bmatrix} \left(h_1h_2 \ldots h_9\right)^T$$

Applying the SVD decomposition formula to the matrix **[A]** we have:

$$A = U\sum V^T = \sum_{i=1}^{9} s_i u_i v_i^T$$

With $\mathbf{s_i}$ being single values and sorted in descending order, $\mathbf{s_9}$ is the smallest value. Then the value of $\mathbf{h_i}$ is equal to the last value of column $\mathbf{v_i}$.

In fact, the input images may have the origin in the left corner of the image, or the origin may be in the centre of the image. Leaving this condition will affect the results of later transformations such as when multiplying the image by a factor or affin-like transformations. The images need to be normalized by rotation and displacement transformations.

RANSAC (RANdom SAmple Consensus) was published by Fischler and Boller in 1981. The main idea of RANSAC is as follows: From the initial data set, we will have two types of noisy and noiseless data (outlier and inlier), so we have to do calculations to find the best model for the data set. The calculation and selection of the best model will be repeated **k** times, with the value of **k** chosen so that it is large enough to guarantee the probability **p** (usually falling to the value 0.99) of the sample data set. random contains no noisy data.

The process of implementing RANSAC algorithm is described as follows:

From the input data set including noise and no noise, we choose from n random data, minimum to build the model:

- Proceed to build a model with those n data, then create a threshold to test the model.
- Call the original data set minus the n data set to build the test data set model. Then, verify the built model with the validation data set. If the result obtained from the model exceeds the threshold, then the point is noise, otherwise it will be the opposite.
- This process will be repeated for **k** times. Where **k** is calculated according to the above formula. At each loop the value of **k** will be recalculated.
- As a result, the model with the most noise-free data will be selected as the best model.

In the problem of creating Panorama images, the Homography matrix is calculated from the set of pairs of corresponding highlights of the two original images that were compared in step two. Then, the corresponding four pairs of salient points are not collinear, the equation **Ah = 0** according to the normalized DLT method described above. In which, **A** is a matrix of size 8*9. From that, we can determine matrix **h**.

With the Homography matrix calculated from four pairs of random points, we have d as the distance measuring the closeness of the compared pairs of points. With the pair of similar highlights **(x, x')** and **d($\vec{a}$, $\vec{b}$)** as the distance of the two vectors, we have the following distance formula:

$$d = d\left(\vec{x}, H\vec{x'}\right) + d\left(\vec{x'}, H\vec{x}\right)$$

Detailed algorithm:

- Initialize number of loops **k**, **distance**, **max$_{inlier}$**, and **p = 0**.
- **for(i = 1:k)**, do the following steps:
- **Step 1:** Choose 4 pairs of random similarities
- **Step 2:** Check if the points are on the same line.
- **Step 3:** Calculate the Homography H matrix from 4 points using the normalized DLT method.
- **Step 4:** Calculate the distance d of the pairs of similar highlights
- **Step 5:** Calculate the number of m pairs of non-random points (inlier) that satisfy the condition: d_i < distance
- **Step 6**: If m > max_{inlier} then max_{inlier} = m, Homography matrix H = $H_{current}$.
- Continue to recompute the matrix H for all similarity pairs that are considered to be inliers by DLT method.

3.6 Image stitching based on Homography

After the Homography matrix is calculated, the final step in creating a Panorama image is to blend the two images together. The idea of this step is to use one image as the centre, and then use the Homography matrix to project the other image to the central image plane.

As shown in Figure 3.16, on the left is the central image used as the projection plane. On the right is the image projected onto the first image plane using the Homography matrix. The gray part is the part of each image, and the black part is the common part of the two pictures. To perform this technique, we use perspective projection to project the original

input images using the homography matrix as the transformation matrix to project onto the projection plane.

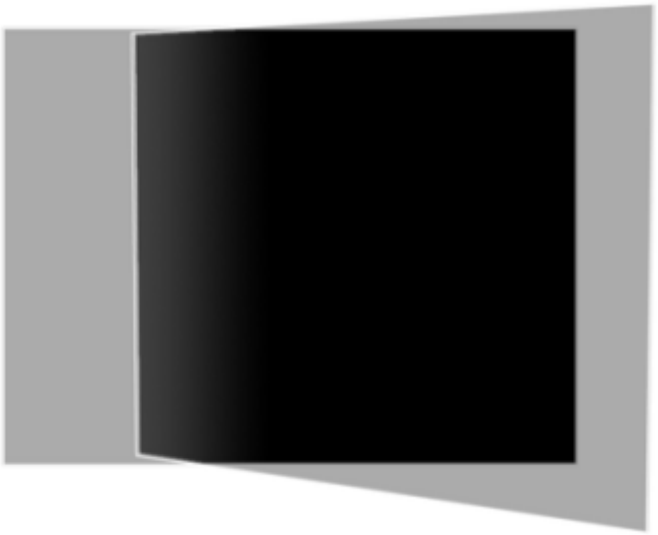

Figure 3.16: Image stitching illustration

Projection is the transformation of points in an n-dimensional coordinate system into points in a coordinate system of less than n-dimensionality. A point P on an object defined in real world coordinates at position (x, y, z) through a perspective projection t is obtained by point P' whose coordinates (x', y', z') lie on projection plane.

4 Creating panorama images

4.1 Image stitching softwares

MATLAB stands for “MATrix LABoratory”, invented by Cleve Moler in the late 1970s, and then chair of the computer department at the University of New Mexico.
MATLAB, originally written in the Fortran language, was until 1980 an internally used part of Stanford University.
In 1983, Jack Little, who studied at MIT and Stanford, rewrote MATLAB in C language and it built additional libraries for designing control systems, toolbox systems, and simulations, etc.. Jack built MATLAB into a matrix-based programming language model.
Steve Bangert is the man who wrote the interpreter for MATLAB. This work lasted nearly 18 months. Later, Jack Little combined with Moler and Steve Bangert decided to turn MATLAB into a commercial project - The MathWorks company was born at this time - in 1984.
The first version MATLAB 1.0 was released in 1984 written in C by MS-DOS on PC and was first released at the IEEE Conference on Design and Control in Las Vegas, Nevada.
MATLAB was originally developed to help students use two libraries LINPACK and EISPACK for two linear algebras without knowing Fortran programming.

4.2 MATLAB interface

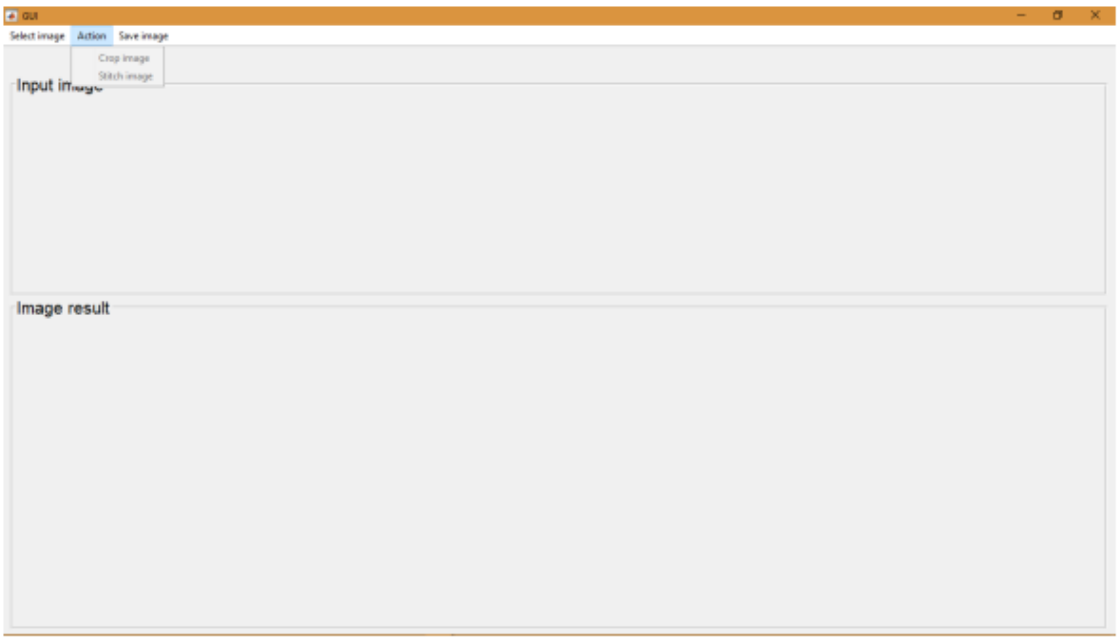

Figure 4.1: The main interface of the program

The interface of the program includes:

- **Select image:** Allows to select the input image set.
- **Crop image:** Allows to crop the image.

- **Stitch image:** Perform image stitching.
- **Panel input image:** where the input image is displayed.
- **Panel image result:** where the resulting image is displayed, including the panorama image.
- **Save image:** Save the resulting image to the computer.

4.3 MATLAB trial run

In the program, we will conduct an experiment to combine three input images, which are Figure 4.2, Figure 4.3 and Figure 4.4 .

Figure 4.1: First input image

Figure 4.3: Second input image

Figure 4.4: Third input image

After clicking the “Select image” menu, the image selection dialog box will appear.

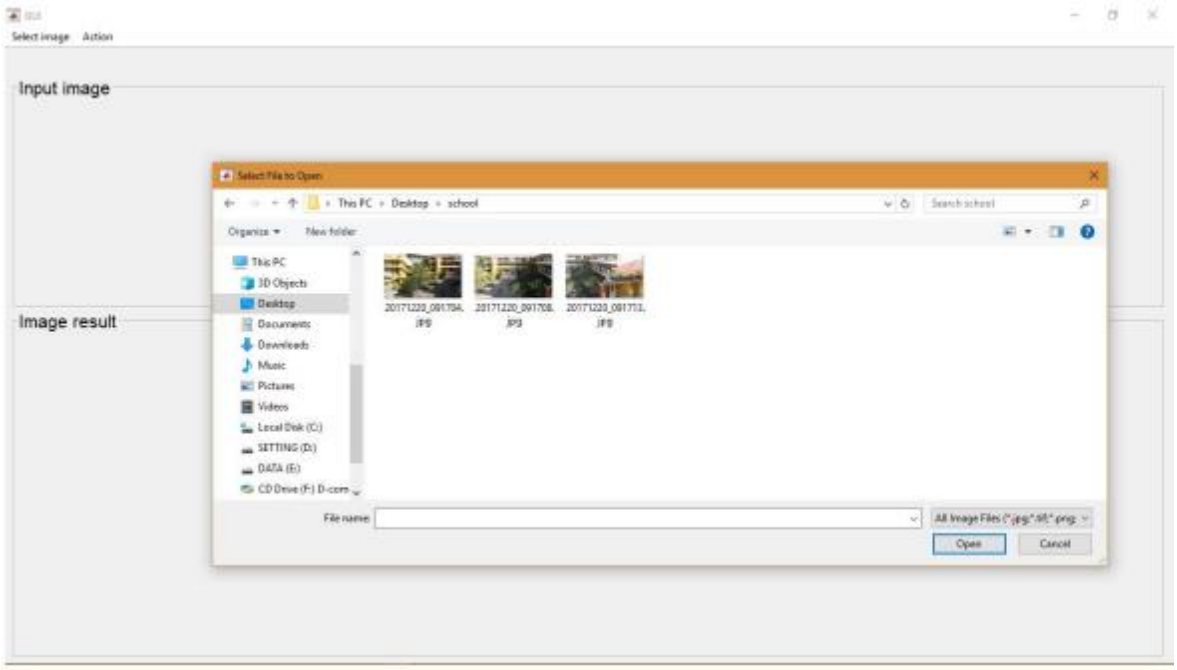

Figure 4.5: Selecting images

After selecting the image, the selected image will be displayed on the "input image" area.

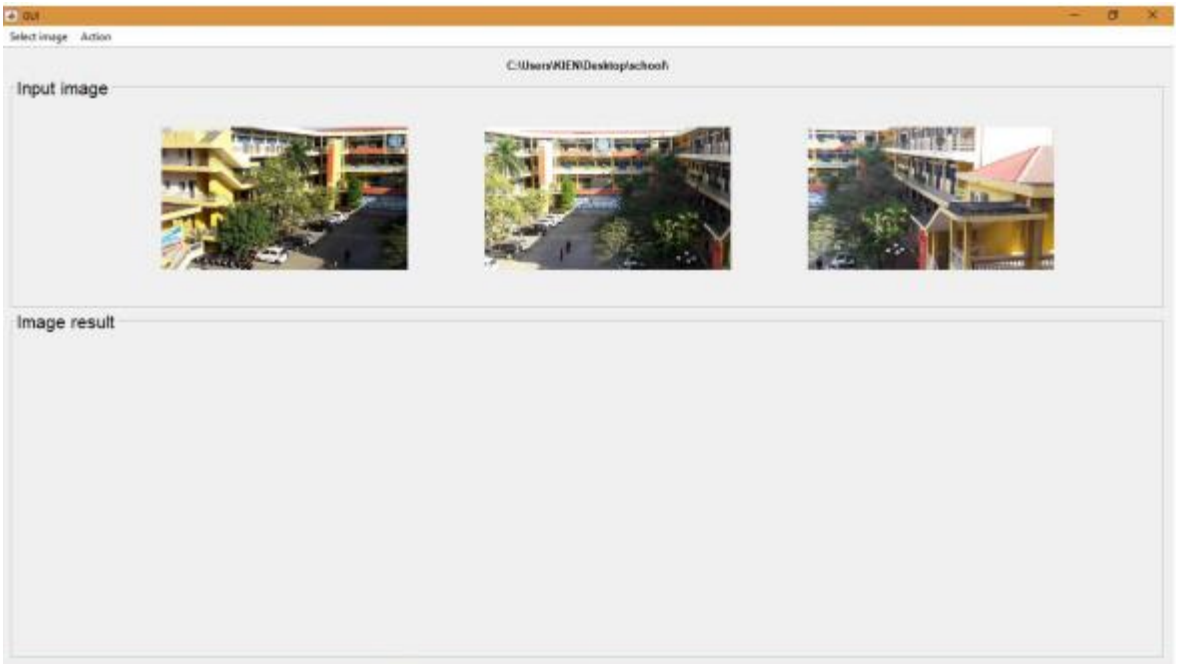

Figure 4.6: Selected images are displayed

Then click the “Action” menu and select “Stitch image” and the program will begin to stitch the image.

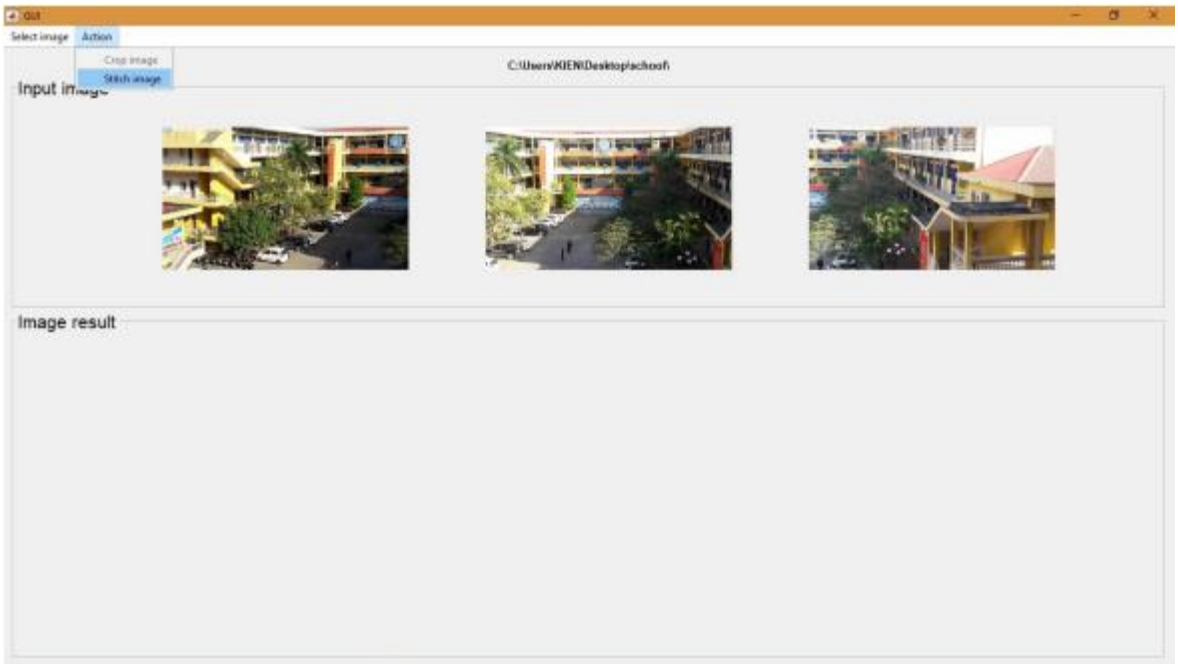

Figure 4.7: Selected images are displayed

The program will first find the angles in two images based on the Harris algorithm. We have the result as shown in Figures 4.8, 4.9 and 4.10 with the corners marked as green pixels.

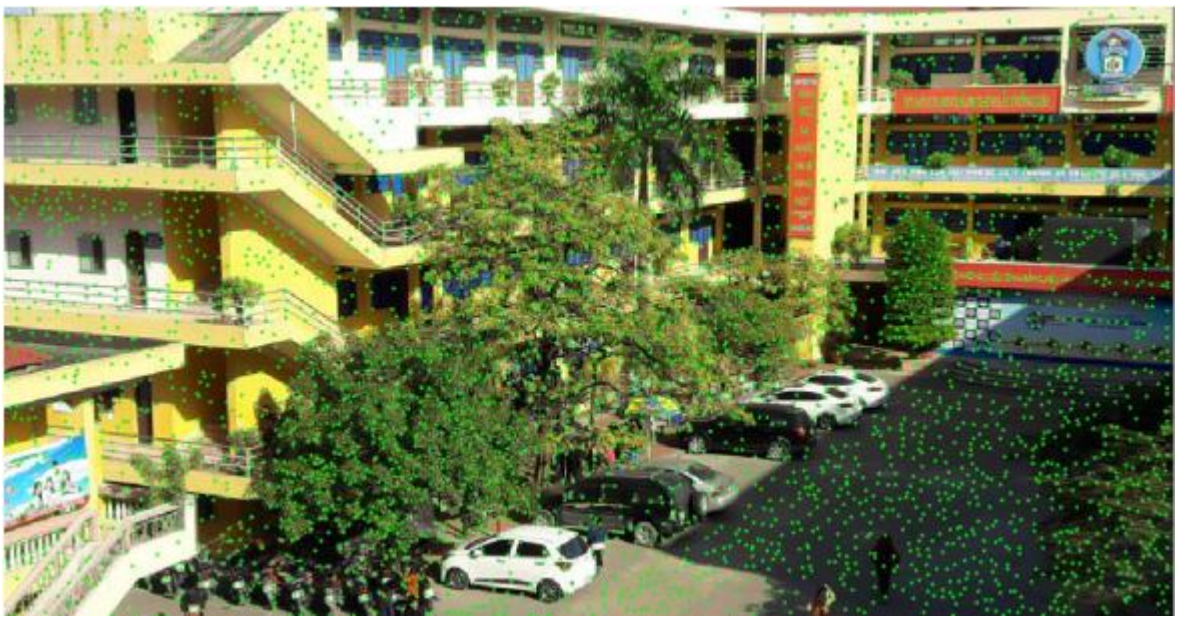

Figure 4.8: Corner search results for the first input image

Figure 4.9: Corner search results for the second input image

Figure 4.10: Corner search results for the third input image

Next, after finding the corners in the pictures. The program calculates the description for the corner points, the results returned are three matrices with sizes m*128 and n*128 and k*128, respectively:

- m, n and k rows are the number of angles and also the number of key points in the first, second and third input image, respectively.
- 128 is the dimensionality of the vector used to describe each corner point or key point.

Based on the description of the feature set of the first and second images. We proceed to match two images based on the set of key points. The results of the feature match function are as shown in Figure 4.11 and Figure 4.12

Figure 4.11: Corner search results for the third input image

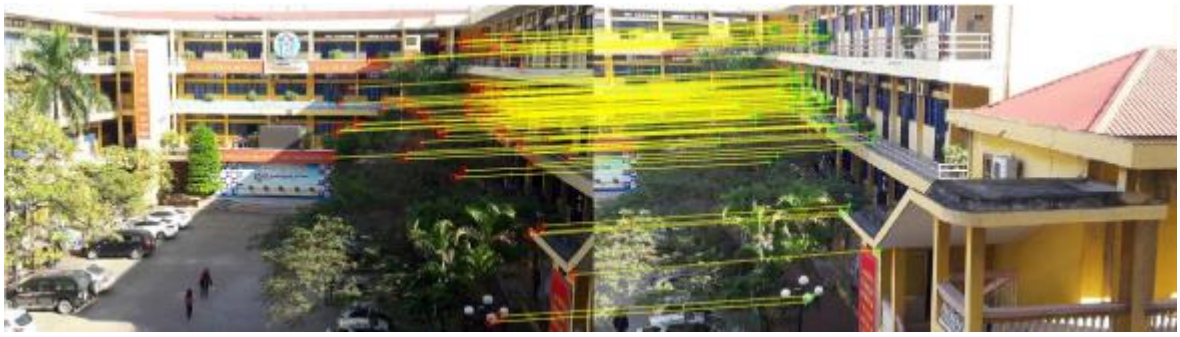

Figure 4.12: Corner search results for the third input image

After matching two images, we have a set of similarities, the next step is to calculate the Homography matrix based on the set of pairs of similarities on the RANSAC algorithm. After calculating the Homography similarity matrix between the two input images, the program will choose the second image as the "center" and transform the first image according to the second image and the third image according to the second image based on the medium Homography matrix find.

Figure 4.13: The first image is transformed according to the second image

Figure 4.14: The second image is the centre so it doesn't change

Figure 4.15: The third image is transformed according to the second image

After transforming the image according to the similarity matrix, the final step of the algorithm is to superimpose three images to form a Panorama image.

Figure 4.15: The resulting panorama

4.4 Trial run result

Based on the experimental results, to be able to combine panorama images, it is necessary to ensure that the content of the second image must contain at least 10% to 15% of the content of the first image.

The content of the second image and the first image overlaps too little, leading to the program not being able to find similarities or finding too few, leading to inaccurate homography matrix calculation.

Two images with different shooting positions can still find similar features and calculate the Homography matrix. However, when merging, some objects are obscured in the first image but not in the second image, so when the image is stitched, there will be the phenomenon of objects being overlapped.

In case the second image is taken diagonally upwards, the resulting returned image is still guaranteed to be acceptable.

The number of similarity features between two images must be at least 4 to be eligible to perform the RANSAC algorithm, thereby finding the homography matrix. To overcome this problem, we need to increase the threshold used in image matching to have more similarities. The greater the number of similarities between two images, the more accurate the resulting image will be.

From the above test cases, it can be concluded that the input image acquisition step plays an important role in determining the output image results. For that reason, it is essential to use a tripod or use a skateboard when taking photos to ensure the best quality of the input image and the least distortion.

5 Conclusion

After a period of researching the topic, the thesis has achieved the following results:

- Learn the principle of image stitching technique.
- Learn how to find and extract key points and SIFT feature representation.
- Learn the feature matching-based panorama stitching process.
- Successfully applied to the installation of the panorama test program.

Limitations encountered in the thesis:

- In feature extraction, the project only learns about Harris algorithm, there is no comparison with other algorithms and matching methods.
- The technique of mixing colors and cropping images has not been implemented yet.
- Due to usage of MATLAB environment, the interface of the program is not user friendly, processing speed is slow.

Future developments:

- Test on many different feature extraction algorithms to get comparison and evaluation.
- Research and apply image smoothing and cropping algorithms in post-processing steps.
- Research and apply optimal algorithms to increase processing speed.
- Applications to create Panorama images on digital photography devices, mobile devices running Android, iOS or Windows Phone operating systems.

References

Markéta Potůčková, “Image matching and its application in photogrammetry ”.

Ms. Vrushali and S. Sakharkar, “Image stitching techniques-an overview ”.

Prof. C. S. Gode and Ms. A. N. Ganar, “Image retrieval by using colour, texture and shape features ”.

Szeliski, Richard, “Image Alignment and Stitching”, 2005

Herbert Bay, Andreas Es, Tinne Tuytelaars and Luc Van Gool, “Speeded-Up Robust Features (SURF) ”.

E Garcia,"SVD and LSI tutorial", MIIslita.com, 2006

Martin F., Robert B. – “Random sample consensus”- A paradigm for model fitting with application to image analysis and automated cartography, Communications of the ACM 24 (1981) 381–395.