Feature-Based Matching in Historic Repeat Photography: An Evaluation and Assessment of Feasibility

by

Christopher Gat
BSc, University of Victoria, 2006

A Thesis Submitted in Partial Fulfillment
of the Requirements for the Degree of

Master of Science

in the Department of Computer Science

Supervisory Committee

Feature-Based Matching in Historic Repeat Photography: An Evaluation and Assessment of Feasibility

by

Christopher Gat
BSc, University of Victoria, 2006

Supervisory Committee

Dr. A. Branzan Albu, Co-supervisor
(Department of Computer Science / Electrical and Computer Engineering)

Dr. D. German, Co-supervisor
(Department of Computer Science)

Dr. E. Higgs, Committee Member
(Department of Environmental Studies)

Abstract

Supervisory Committee

Dr. A. Branzan Albu, Co-supervisor
(Department of Computer Science / Electrical and Computer Engineering)

Dr. D. German, Co-supervisor
(Department of Computer Science)

Dr. E. Higgs, Committee Member
(Department of Environmental Studies)

This study reports on the quantitative evaluation of a set of state-of-the-art feature detectors and descriptors in the context of repeat photography. Unlike most related work, the proposed study assesses the performance of feature detectors when intra-pair variations are uncontrolled and due to a variety of factors (landscape change, weather conditions, different acquisition sensors). There is no systematic way to model the factors inducing image change. The proposed evaluation is performed in the context of image matching, i.e. in conjunction with a descriptor and matching strategy. Thus, beyond just comparing the performance of these detectors and descriptors, we also examine the feasibility of feature-based matching on repeat photography. Our dataset consists of a set of repeat and historic images pairs that are representative for the database created by the Mountain Legacy Project www.mountainlegacy.ca.

Table of Contents

List of Tables

List of Figures

Acknowledgments

The catalyst to the event that is this thesis was an out-of-the-blue email from Daniel German asking if I would be interested in pursuing a master's degree. I was roughly 16 months into a voyage through South America, and a person in search of what would be next. While the significance of that email at the time may have seemed marginal, it was the start of a new journey that will culminate with the defense of this thesis. Several people have influenced this work and I would like to take the opportunity to acknowledge their contribution.

Daniel German did more than just send an email, but gave important guidance and a valued perspective on the my research as well as assisting my studies with finance support. Alexandra Branzan Albu gave constant support, encouragement, and thoughtful advice throughout the process, in addition to sparking my interest in Computer Vision. Eric Higgs, the founder and director of the Mountain Legacy Project, not only provided me with employment and a database of repeat imagery, but also a project worthy of my interest and pursuit. I am also grateful for the support from outside the university environment: my partner, Vladimira Lackova, constantly reminded me that I could do it, even when I could not see the light at the end of the tunnel, my brothers and their partners, family that I consider lucky to have, and lastly, my parents, Barb McKrow and Gabriel Gat, whose unconditional love and support knows no bounds.

Chapter 1 Introduction

Repeat photography is the process of recapturing a photograph from the same vantage point of a reference photo. When the time between the repeat and the reference image is substantial, the differences in content can be dramatic and alarming. Repeat photographs have been used in *New York Changing* [1] and the *Then and Now* [2] book series to exhibit significant change in urban areas over time. Changes in landscape have also been recorded in several repeat photography projects. The *Second View* and *Third View* books [3], [4] have repeated landscape images in the American Midwest. The Mountain Legacy Project (MLP) is devoted to incrementally repeating photographs from the Library and Archives Canada and British Columbia Archives collection of historic surveys images (est. over 140,000). These repeat photograph pairs have been used as evidence for climate change and research in environmental studies and ecological restoration [5-8].

Typical tasks in the repeat photography process involve the determination of geographical location and field of view of the original photo (e.g. near the top of a certain mountain) and the manual matching/alignment of the original and repeat images. Computer Vision can play a major role in both tasks. Our focus here is on image matching. Repeat photographs are rarely taken with the same camera and lens as the original, therefore even when a historic image has been well repeated, the images need to be scaled, rotated, and translated to produce the final alignment of the images.

Feature-based matching approaches are well suited for image alignment. They involve three stages: detection, description, and matching. The detection phase is concerned with finding salient points within an image that may be associated with a scale and shape.

These feature points or regions can then be processed by a descriptor, a method that characterizes the local structure of the point or region as a feature vector. The matching stage defines how these features vectors are compared and associated.

This thesis investigates the feasibility of feature-based matching in the context of repeat photography. We report on the quantitative evaluation of four state-of-the-art feature detectors and two feature descriptors on a representative subset of the Mountain Legacy database. Our work is important due to the unique content of our image set: there is considerable potential for digital change detection methods on landscape repeat photography for numerous research disciplines (e.g. environmental science, landscape ecology). A critical prerequisite for the detection of change, however, is the detection of similarity, which is inherent in the feature-based matching process. Furthermore, the results of our evaluation can be directly applied to a practical problem within the MLP, namely the alignment of historic and repeat photographic pairs after the repeat. The first two subsections elaborate further on these motivations. This is followed by a section rationalizing our work. Finally, we conclude this chapter by summarizing our contributions and giving an overview of the remainder of the thesis.

1.1 The Potential of Landscape Repeat Photography

The monitoring of change in the Earth's surface via remotely sensed data provides a foundation for improved understanding of the interaction between humans and natural phenomena and helps us better manage and use resources [9]. The monitoring of such change is actively researched in the field of change detection. Change detection, at its simplest definition, is the process of identifying change in phenomena or objects over some temporal range [10]. The type of change depends on the type of research being

done. For example, remote sensing multi-temporal data sets have been used to identify change in the following categories: land-use and land-cover; forest or vegetation change; forest mortality, defoliation and damage assessment; deforestation, regeneration and selective logging; wetland change; forest fire and fire affected area of detection; landscape change; urban change; environmental change; crop monitoring; shifting cultivation monitoring; and change in glacier mass balance and glacier [9]. The principal drawback to satellite based remotely sensed data is the lack of temporal range; change can only be monitored in the last 40 years. Furthermore, our ability to understand the driving forces and processes controlling change are directly influenced by our ability to monitor long-term change [11]. The use of aerial photography can improve the temporal range of change detection, however, the oldest photographic recordings of landscape are found in ground oblique photographs.

The oblique nature of such photographs makes them difficult to directly compare against modern methods of landscape data representation (remote sensing, GIS). However, we have a direct way to compare such scenes if we have a current repeated photo of the same scene. Thus, the potential of the MLP dataset for examining historical change is evident, however, for the data contained in these photos to be fully utilized by researchers two problems must be solved: the geo-referencing of pixels, and digitally aided identification and description of change. The former problem is one that would allow a researcher to make spatially explicit assertions about the changes occurring between the images. The latter problem is one that spawns from a desire for convenience; manual interpretation of the images is generally more accurate, but considerably more

time consuming, requires expert interpreters, and is subjective in nature[12]. In both cases, a pre-processing step to accurately register the images is needed.

1.2 Manual Alignment of MLP Repeat Photographs

The MLP field crew last season captured over 900 repeat photos. Currently, the process to align these photos has been a manual one: lab assistants would use Adobe Photoshop CS4 to overlay the images, using basic transformations (rotate, scale, and translation) available in the Free Transform tool. The task is difficult since the act of overlaying the historic image onto the repeat involves both transforming the image and adjusting the transparency to assess the degree of alignment that has been achieved. Individually, the image pairs can be aligned within a few minutes, but with the volume of repeats that are taken in the field season, the process becomes time consuming, tedious, and costly. For example, typically this task would take a lab assistant 2 to 3 weeks of full time work.

1.3 Rationale for our Work

The primary rationale for our work is that the performance of feature detectors and descriptors may vary for different applications and scene types, and thus, the results from previous evaluations cannot be applied to the conditions of repeat photography.

We also propose a novel evaluation technique. Ideally, we would adopt a previously used method that best matches our objectives. However, the context of our dataset and the nature of differences between images in repeat photography make such a strategy difficult.

The primary differences that cause this problem are:

- Our dataset does not contain image sequences of the same scene over some controlled transformation, therefore we cannot analyze a detector-descriptor's robustness to small increments of change.
- A repeat photographic pair can differ by change in the scene (e.g. trees might grow or disappear over time). Such differences can lead to accidental correspondences (features detected on the same respective location, but for different reasons), which have implications in using the repeatability rate and recall criteria (as have been used previously) for evaluation.

1.4 Contributions

The contributions of this work are:

- The identification of performance criteria that is best suited for the simultaneous evaluation of detectors and descriptors in the conditions of historic repeat photography.
- Two novel approaches, pass rate and rank score, to assessing these performance criteria over a large dataset of image pairs.
- The evaluation of 4 state-of-the-art detectors and 2 descriptors with historic repeat photography. Specifically, we assess the ability of a these feature-based matching systems to find matching features between image pairs, and present the data theoretical manner, not making assumptions about the number of correct matches or the precision that might be required by an application.
- The feasibility of solving a practical problem within the MLP using a feature-based approach. We also examine if *a priori* information can be introduced to help improve results.

The evaluation does not take into consideration the efficiency of the algorithms used for detection and description, since the computational costs are only marginally dependent on the content within the images. Furthermore, detectors and descriptors can be configured and adjusted in numerous ways. We do not adjust these parameters, but use each method in their default setting, as has been done in other evaluations. Finally, we do not explicitly evaluate the localization accuracy of detectors.

1.5 Overview

This subsection outlines the structure of the remainder of this thesis. In Chapter 2, we briefly explain and discuss the background of the Mountain Legacy Project. Included in this section is a detailed explanation of the MLP data, and some definitions related to MLP terminology. In Chapter 3 we present a review of literature related to this thesis. Chapter 4 introduces the reader to detectors, descriptors, and the matching approach used in this evaluation. In Chapter 5, our proposed evaluation methodology is presented, subsequently followed by the experiment results (Chapter 6), and overall conclusions and future work (Chapter 7).

Chapter 2 Background

2.1 The Mountain Legacy Project

In a period between 60 to 120 years ago, surveyors in Western Canada used photographic surveying techniques to create the first detailed topographic maps of the area[5]. The Mountain Legacy Project, originally called the Rocky Mountain Repeat Project, is devoted to the systematic repetition of these photos[7] for the purpose of research in environmental and ecological research. The collection of historical photos is estimated to be roughly 140,000, of which approximately 4000 have been digitalized and repeated (as of 2010).

The historic photographs are originally in the form of glass plate negatives. These glass plates are scanned with a ScanMate F8 Plus scanner at 1800dpi with a 16 bit depth. With the exception of the black and white large format photographs taken in the early days of the project, repeats are originally archived in their raw format (3FR file). Raw files are then converted into 16 bit tiffs using Phocus (Hasselblad's proprietary raw processing software) , which are then manually overlayed (registered) in Photoshop CS4. The final registered repeats and historic images are saved in 8-bit form. It should be noted that the registration process only involves translation, scaling, and rotation.

An example of an image pair from the MLP dataset is shown in Figure 1.

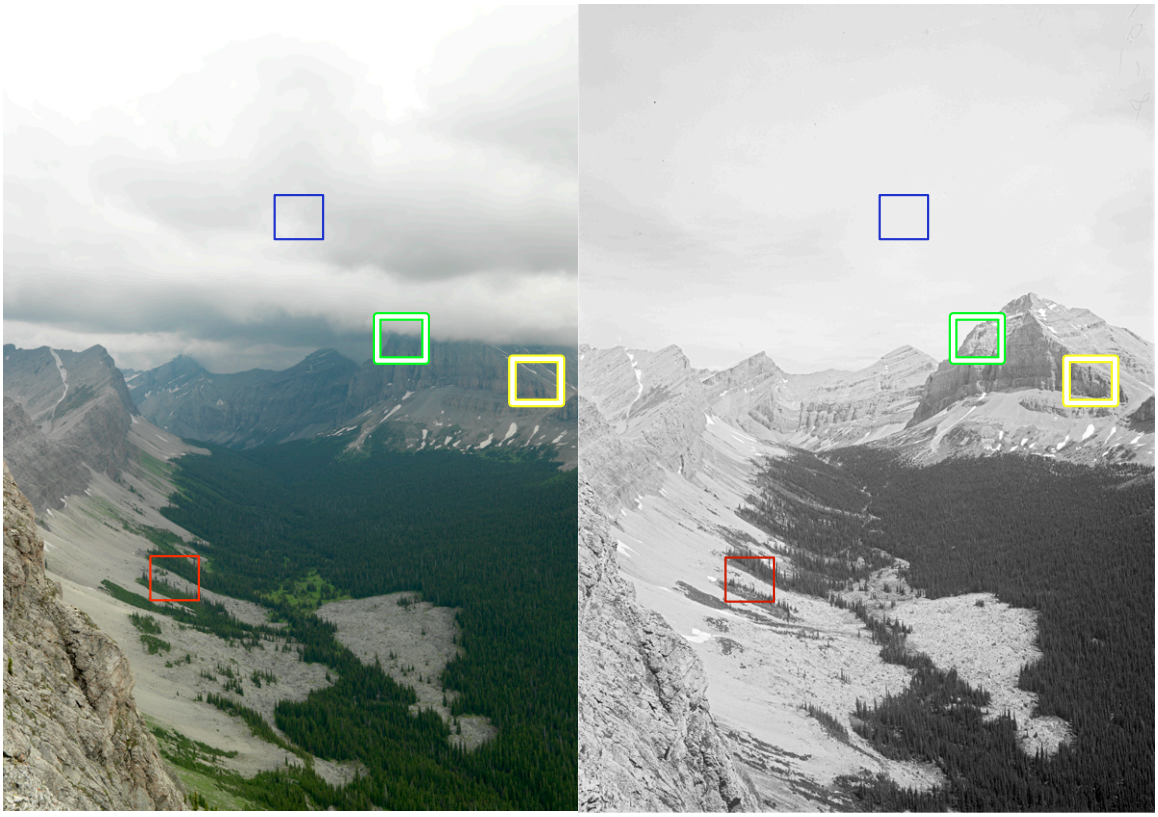


Figure 1 An image pair from the Mountain Legacy dataset. The coloured boxes refer to magnifications presented in Figures 2 through 5.

2.1.1 Terminology

- *Historic Image* – a historic image can be any image that is from a historic survey. The project internally refers to these images as "originals", however, in this thesis I have chosen to use to term "historics" to refer to the images because the term is generally clearer.
- *Repeat Image* – a repeat image is a photograph of the historic scene that has been recaptured from the same viewpoint.
- *Registered Image* – a registered image is one that has been manually aligned by the MLP lab assistants. A registered version of both the historic image and repeat image exist. These images are referred to as the registered historics and registered repeats.

- *Raw Image* – a raw image is a version of a historic or repeat image that has not been manually aligned. These images are referred to as the raw historics and raw repeats, respectively.
- *Survey* – each historic image comes from a specific historic survey. A survey involved taking photographs over a specified area, sometimes over several years, usually with the intention of creating maps with the information contained in these photographs. The name of the survey is usually associated with the name of the principle surveyor, year(s) the survey took place, and the general area that the survey covered.
- *Station* – a station is a specific location in the survey at which the historic photographs were taken. A station usually includes several photographs, taken in different directions and locations that are within walking distance from each other.
- *Surveyor* – a surveyor is the person who organized and led the survey.

2.1.2 Data Set Characteristics

The historic photographs were the result of a variety of different surveys by different organizations, so it is difficult to determine the camera used for each specific photograph. However, a guide, *Photographic Surveying*, written by one of the most prominent surveyors, M.P. Bridgland, does reveal information pertaining to equipment used on at least some of the surveys before 1924[13]. The book suggests that a large-format camera, 4 ¾" x 6 ½" glass plate negatives with 164mm Zeiss Tessar Series III lens with B/W panchromatic emulsion, and Wratten and Wainwright "G" filter (yellow) were used during Dominion Land Surveys in the early 20th century. As early as 1895, surveyors

were using techniques to remove the effect of haze in the landscape photographs. This was done by using orthochromatic plates (which were less sensitive to the blue-violet end of the spectrum) coupled with a deep orange tint screen (pre 1915) or panchromatic plates coupled with a yellow filter (approx. 1915-1924). As a result, blue haze was mitigated, but shadows were darkened since these areas are predominantly lit by the scatter atmospheric light. A property of survey images is that they are taken level to the horizon (tangent to the earth's surface) in accordance with survey guidelines.

The repeat photographs have been acquired with several cameras. The first repeats taken were in 1997 by Jeanine Rhemutalla and Eric Higgs using a Linhof Technika large-format film camera, 90mm Schneider-Kreuznach Angulon 1:6.8 lens, No. 85 Wratten filter (pale orange), a "polarizing and haze" filter, and T-max 100 black and white film[8]. The project eventually switched to digital when the Hasselblad H1 with the Imacon iExpress 16 megapixel digital back was introduced in 2006. Both a Hasselblad H3D-39 and H3DII-39 are in use today. The H3D was introduced in 2007 and the H3DII was introduced in 2009. Both these cameras have the same 39 megapixel 36.8x49.1mm sensor. The lens used to take the photos is the Hasselblad HC 3.5/35 Lens (actual focal length is 35.8mm) with a 63.88 degree angular field of view. The majority of the repeats have been taken with the H3D models. Only UV filters were used the Hasselblad camera configurations. The resolution of the raw Hasselblad image files is 7216x5214 pixels, while the raw scanned glass place negatives (historics) have a resolution of approximately 11,500x8,500 pixels (the scanning procedure does not use a fixed cropping, so the actual resolution differs from image to image).

2.1.3 Characteristic Differences Between Image Pairs

The data set has a particular set of characteristics that are very important when attempting to use them in the context of feature detection and description. In the following we note some of the most pertinent differences between the historic and repeat images.

- *Image capture device*: here we refer to the camera system used to capture the image. The most important aspects of the camera system are the lens and the sensor or film (for the sake of convenience we will refer to both digital sensors and film as "sensors"). By changing the capture device, the spatial and spectral resolutions, in addition to the noise characteristics, differ between the images captured with those devices.
    - *spatial resolution*: The spatial resolution is the ability of the camera system (sensor and lens) to resolve details. Both the current day and historic systems were able to resolve an exceptional amount of detail. Differences in spatial resolution are therefore small, especially with the downsized images that are used in this evaluation. An example of detail captured in these images can be seen in Figure 2.

Figure 2 An example of the significant detail (spatial resolution) obtained in both images and subtle changes in landscape. These regions are associated with the red boxes in Figure 1.

- *spectral resolution*: This type of resolution relates to bandwidth of light frequencies captured by the sensor [14]. As noted previously, filters may have been used in some of the historic surveys as well as plates that were less sensitive to the blue/violet range of the spectrum. In contrast, the repeat photographs are able to capture the visible spectrum; capturing ranges of this spectrum in three separate channels associated with the colors red, green, and blue.
- *noise*: Any image, whether captured on film or a digital sensor, is associated with some degree of degradation. Such degradation can generally be referred to as noise. One cause is from random errors associated with the image capture device, transmission, or processing. Figure 3 shows an example of such noise. Notice that the historic image is noisier than the repeat. The historic images tend to also include non-random noise related to physical damage of the glass plate negatives. This includes scratches and speckles in the image.

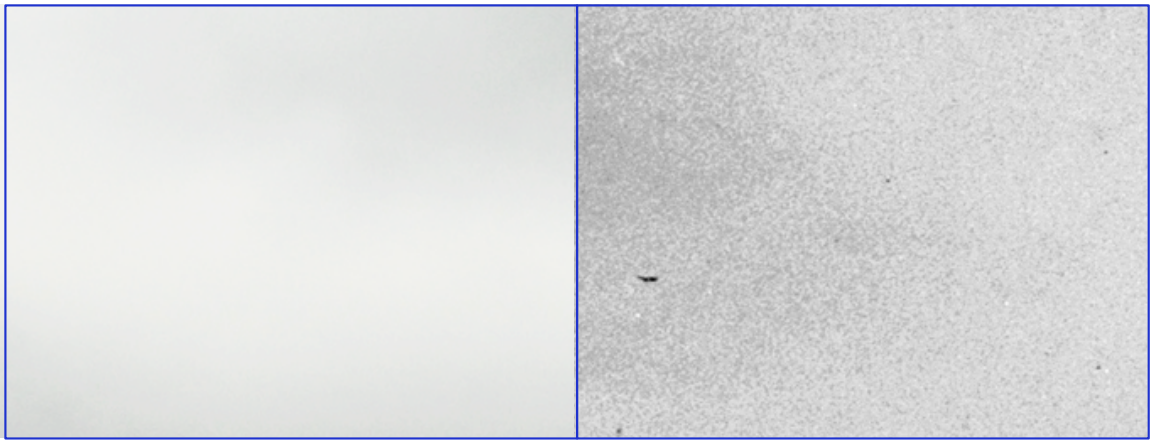


Figure 3 An example of the difference in noise characteristics between these images (contrast has been increased to amplify noise for viewing). These regions are associated with the blue boxes in Figure 1.

- *radiometric*: The radiometric attributes of an image are the ones associated with the visible light used to form the image on the sensor. The repeat photographs have been taken without regard to the light conditions of the historic image. Therefore, we need to take into account the change in light source (illumination) and the disturbance of light (atmospheric).
    - *illumination*: The landscape images in our dataset are lit by the sun. The strength and direction of the sunlight illuminating the scene differ depending on the time of day/year and whether the light reaching the scene object is indirect or direct. The implication of a change in direction of the source light may cause shadows to occur in different parts of the scene, possibly changing the gradients in those areas as well. These shadowed areas still receive light, but indirectly through reflection of surrounding objects or through the scattering of light in the atmosphere. See Figure 4 for an example of illumination change.

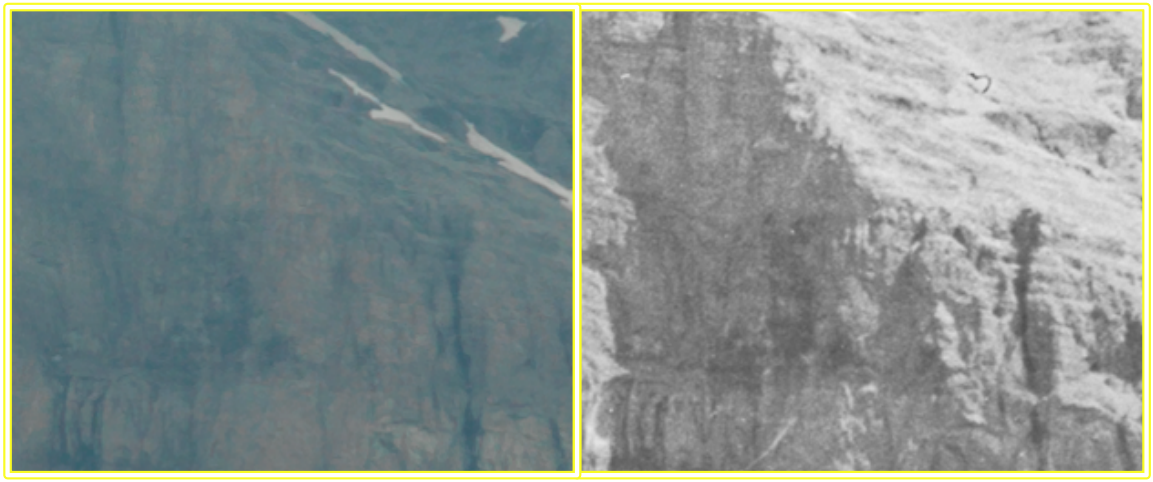


Figure 4 An example of illumination change. The historic area has both direct and indirect lighting (in the shadow), while the repeat area is lit by diffused sunlight from the overcast sky. These regions are associated with the yellow boxes in Figure 1.

- *atmospheric*: The sunlight directed onto the scene must pass through atmosphere that affects that light that is seen by the viewer. Weather is the best example of atmospheric interference, however, dust and smoke are common disturbances as well. It affects the contrast of the scene at different depths when haze is present and occlude direct sunlight from parts or of the scene when clouds are present. See Figure 5 for an example of atmospheric interference.

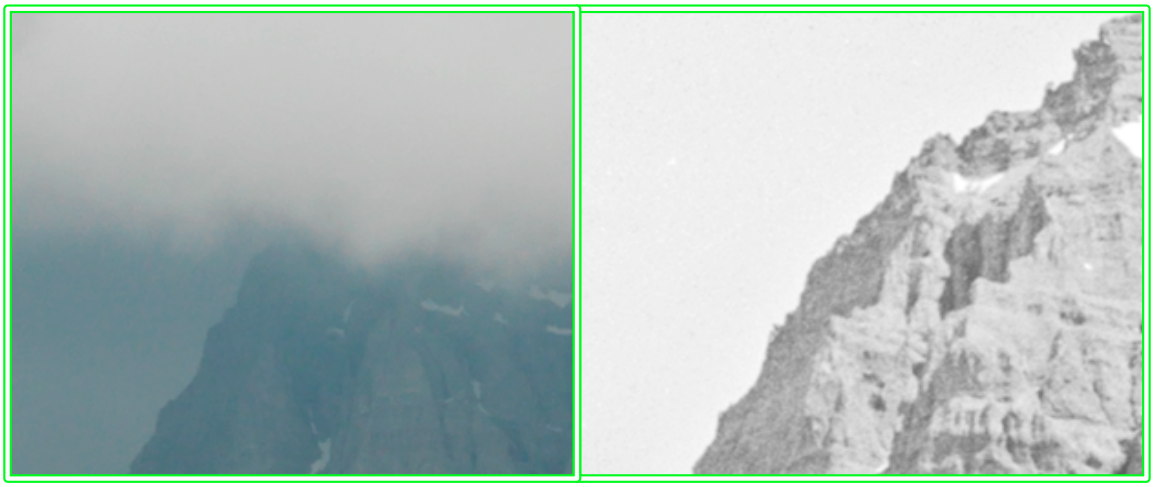


Figure 5 An example of atmospheric change. At this depth the scene begins to lose contrast in the repeat compared to the historic. Furthermore, atmosphere occludes parts of the mountain. These regions are associated with the green boxes in Figure 1.

  - *phenological*: Phenological change (seasonal differences in plant life) is not taken into account when a photograph is repeated. Therefore the phenological attributes of the scene may differ. This difference is dependent on the scene and the plant life within. In the context of image processing or computer vision, this would be associated with a change in texture, tone, or colour.

- *perspective*: While great care is taken to find the exact place a historic photograph was taken, it is natural to assume a certain amount of error. Perspective change, both from change in camera location and optical center misalignment, distorts the scene of the repeat such that exact alignment cannot be achieved with a simple scale, rotation, and translation.
- *landscape change*: Landscapes are spatially heterogeneous geographical areas characterized by diverse interacting patches [15]. From the point of view of a landscape ecologist, a change in landscape depends upon the context at which

the landscape is compared. For example, the large-scale structures (the mountains) in Figure 2 remain persistent in both the historic and repeat versions of the image. However, subtle changes such as tree growth or tree line movement may also be regarded as important aspects of change. In the context of computer vision however, we are interested in how the local intensity patterns change within the image in association with some physical change within the scene. A change in landscape may change the texture or contours of corresponding areas of the images. In some cases, the texture and contour changes related to landscape change in the MLP images are significant.

Chapter 3 Related Work

This chapter gives an overview of the work related to this thesis. The first section covers work that has been done in computer vision in regards to repeat photography images and complex scene changes. Section 3.2 introduces the reader to previous evaluations on detectors and descriptors.

3.1 Historic Repeat Photography and Complex Scenes In Computer Vision

Very little work has been done in computer vision that has directly used repeat photographic image pairs. The most notable study was conducted by Bae and Agarwala [16], when they constructed an application that aided users in relocating the position where a historic photograph was taken. Specifically, a user took two wide baseline images of the scene, and then manually identified 8 corresponding points between one of these images and the historic photograph. Given this information, they were able to use epipolar geometry to resolve the users location with respect to the original location the photography was taken. A series of directional cues were then given to the user to guide them towards the correct location and pose to repeat the photograph.

Some image sets, while not explicitly defined as historic repeat photography, still contain a similar set of variable change that is comparable to ours and worth mentioning. Photo tourism is an image based rending explorer that leverages SIFT[17] to match points and structure from motion to reconstruct 3D representations of a scene from a large variety photos[18]. The dataset consists of large collections of photos over a large scene (e.g. a city plaza in Prague) taken by tourists at different perspectives, lighting conditions, time, and weather. The matching problem in these complex situations is overcome by the transitivity of the dataset. That is, one image corresponds to a series of

images and therefore a feature need not match for every single image, but rather it just needs to match with the same feature in at least one other image which is then linked to the others. While this technique might be able to be used on repeat photography that is taken periodically, this is not the case with our dataset, where we assume a scene has been repeated only once. [19] proposed a similar application, but focused on a method to temporally sort historic urban photographs based on change in scene structure, adding the dimension of time to a photo tourism type experience. Correspondences in the historic images were done manually rather than automatically generated.

3.2 Evaluation Techniques for Feature Detection and Description

The evaluation of both feature detectors and descriptors has been actively researched over the past 10 years. The evaluation process can be divided into following components: the creation of ground truths (Section 3.2.1), performance criteria (Section 3.2.2), aggregation methods (Section 3.2.3), and dataset characteristics (Section 3.2.4). At the end of this section we've included 3 overview tables that organize the most pertinent aspects of the evaluations reviewed.

3.2.1 Ground Truth Creation

The basis of every feature detector and descriptor evaluation is the ground truth[1] data; a sequence of two or more images of the same scene where the association of real world points between the images is known. This association between points is called correspondence. The correspondence between points in a 2D scene can be defined through homographies (projective transformation). Once the geometric relationship

[1] This is not to be confused with the ground truth of an ecological study, for which the term refers to field-based sampling.

between corresponding points is defined, an error tolerance, both for location and region overlap, is defined.

A homography can describe both the forward and inverse geometric relationship between photographs of real world scenes of planar surfaces (e.g. a wall with graffiti on it) or scenes in which the camera viewpoint is held constant, but transformed through scaling or rotation (e.g. using the zoom of a camera).

Schmid *et al.* [20] were the first to apply this concept for the purpose of establishing a ground truth for an evaluation on interest point detectors on planar scenes. To create the homography that describes the geometric relationship between images a few corresponding points are needed between images. Once these correspondences are known, the transformation can be modelled using a method such as Direct Linear Transformation [21]. Rather than selecting the corresponding points manually, which Schmid deemed subjective, she used a more systematic approach. For sequences of images, the scene was fixed and the camera adjusted based on the transformation being tested. Each test image was also photographed with a black dot grid projected onto its surface. A template matching method was then used to extract locations of black dots on each image, and these locations were manually paired for use in the algorithm that modelled the homography transformation. This method worked well in a studio setting, however was difficult to apply to real world (as opposed to a controlled studio setting) images. [22] used a process to allowed for such images. Corresponding points were manually selected between image pairs to approximate a homography and then further refined with a small-baseline homography estimation algorithm [21].

Once the homography is defined, a homogenous coordinate from one scene can be transformed to the other. This attribute gives us the ability to compare whether two local features from two images correspond to the same point in the scene. This comparison is formulated by using the Euclidean distance measure between two coordinates, where one of the points has been transformed by the homography.

$$\varepsilon_l = dist(H\mathrm{x}_1, \mathrm{x}_2) \tag{3.1}$$

where H is the homography from image 1 to image 2, x_i is the homogenous coordinate from image i , and *dist* is the Euclidean distance function. ε_l is the location error between two features. Ideally, a feature detector locates features such that the location error is zero. However, in practice such accuracies are not achieved. Mikolajczyk and Schmid[23], [24] tolerated a location error of less than 1.5 pixels when considering the correspondence of two points.

The above ground truth criterion can be used for feature points, but is insignificant for scale-invariant and affine-invariant features. In these cases, the consistency between feature region size and shape must also be taken into consideration. Consider the example shown in Figure 6. We have two images: a reference image and a comparison image, which differ solely by some scale factors *s*. The image of a white circle within a larger black circle is a trivial example of two blob type features located at different scales. The coloured based lines indicate the rough boundaries of these features. The task of determining the ground truth is establishing which features from the reference image should be defined as corresponding with the features of the comparison image. If only the location error that was previously mentioned were to be used, there would be 4 correspondences, since each feature would correspond with both features of the paired image. However, if the size and location of the feature is consistent over the

transformation between the images, then the ground truth correspondences can easily be established. The table on the right side of Figure 6 shows all possible corresponding features. The scale factor *s* has been applied to the features in the reference image so that they can be compared. The last column shows the two features overlapping each other, where the light red colour indicates parts of the surface that are inconsistent. This area is also known as the overlap error.

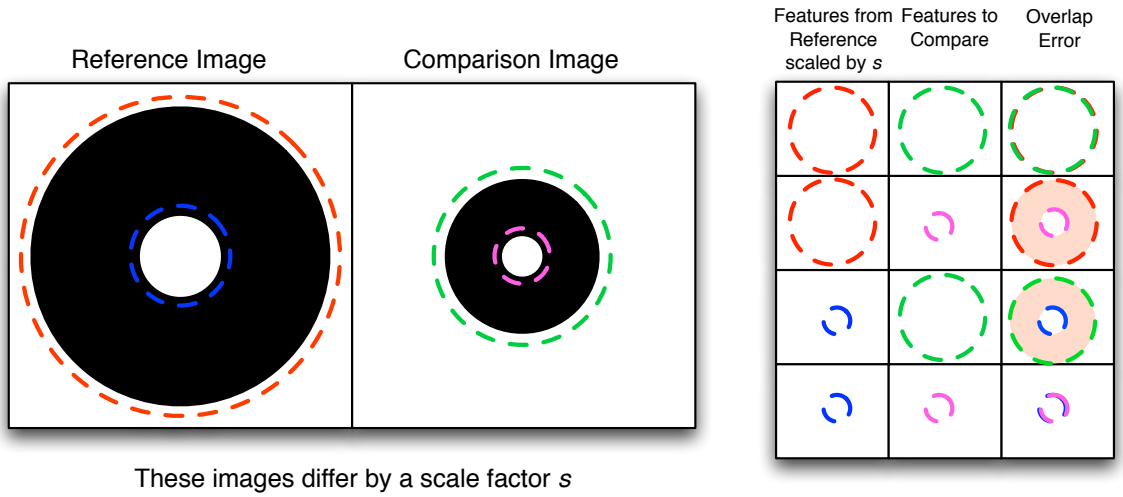


Figure 6 Feature Overlap. Two local features are denoted by dashed coloured lines in the reference and comparison images. Without a measure of overlap error between the regions, the ground truth correspondence cannot fully be defined. The table on the right shows all possible corresponding features. The scale factor *s* has been applied to the features in the reference image so that they can be compared. The last column shows the two features overlapping each other, where the light red colour indicates parts of the surface that are inconsistent. This area is also known as the overlap error.

The overlap error can formally be defined as

$$\varepsilon_O = 1 - \frac{\mu_a \cap \mu'_b}{\mu_a \cup \mu'_b} \tag{3.2}$$

where μ_a is a feature region from image A and μ'_b is a feature region from image B transformed according to the homography. The feature region area, μ, is computed by

counting the pixels that lay within the region boundaries (which can be defined geometrically). This approach has been applied in [22], [23], [25], [26], where the overlap error tolerance level was set to 40%. In evaluations for affine-invariant features detectors[22], [23], the location error was ignored, since the overlap error of the elliptical regions used to represent affine features were considered to be the much more dominant source of error.

A problem with the above approach is that it is biased towards larger regions. This is demonstrated in Figure 7. The overlap error between two regions can always be made smaller by increasing their scale. [22] proposed a solution to this problem. Instead of comparing regions at their detected size, they are normalized for comparison. A reference region is first normalized to a constant size (a 30 pixel radius in their experiments). This normalization produces a scale factor, which is then used to scale down the region of comparison. Such normalization is only desirable in the comparison of detectors as opposed to a matching experiment, where description takes place as well.

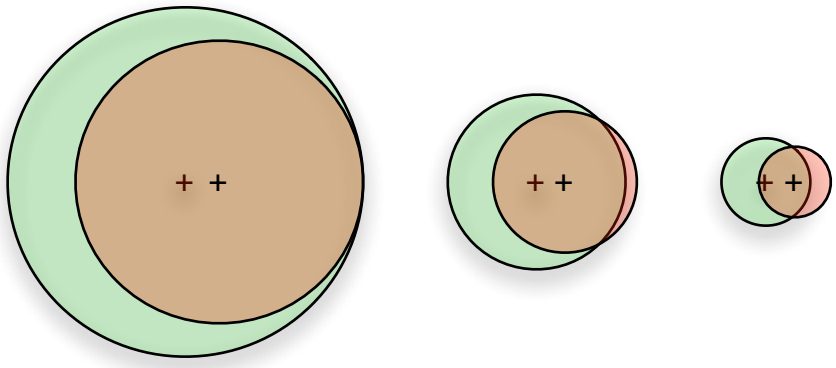


Figure 7 Large Region Biased. The figure shows the overlap between two features at small, medium, and large scales. As the scale of the regions increase, the overlap error between them decreases.

The ground truth created with the location and overlap error can also be used in the evaluation of descriptors. In a matching scenario, two groups (one for each image) of feature vectors are compared. The ground truth is the set of feature vectors from one image that are known to match a set of feature vectors from the other. Since corresponding regions contain the same content, or at least coverage of the same objects, this mapping can also be used as the ground truth for descriptor evaluation.

[22] conducted an overlap error experiment to establish which overlap error tolerance level was able to capture the majority of the correct matches and corresponding regions, while not sacrificing too much of the recall statistic (the larger degree of error, the greater the chance of false positives). The experiment was done on a single image pair that differed on viewpoint. They plotted a graph showing the number of correct matches and corresponding regions over an interval of change. For example, the score for 30% overlap error was computed for the number correct matches whose error was between 20% and 30%. Correct matches were defined by fixing the distance threshold for each descriptor so that it returned a 50% precision. The graph demonstrated that most corresponding regions were found between an error rate of 10% and 60%, while most correct matches were found between an error rate of 10% and 40%. According to this data, an overlap error of 50% was chosen for the remaining experiments conducted in the study.

The creation of ground truths for 3D scenes are more elaborate and beyond the scope of this thesis. For further information regarding 3D ground truths see the work of Moreels and Perona [27] and Fraundorfer and Bischof [28].

A study by Valgren [29] comparing the effectiveness of SIFT and SURF [30] on outdoor scenes which incur seasonal change, determined the ground truth of matching

points by having a human judge their similarity in location based on the criteria that features matched if they "clearly correspond[ed] to the same object in the images".

3.2.2 Performance Criteria

Repeatability has been noted as the most important quality of a feature detector [31]. It is unsurprising then that it has been the most actively used criterion in evaluations of feature detectors. Repeatability is the persistence of a detected feature over some geometric or photometric transformation [31]. The rate of repeatability is the number of corresponding features, as defined by the ground truth, with respect to the number of detected features [20]. Formally this can be described as:

$$R_{AB} = \frac{C_{AB}}{\min(D_A, D_B)} \tag{3.3}$$

where R is the repeatability rate between images A and B, C is the number of corresponding features between the images, D_i is the number of detected features for image i, such that they have the potential of being repeated. For a feature to have the potential of being repeated, the scene object that it is located on must appear in both images that are being compared. Figure 8 gives an example where a detected region in one image is not included in D. The feature denoted by the red ring is not common to both images, therefore it is not considered when determining the repeatability rate. A similar situation occurs when photos are taken at different viewpoints on a 3D scene, when features detected on areas that have been occluded must be removed.

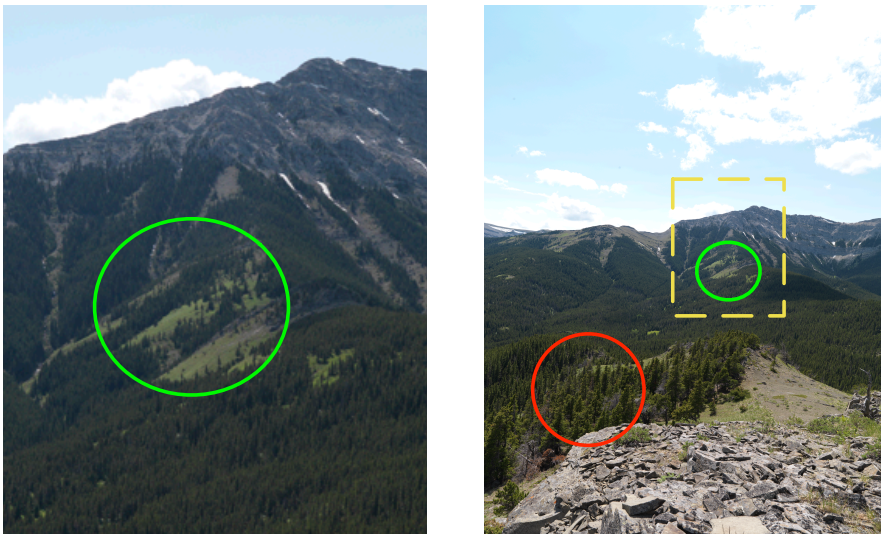


Figure 8 Features common to the image pair. The image pair above gives an example of two features that are common to both images (the green circles), and one feature that occurs on an uncommon area (the red circle). When the repeatability measure is used, it is important to consider only features that are common to both images.

Mikolajczyk *et al.* and Fraundorfer and Bischof [22], [28] used the repeatability rate along side the number of correspondences, number of correct matches, and the matching score as performance criteria of detectors. The matching score was defined as the number of correct matches with respect to the number of detected regions. The SIFT descriptor was used to define a match. Notice that if the descriptor performed a perfect matching, the matching score would be the exact same as the repeatability rate, since the number of corresponding regions would equal the number of correct matches. With this in mind, if the matching score did not deviate from the repeatability rate, then the detector was considered to find more distinct regions than ones where the opposite was true.

In addition to defining the repeatability rate, Schmid *et al.* [24] also introduced the performance criterion of information content, a formal measure of distinctiveness for detectors. Distinctiveness is the likelihood of a descriptor computed at the point within the population of all the observed described local features. That is, each interest point

detector evaluated finds a set of points within an image. Each of these points can be described by a feature descriptor, and represented as a feature vector. Each feature vector represents a point in descriptor space. The information content property measures the distribution of the descriptors in descriptor space. The more spread out the descriptors, the higher the entropy (the more random the descriptor) and the greater the distinction between feature vectors, while the less spread out, the lower the entropy and the lower the distinction between feature vectors. Entropy is calculated based on probabilities:

$$H(A) = -\sum_i p_i \log(p_i) \tag{3.4}$$

of partitions of descriptor space. The probability of a partition is the number of descriptors contained within the partition as compared to the total number of descriptors. In order to remove bias to a certain scene type, interest points were extracted and described from a 1000 images of different types (aerial photography, images of toys, and images of paintings) and descriptor vectors were calculated and normalized for each. This measure has also been used in [32].

The evaluation of feature descriptors has typically been done in the context of either image matching [33], [34] or image retrieval [34-36]. In each circumstance, we have a set of feature vectors that need to be labelled as a match or non-match. However, the case of image retrieval the problem is framed such that one image is used to query a database of images. In this circumstance, a much higher potential for false positives exist. This subtle difference has given rise to the use of different performance criteria.

In the context of image matching, feature vectors that need to be matched come from two images that share the same scene content, but differ in some regard (e.g. viewpoint). The goal in the image matching scenario is to find the matching locations between the

images. The performance criteria of *precision*(alternately, *1-precision*) and *recall* have been used to evaluate descriptors and detectors in this context [33], [34], [37]. *Precision* and *recall* can be defined as follows:

$$recall = \frac{\#\,correct\ matches}{\#\,correspondences} \tag{3.5}$$

$$precision = \frac{\#\,correct\ matches}{\#\,total\ matches} \tag{3.6}$$

where a correct match is two feature vectors that are deemed correct according to the ground truth, and the total matches are summation of both the number correct matches and false matches. *Precision* is useful for image matching because the performance of outlier removal methods such as RANSAC are largely dependent on the ratio of correct matches to total matches returned. *Recall* is relevant because an ideal descriptor can correctly match all of the true matches in the ground truth. Typically, *precision* can be improved at the cost of *recall* by adjusting an acceptance threshold in the matching process. As an alternative to *recall,* Valgrin and Lilenthan [29] used the quantity of correct matches in conjunction with *precision*.

While the goal of image matching is to find matching points between two images, the objective of image retrieval is to find a matching image in a collection of images given some query image. The feature vectors of the query image need to be compared against a database of feature vectors that come from a variety of different images. As with the image matching problem, the goal is still to maximize the number of correct matches, and minimize the number of false matches. An alternative to precision and recall that describes the same goal is the Receiver Operating Characteristic (ROC) curve. Such an approach was used by [27], [35], [38] for the evaluation of descriptors.

The performance criteria for specific evaluations are summarized in Table 1.

3.2.3 Aggregation Methods

The aggregation of performance criteria over a dataset is another interesting characteristic of an evaluation. In the simplest case no aggregation needs to be done since the dataset is small enough for the analysis of individual results. [20], [22], [28] are good examples of such a scenario, where each image in a sequence represents one transformation, associated with one repeatability result, which is plotted on a graph for that particular image sequence. However, when the number of scenes for an evaluation is expanded each sequence cannot be analyzed individually. There is a desire to aggregate the results. In the case of the repeatability rate, aggregation has been done by averaging results obtained at the same transformation over different image sequences. For example, two image sequences may capture a viewpoint change on two separate scenes. The second photo of each sequence captures a viewpoint transformation of 20^{o}. The repeatability for each image transformation can then be averaged. Such an approach has been used in [23], [37] . Aggregation of performance criteria associated with precision vs. recall and ROC curves is less trivial. The previous case (repeatability), a specific transformation could be represented as a single value. In the case of descriptor type performance criteria, a specific transformation is represented as a curve, defined by varying a distance threshold. Two techniques have been applied to aggregate precision vs. recall [37]and ROC [27] data.

Rather than defining a precision vs. recall plot for a specific transformation, [37] defined the precision and recall values over an entire image sequence. For example, consider you have a 10 image sequence, each viewing a scene at 5^{o} increments. Interest points in each of the images can be described, and these descriptions occupy a common descriptor space. The ground truth dictates that the related points are known, and these

points can be used to form clusters in descriptor space. Rather than associating a feature vector with a distance to its nearest neighbour, a feature vector is associated with a distance to its nearest cluster center. It is this distance that is applied to a threshold which defines it as a correct or false match, allowing a precision and recall value to be obtained for an entire image sequence. Since more than one sequence was examined in this study, an extra step was needed to aggregate over sequences. The authors do not explicitly mention the methods used for aggregation, but the assumption is made that one or two approaches were used: either both the precision and recall values were averaged at common distance thresholds over different image sequences, or the precision and recall values were calculated based off of the number of correct matches, total matches, and ground truth matches of all the image sequences together, rather than at just one. It should be noted that in such an approach the performance of descriptors and detectors as a function of transformation (viewpoint in this case) cannot be analyzed.

In [27], the number of scenes (100, where a scene in this case is defined by a singular object on a turntable) and sequences (2 per object) was much greater than any other evaluation, therefore requiring some form of aggregation. Unfortunately, the method used to aggregate the results that created the ROC curves was not explicitly documented. It is a reasonably assumption that the detection rate and false positive rate were averaged at common distance thresholds (a distance ratio threshold for the matching approach in this study) for all pairs of images for all sequences. Unlike the previously mentioned study, a subsequent analysis graph was created to examine the performance over transformation increments. This was done by fixing the distance threshold based on a specific false

positive rate (10^{-6}) and comparing the averaged detection rate over a series of viewpoint transitions.

3.2.4 Characteristics of Data Sets

The characteristics of a dataset used for evaluation differ depending on the application. Since this evaluation is done in the context of image matching, only datasets relevant to this domain are presented. The characteristics of an image matching dataset help define the performance under certain image conditions such as geometric or photometric transformations and scene type. Ideally, a sequence of images of the same scene exists where each image represents an incremental increase in some transformation property. A criterion value can then be evaluated as the transformation is increased. Figure 9 gives an example of such an image sequence that varies on viewpoint and rotation. The content of the scene plays an important role as well, since the attributes of certain detectors make them better suited for different types of features. Thus, numerous image sequences taken on various types of scenes are also advantageous.

The Mikolajcyzk's Oxford Graffiti dataset has been used in a variety of studies [22], [26], [33], [34]. The dataset contains 8 image sequences, each containing 6 images, all of which can be related via a homography(planar scenes or fixed camera position). Each image sequence is associated with one of the following transformations: scale+rotation, viewpoint, image blur, illumination, and JPEG compression. A subset of transformations (rotation, scale, and viewpoint) are applied to two image sequences, where the scene content differ and are categorized as either textured or structured (distinct edges). [22] examined the repeatability rate and matching score of feature detectors on each image sequence, while [33], [34] evaluated *precision* and *recall* on a select pair of images from

each sequence. Some of the image sequences in the Oxford Graffiti dataset are from the much larger INRIA dataset of planar scenes, which contains 449 images. However, the INRIA dataset in its entirety has not explicitly been used to evaluate feature detectors or descriptors.

Fraundorfer and Bischof [28] evaluated the repeatability and matching score on 2 viewpoint change image sequences, each containing 19 images. This dataset was unique compared to the Oxford Graffiti dataset because the viewpoint change was more extensive (0^o to 90^o degrees) and applied to a 3D scene.

Moreels and Perona [27] also extended the evaluation of feature detectors and descriptors to 3D scenes. Their study, however, contributed to the creation of a much larger dataset than any of the other previous evaluations. Instead of testing only a handful of scenes, their tests included 100 object types viewed from 144 calibrated viewpoints under different 3 lighting conditions.

Gil *et al.* [37] created a dataset under the condition of vision-based simultaneous localization and mapping (visual SLAM) for the evaluation of both detectors and descriptors. Viewpoint, scale, and illumination were the only transformations examined. Viewpoint change was examined for 12 image sequences, each containing 21 images (2.5^o degrees of separation). However, only 4 (one 3D and three 2D) unique scenes were used. The remaining sequences were generated by changing the resolution of the images, which ranged from 320x240 to 1,280x960 pixels. A similar approach was used for the 14 scale change image sequences generated. Two image sequences were used for the analysis of illumination change.

Valgren and Lilienthan [29] studied the performance of SIFT and SURF for use in the application of topological localization on a dataset of images that differed by seasonal change. An omni-directional camera was mounted on a robot, which navigated a route around the local university campus. Every few meters (the actual amount varied) a 360^0 photograph was taken. This exercise was completed 7 times in 9 months, capturing the summer, winter, spring, and 4 variations of the fall (without leaves, with leaves, with leaves and some colour change, less leaves with more colour change). Variations in weather also existed between seasonal sets. For the purposes of comparing the detector-descriptor methods used in the evaluation, 5 location images from each seasonal set were used. In other words, 5 image sequences of outdoor seasonal change, each containing 7 images. Note that in previous image sequences, the transformation was usually the change of one well-defined element (such as viewpoint). In this case, the change is less explicit because images vary by a set of uncontrolled elements related to time of year, day, and weather conditions. The shared location between images in each sequence was not exact, differing by as much as 10 meters.

A summary of the number of scenes, image sequences, and number of images per sequences for each dataset can be seen in Table 2. The results of these studies, categorized by the image condition (i.e. the transformation) can be seen in Table 3.

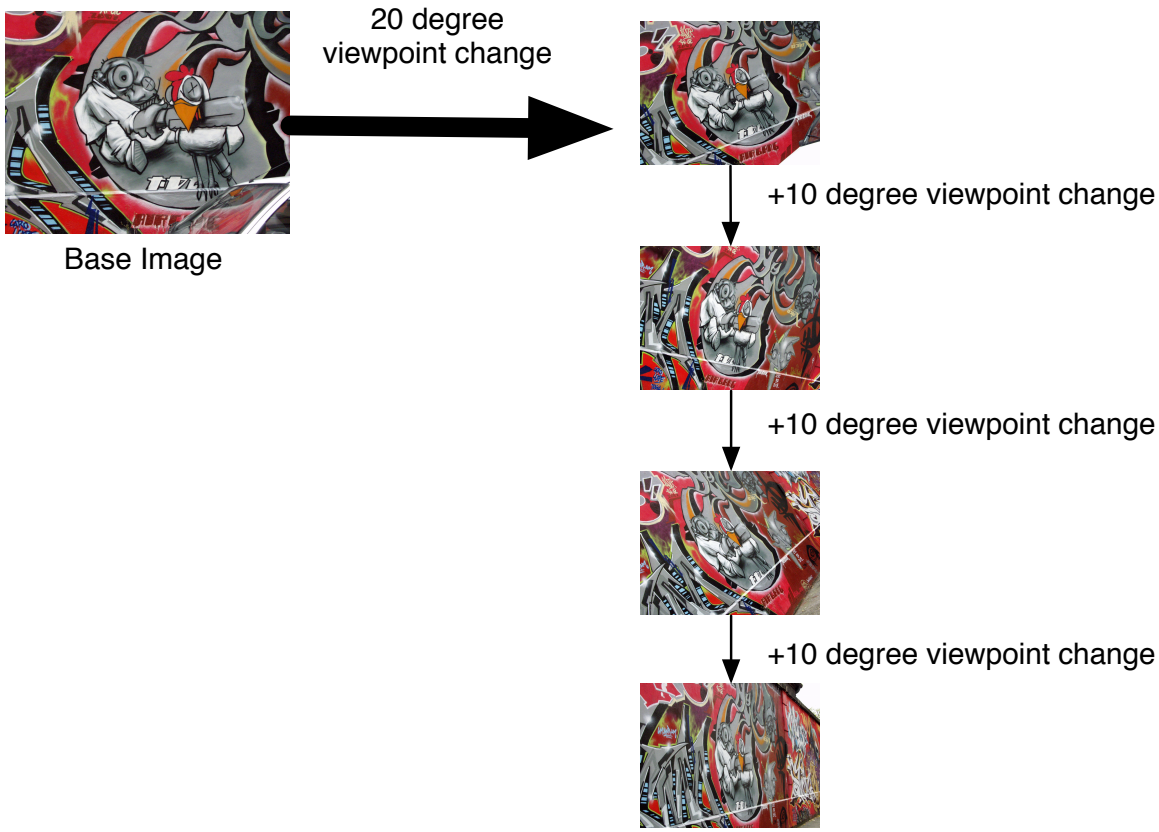


Figure 9 Image Sequence. A typical dataset uses a sequence of images with known change between each image. In this example, Mikolajczyk *et al.*[22] took photographs of a wall at different viewpoint angles, and captured the change by calculating the homography from the base image to each transformed image.

Table 1 Evaluation Context Overview

Evaluation	Detectors	Descriptors	Context	Matching Strategy	Performance Criteria
Mikolajczyk *et al.* [22]	Harris-Affine, Hessian-Affine, MSER, IBR, EBR, Saliency Regions	SIFT	Image Matching	Nearest Neighbour Euclidean Distance	Repeatability, Matching Score, #Correspondences, #Correct Matches
Mikolajczyk and Schmid [23]	Harris, Harris-Laplace, Laplacian, DoG, Gradient, Hessian, Harris-Affine	Steerable Filters	Image Matching	Nearest Neighbour Mahalanobis Distance	Repeatability, Location Error, Overlap Error

Evaluation	Detectors	Descriptors	Context	Matching Strategy	Performance Criteria
Mikolajczyk and Schmid [35]	Harris, Harris-Laplace, Harris-Affine, DoG	SIFT, Steerable Filters, Differential Invariants, Complex Filters, Moment Invariants, Cross Correlation	Image Retrieval	Similarity Threshold Mahalanobis Distance Euclidean Distance	Detection Rate vs. False Positive Rate
Mikolajczyk and Schmid [33]	Harris, Harris-Laplace, Hessian-Laplace, Harris-Affine, Hessian-Affine	SIFT, PCA-SIFT, GLOH, Spin Images, Steerable Filters, Differential Invariants, Complex Filters, Moment Invariants, Cross-Correlation	Image Matching	Similarity Threshold Nearest Neighbour and Nearest Neighbour Ratio Distance Mahalanobis Distance Euclidean Distance	1-Precision vs Recall
Moreels and Perona [27]	Harris-Affine, Hessian-Affine, MSER, Multi-Scale Harris, Multi-Scale Hessian, Saliency Regions, DoG	SIFT, Shape Context, Steerable Filters, Differential Invariants, PCA-SIFT	Image Retriveal, Image Matching	Nearest Neighbour Ratio Distance Euclidean Distance Mahalanobis Distance	Detection Rate vs False Postive Rate
Gil *et al.* [37]	Harris, Harris-Laplace, MSER, DoG, Saliency Regions, Hessian (SURF), SUSAN	SURF, Upright SURF, Extended SURF, SIFT, GLOH, Zernike Moments, Gray Level Patches, Orientation Histograms	Image Matching	Nearest Neighbour Cluster Euclidean Distance	Precision vs. Recall, Repeatability
Schmid *et al.* [20]	Improved Harris, Harris, Foerstner, Cottier, Heitger, Horaud	Differential Invariants	Image Matching	NA	Repeatability, Information Content
Fraundorfer and Bischof [28]	Harris-Affine, Hessian-Affine, MSER, IBR, EBR, Saliency Regions	SIFT	Image Matching	Neartest Neighbour Euclidean Distance	Repeatability, Matching Score, #Correct Matches
Valgren and Lilienthal [29]	DoG, Fast-Hessian (detector for SURF)	SIFT, SURF, Upright SURF, Extended SURF	Image Matching / Image Retreival	Nearest Neighbour Ratio Distance	Precision and Number of Correct Matches

Table 2 Evaluation Dataset Overview

Evaluation	#Scenes	#Sequences	#Images per Sequence

Evaluation	#Scenes	#Sequences	#Images per Sequence
Mikolajczyk *et al.* [22]	8 (approx. 2 scenes per transformation)	8	6
Mikolajczyk and Schmid [23]	10 (scale) 6(viewpoint)	10 (scale) 6(viewpoint)	10
Mikolajczyk and Schmid [35]	4	4	3
Mikolajczyk and Schmid [33]	10 (approx. 2 scenes per transformation)	10	2
Moreels and Perona [27]	100(viewpoint and illumination) 5(scale)	200 (viewpoint, 2 camera views with a image every 5 degrees) 42600(illumination 3 light changes, 142 viewpoints, 100 objects) 284 (scale, 4 scales levels, 142 viewpoints, 5 objects)	71(viewpoint) 3(illumination) 4(scale)
Gil *et al.* [37]	4	12(viewpoint) 14(scale) 2(illumination)	21(viewpoint) 12(scale) 13(illumination)
Schmid *et al.* [20]	2	8	16(viewpoint) 10(scale) 22(illumination)
Fraundorfer and Bischof [28]	2	2	19
Valgren and Lilienthal [29]	5	5	7

Table 3 Evaluations Results Overview

Evaluation	Viewpoint	Scale	Rotation	Illumination	JPEG Compression	Blur	General	Comments
Mikolajczyk *et al.* [22]	MSER	Hessian-Affine (Scale +Rotation)		MSER	Hessian-Affine	Hessian-Affine		
Mikolajczyk and Schmid [23]	NA	Harris-Laplace	NA	NA	NA	NA		Harris-Laplace outperformed Harris-Affine in Scale experiment
Mikolajczyk and Schmid [35]	SIFT	SIFT (Scale + Rotation)	SIFT Steerable Filters Cross-Correlation	Steerable Filter	NA	NA		In the rotation experiment, descriptors paired with the Harris performed better than descriptors paired with the Harris-Laplace In the scale experiment, descriptors paired with DoG slightly outperformed the Harris-Laplace Illumination experiment suggests photometric changes have less significance than geometric transformation (for descriptors)
Mikolajczyk and Schmid [33]	GLOH (structured scenes) SIFT (textured scenes)	GLOH	GLOH SIFT Shape Context	GLOH SIFT	GLOH SIFT PCA-SIFT	GLOH PCA-SIFT	Generally, both GLOH and SIFT showed the best results. For structured scenes, Shape Context also performed well.	In the scale experiment, the Hessian-Laplace performed better than the Hessian-Affine. In the viewpoint test, the Hessian-Affine outperformed the Harris-Affine. viewpoint, blur, and jpeg compression were the most difficult cases for the descriptors to handle.
Moreels and Perona [27]	Hessian-Affine/SIFT (3D scenes) DoG/SIFT (3D scenes) MSER/SIFT(2D scenes)	Harris-Affine/SIFT	NA	Harris-Affine/SIFT	NA	NA	Overall best overall choice is Hessian-Affine or Harris-Affine paired with SIFT or Shape Context	
Gil *et al.* [37]	Harris GLOH, SURF	Harris(3D scenes) MSER(2D scenes) GLOH(2D,3	NA	Harris GLOH SIFT SURF	NA	NA	Harris performed best for detectors in the context of repeatability, while GLOH and	Illumination change was non-linear. In this circumstance, MSER performed very poor. Harris features were used for all descriptor experiments

Evaluation	Viewpoint	Scale	Rotation	Illumination	JPEG Compression	Blur	General	Comments
		D) SURF(3D)					SURF performed best for local descriptors	
Schmid *et al.* [20]	Improved Harris	Improved Harris	Improved Harris	Improved Harris (both unifrom and non-uniform illumination)	NA	NA	Improved Harris also had the highest information content rating	A camera noise test was also done. The majority of the detectors tested were fairly resilent to such change.
Fraundorfer and Bischof [28]	MSER (piecewise planar scene) DoG (piecewise planar and complex scene) IBR (complex scene)	NA	NA	NA	NA	NA		Performance improved when using all the detectors together. In the complex scene (an office), viewpoint changes over 30 degrees return very few correct matches
Valgren and Lilienthal [29]	NA	NA	NA	NA	NA	NA	SIFT was able to find more correct matches, but at a lower precision compared to the SURF methods. U-SURF was determined to perform best when both criteria were taken into consideration	Difficult cases related to large changes in lighting direction (backlit compared to frontlit scenes) and the introduction of snow coverage.

Chapter 4 Feature-Based Matching

This thesis evaluates the performance of feature based matching system on repeat photography. This chapter is dedicated to the explaining the different detectors and descriptors in addition to the matching approach used to form these matching systems. The output of such a matching system is a set of proposed matches (a mixture of correct and false matches). Therefore, a final stage in matching is usually applied to remove outliers (false matches) and estimate the transformation between images. A common method used for such a task, RANSAC, is discussed in Section 4.4.

4.1 Feature Detection

4.1.1 Local Features and Feature Detectors

Local features are pertinent locations within an image that are significant because of an image pattern oriented around the feature location. What constituents a pattern of relevance depends on the method and the application. These patterns can traditionally be thought of as edges, corners, junctions, blobs, or segments. The advantage of being able to extract features from an image is that further methods can be used on a subset of areas within an image, rather than having to process each location of the image canvas. Consider the example of a landscape photograph, where half of the image shows blue sky, and the other half shows land. The homogenous nature of the sky provides very little information for further analysis (unless this is specifically being analyzed), while the ridges and texture of the land are much richer in information.

A local feature can be a point or a region (centered on a point). A feature can be covariant for certain transformations. A feature is covariant when its shape or size changes with the transformation that is applied to an image. For example, if a feature is

detected with a 10 pixel radius circular region, and then the same image is scaled by a factor of 2, the newly detected feature is said to be covariant if it is located at the same position and has a radius of 20 pixels. In literature this attribute is usually referred to as invariant (e.g. scale-invariant or affine-invariant). The misuse of this term was pointed out by[31], though invariance is still the convention and thus is used in this thesis.

The method that extracts local features from images is called feature detection.

4.1.2 Properties of Local Features

While the exact definition of a local feature depends on the method, the qualities of good features and the detector that produces them have been well classified by [31]:

- *Repeatability:* A repeatable feature is one that is persistent over transformation. The easiest example is that of a viewpoint change. By changing perspective the feature should still be detected as an area of interest; it should be repeatable. This quality is synonymous with stability, and is generally regarded as the most important.
- *Distinctiveness/Informativeness:* The patch where the local feature is located should be distinct in its intensity patterns in comparison to other features. The more distinct a feature, the easier it is to match and distinguish.
- *Accuracy:* The detected features should be accurately localized both in image location (coordinates) and scale. Scale determines the size of the local feature. More specifically, a feature is more accurate if its transformed counterpart is localized exactly in location, scale, and shape.

- *Locality:* The local area defining the feature should be small enough as to not encompass occlusions. This quality comes at the cost of distinctiveness (larger regions have more information).
- *Quantity:* Generally, the more features the better, but this is application specific. It is desirable to have a simple threshold that can control the amount of features detected.
- *Efficiency:* It is desirable to have a feature detector that runs efficiently. However, how efficient and the cost at which that efficiency is obtained depend on the application.

As emphasized throughout, the quality of a detection method and its associated features can depend highly on the application. In the context of matching scenarios, repeatability, accuracy, and distinctiveness are regarded as the most important, and this is demonstrated through their use as performance criterion in a variety of evaluations.

4.1.3 Feature Detectors used in this Evaluation

When choosing a feature detector for an application it is important to select detectors that are known to perform well on certain types of image structures and handle appropriate levels of invariance. While some evaluations have considered different image types [22], [23] (textured and structured images) these categories were loosely defined and only tested on a few image sequences. Explicitly categorizing our dataset as textured or structured (well defined edges) is difficult. While landscape scenes are inherently very textured, it is the texture that is susceptible to change, while large-scale geographical contours are pertinent over time. Instead of choosing feature detectors based on image

type, we choose 4 detectors that fall into 3 different image structure categories: corners, blobs, and regions.

Since our images are oblique landscape photographs, they inherently capture objects at a variety of scales. Furthermore, as mentioned in the introduction, a practical problem within the MLP operating process is the alignment of repeat image pairs that differ in scale. We are therefore motivated to use scale-invariant detectors.

The detectors and implementations that have been used in this evaluation are given in Table 4. We also include the type of feature (as specified in [31]) for which each method is categorized.

Table 4 Feature detection methods, implementations, and category of feature type.

Method	**Implementation**	**Feature Type**
Harris-Laplace [23]	h_affine.ln [39]	Corner
Hessian-Laplace [22]	h_affine.ln [39]	Blob
Difference of Gaussian (subsequently referred to as DoG) [17]	VLFeat toolbox Matlab [40]	Blob
Maximally Stable Extremal Regions (subsequently referred to as MSER) [41]	computer_descriptor.ln(2010) [42]	Region-based

These detectors have been well evaluated and are commonly used in literature [31]. Three of the detectors evaluated, the Harris-Laplace, Hessian-Laplace, and DoG, are

based on differential functions of the images. These forms of detection are grouped into a 3-stage process of saliency map creation, multi-scale representation, and automatic scale selection. In the proceeding sections, these stages are discussed while describing the Harris-Laplace and Hessian-Laplace detectors. While the DoG detector does follow a similar scheme, the simultaneous choice of scale and salient features make it difficult to incorporate into these groups, and therefore it is discussed separately. Finally, the MSER detector, which is conceptually different than the others, is reviewed. Examples of detected features for each of the detectors are presented in Figure 10.

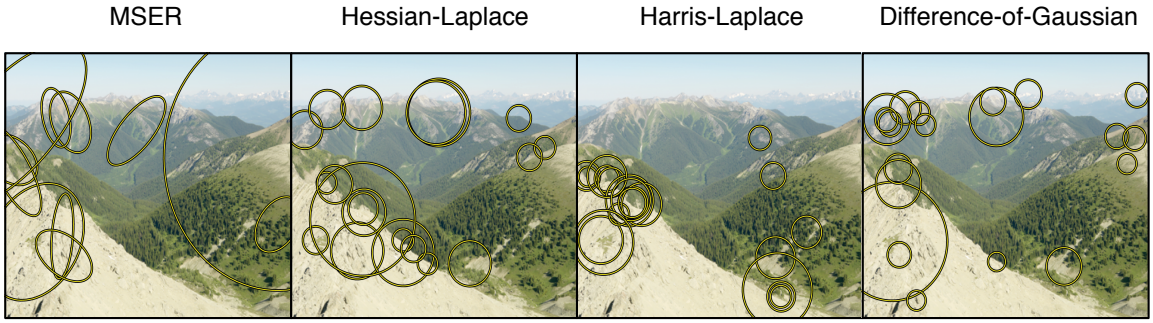


Figure 10 Detected features for each of the detectors used in this evaluation. Hessian-Laplace, Harris-Laplace, and DoG are scale-invariant features represented by circular regions associated with scale. MSER features are affine-invariant and represented by ellipses.

4.1.3.1 Saliency Maps

In the first step of feature detection an image is processed with an operator, which produces a value at each pixel that is a function of the local information content (e.g. gradients, gray-level values) around that pixel. The result can be described as a saliency map[27]. Features are then selected based on the local extrema of the saliency map and the requirement that the saliency value surpass a specified threshold. One aspect that distinguishes the Harris-Laplace and Hessian-Laplace detectors is the function they use to define their saliency map.

4.1.3.1.1 Harris

The Harris detector is based off the second moment matrix (also known as the auto-correlation matrix). This matrix is formed with the first derivatives of a local region and describes the local gradient distribution within an image patch.

$$M(\mathrm{x},\sigma_I) = g(\sigma_I)\cdot\begin{bmatrix} I_x(\mathrm{x})^2 & I_x(\mathrm{x})I_y(\mathrm{x}) \\ I_x(\mathrm{x})I_y(\mathrm{x}) & I_y(\mathrm{x})^2 \end{bmatrix} \tag{4.1}$$

with

$$g(\sigma) = \frac{1}{2\pi\sigma^2} e^{-\frac{x^2+y^2}{2\pi\sigma^2}} \tag{4.2}$$

Where M is the second moment matrix, g is Gaussian kernel, and I_i represents the first derivative for the i^{th} dimension. M is parameterized by a location **x** in the image, and the Gaussian parameter (known as the integration scale), which defines the characteristics of the kernel used to perform convolution on the gradient images. We mentioned in the previous paragraph that this matrix describes the *local* gradient distribution. With this in mind, the integration scale parameter defines how *local* of an area the user wants to use to calculate their Harris measure. This is not to be confused with the scale associated with a scale space level, which is discussed later. The principle eigenvector of this matrix represents the principle direction of gradients in the neighbour defined by **x** and σ, while the second eigenvector is naturally orthogonal to this direction. The size of the eigenvalues represent the strength of the gradients in these directions. Thus, for a pixel located on a 2D corner, the eigenvalues have a near equal size, since the gradients of a corner are directed orthogonal to each other. Alternatively, a pixel located on an edge has eigenvalues of disproportionate size, since the gradient is dominant in only one direction.

Harris proposed a more compact and efficient version of this concept with the *cornerness* metric, which can be described as:

$$cornerness = \det(M) - \lambda \, \mathrm{trace}(M)^2 \tag{4.3}$$

This metric is motivated by an important property of the second moment matrix: the determinant of the matrix is equal to the product of the eigenvalues and the trace equal to the sum. Thus, a large *cornerness* value correspond to large eigenvalues [31]. The λ is used to mitigate response of the function to straight edges with strong gradients. To find the final set of Harris points, local maxima are selected from the 8-neighbourhood of each point. Additionally, a threshold is applied to the cornerness value at each of these maxima to filter out interest points of small cornerness. These points are considered invariant to rotations[31] because of the symmetric nature of the matrix. The saliency map for the Harris detector is the cornerness value for each pixel location of the image.

4.1.3.1.2 Hessian

The 2x2 Hessian matrix for a point **x** in image I can be expressed as:

$$H(\mathrm{x}) = \begin{bmatrix} I_{xx}(\mathrm{x}) & I_{xy}(\mathrm{x}) \\ I_{xy}(\mathrm{x}) & I_{yy}(\mathrm{x}) \end{bmatrix} \tag{4.4}$$

where I_{xx} is the second derivative of the image. A commonly used operator based on the Hessian matrix is the determinant of the Hessian:

$$\det(H) = I_{xx}(\mathrm{x}) \cdot I_{yy}(\mathrm{x}) - I_{xy}(\mathrm{x})^2 \tag{4.5}$$

The determinant of the Hessian is computed for each point **x** in the image. If the point is a maxima in its local 8-neighbourhood and above a threshold value, then it is considered a Hessian interest point, forming the final saliency map. These extrema are localized at the centre of blobs, points and/or regions that are brighter or darker than the surrounding area.

4.1.3.2 Scale Space Representation

In order to handle features at multiple scales, a feature detector must be invariant to scale change. This invariance can be achieved by applying a saliency map type detector discussed above to a scale space representation of image, which can be defined as a one-parameter family of blurred images parameterized by the size of a smoothing kernel which suppresses finer structures as the kernel size is increased [43]. Koenderink [44] showed the unique kernel for generating space scale was the Gaussian. Hence, a scale space level can then be formed by

$$I(\mathrm{x},\sigma) = g(\sigma)\cdot I(\mathrm{x}) \tag{4.6}$$

the convolution of the image with the 2D Gaussian, where σ is a parameter to the Gaussian function that defines the size of the kernel and correspondingly the scale space level. This notion of scale needs to be incorporated into the Harris and Hessian function defined earlier. Fortunately, the differentiation of an image and the creation of scale can be done simultaneously with normalized Gaussian derivatives, making definition of the multi-scale versions of these detectors fairly straightforward. The non-normalized Gaussian derivative of I can be expressed as

$$I_{i_1\ldots i_m}(\mathrm{x},\sigma_D) = \frac{\partial}{\partial i_1 \ldots i_m} g(\sigma_D)\cdot I(\mathrm{x}) \tag{4.7}$$

With non-normalized Gaussian derivatives, the amplitude of the filter decreases with an increase in scale, which decreases the derivative value of x when applied to the image. In order for this function to be fully scale-invariant, this decrease must not occur (the response of the function should constant, independent of a scale change). A normalized Gaussian derivative achieves this property and can be defined as

$$D_{i_1\ldots i_m}(\mathrm{x},\sigma) = \sigma^m I_{i_1\ldots i_m}(\mathrm{x},\sigma) = \sigma^m g_{i_1\ldots i_m}(\sigma)\cdot I(\mathrm{x}) \tag{4.8}$$

where *m* is the order of the derivative and subscripts *i* denote the Cartesian planes of the partial derivative. This allows for the formation of multi-scale saliency maps for both the Hessian and Harris methods discussed earlier. For the multi-scale Harris is based off the multi-scale second moment matrix

$$M(\mathrm{x},\sigma_I,\sigma_D) = g(\sigma_I)\cdot\begin{bmatrix} D_x(\mathrm{x},\sigma_D)^2 & D_x(\mathrm{x},\sigma_D)D_y(\mathrm{x},\sigma_D) \\ D_x(\mathrm{x},\sigma_D)D_y(\mathrm{x},\sigma_D) & D_y(\mathrm{x},\sigma_D)^2 \end{bmatrix} \tag{4.9}$$

The matrix is parameterized by location **x**, the integration scale σ_I. and the new differentiation scale σ_D. Note the difference between σ_I and σ_D; the integration scale is associated with defining the local area used to calculate the Harris measure at a point, while the differentiation scale is the scale space parameter (computing scale invariant gradients). These two parameters are merged into one by making the differentiation scale a multiple of the integration scale. The *cornerness* measure is still defined the same way. The difference is that *cornerness* values can be computed at multi-scales. Baumberg [45] and Dufouraud *et al.* [46] used the multi-scale Harris method for matching.

The multi-scale Hessian can be formed by using the normalized derivatives.

$$H(\mathrm{x},\sigma_D) = \begin{bmatrix} D_{xx}(\mathrm{x},\sigma_D) & D_{xy}(\mathrm{x},\sigma_D) \\ D_{xy}(\mathrm{x},\sigma_D) & D_{yy}(\mathrm{x},\sigma_D) \end{bmatrix} \tag{4.10}$$

The determinant of Hessian still calculates the saliency measure, but now at multiple scales.

The problem introduced by having a multi-scale representation of the saliency maps is that the number of features increases considerably and multiple features may occur on the same objects, but only at slightly different scales. To remedy this problem, automatic scale selection is used.

4.1.3.3 Automatic Scale Selection

Automatic scale selection is performed by searching for local extrema of a differential function in scale space[47]. It is used to locate the scales at which the local structure is most characteristic. Lindeberg showed that a variety of different functions (based on image derivatives) can be used to search for the characteristic scale. Mikolajczk [48] evaluated several different functions (the squared gradient, Laplacian, determinant of the Hessian, and the Harris function) for the purpose of scale selection and found that the Laplacian performed best, forming the motivation for the Harris-Laplace and Hessian-Laplace feature detectors. The multi-scale versions of the Harris and Hessian have automatic scale selection performed with the Laplacian. That is, if a feature located at **x** is considered to be the characteristic scale if $F(\mathrm{x},\sigma_n) > F(\mathrm{x},\sigma_{n+1})$ and $F(\mathrm{x},\sigma_n) < F(\mathrm{x},\sigma_{n-1})$, where the subscripts $n\text{-}1$ and $n+1$ denote the scale levels above and below the scale level n, and F is the normalized Laplacian function

$$F(\mathrm{x},\sigma_D) = \left|D_{xx}(\mathrm{x},\sigma_D) + D_{yy}(\mathrm{x},\sigma_D)\right| \tag{4.11}$$

4.1.3.4 Difference-of-Gaussian

The DoG method, like the Hessian-Laplace and Harris-Laplace is a scale invariant saliency map based method, but differs on the previous approaches in that it uses the same function to detect both the saliency map and scale space selection in addition to having a much more efficient design. The DoG as a method for detecting interest points has been proposed by several authors [49], [50], but was later refined by Lowe in [17], [51] as the feature detector in SIFT. In reference to the DoG technique in this thesis, we refer directly to the Lowe refinement of this method. The DoG function can be described as follows

$$DoG(\mathrm{x},\sigma) = \left| g(\sigma)\cdot I(\mathrm{x}) - g(k\sigma)\cdot I(\mathrm{x}) \right| \quad (4.12)$$

where g is the Gaussian kernel, σ is the scale parameter and k is a constant scale factor between scale levels. The saliency map for a particular scale level is the difference of two Gaussian smoothed images that correspond to discrete neighbours in scale space. Lowe points out two benefits to this approach. The first is that the scale space levels of image I need to be created for the feature description phase regardless of the function used for feature detection. Therefore, the subsequent step of subtraction between two scale space layers comes at a relatively low cost (compared to the convolution of a normalized derivative). The second advantage is that the DoG function is an approximation to the normalized Laplacian operator, meaning that it is also scale-invariant and a good operator for scale selection.

The Harris-Laplace and Hessian-Laplace select salient features in each scale level by searching for local extrema in 8-neighbourhoods and then selecting scale based on local extrema of neighbouring levels. In the DoG approach, scale and saliency are selected simultaneously. A feature is detected at a 3D local extrema in scale space. This concept is shown in Figure 11. Each of the layers shown is associated with a DoG scale space layer. The location denoted by x is chosen to be a salient point if it is greater or less than its 24 neighbours (denoted by dots). These local extrema are further filtered by assuring that the eigenvalue ratio of the Hessian matrix at these points was above a threshold, which ensures that features are not located on straight edges.

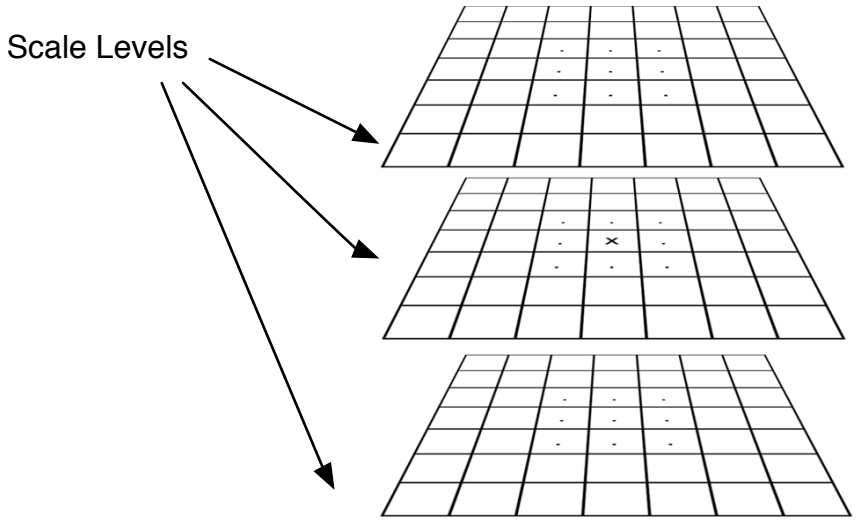


Figure 11 An example of how saliency points are chosen for the DoG detector. If a DoG value is greater than its 24 neighbours in scale space, then it is chosen for further processing.

The DoG method developed by Lowe also used a pyramid scale space representation, a more efficient manner of storing scale information in addition to requiring less computation.

In Lowe's implementation, each pyramid level contains images with half the resolution or the previous pyramid level, with the exception of the base level where the images are at the native resolution. Hypothetically, each pyramid level could contain only one image, but then the scale space would not be sufficiently dense; only scales of a factor of $2^{1/p}$ would be represented, where p is the number of pyramid levels. To increase the density of scale space a Gaussian kernel of increasing size is used to create intermediate scale levels in each pyramid level. These intermediate levels at each pyramid level are called the octaves (the set of images at each pyramid level). The density of DoG scale space is then specified by two parameters: the number of levels per octave, and the number of octaves. A pictorial representation, similar to the one shown in [17] can be seen in Figure 12.

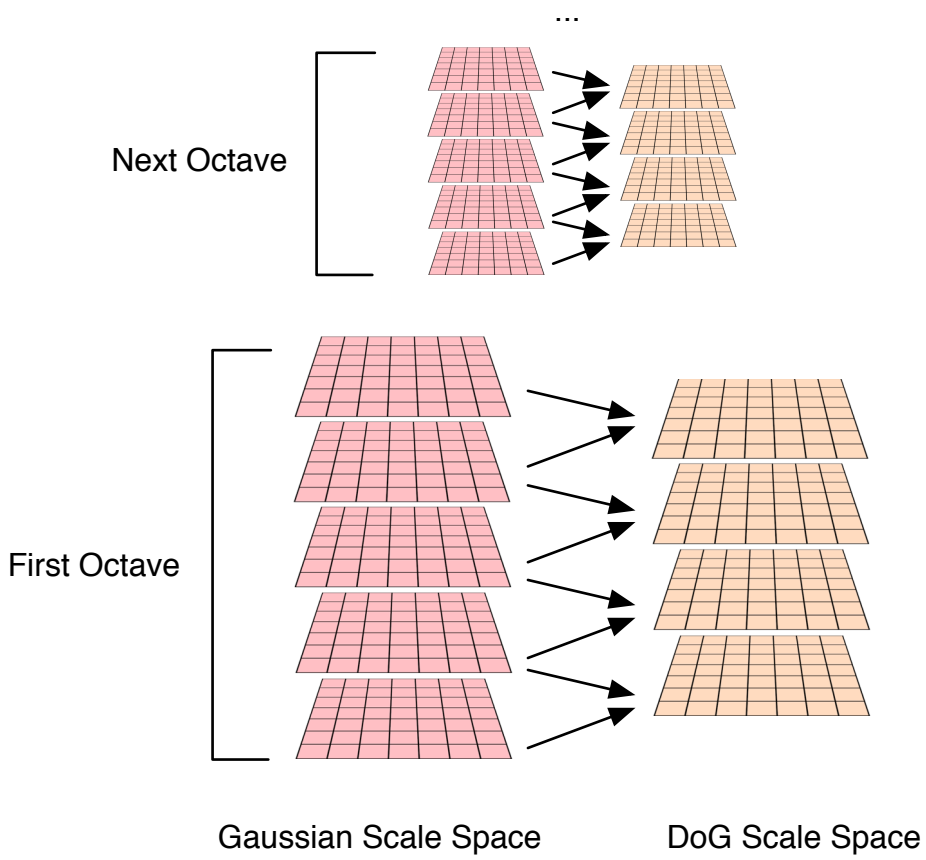


Figure 12 The creation of DoG scale space. First a set of images are convolved with the Gaussian kernel with a progressively increase scale space parameter. The difference of each of these images is used to form the DoG scale space. When the scale parameter is set to twice its initial setting, the image is resampled, and the process is repeated to form the next octave.

4.1.1.5 Maximally Stable Extremal Regions

Maximally Stable Extremal Regions [41] are conceptually different from the saliency map type detectors discussed previously in that they are not based on differential operators. Tuytelaars and Mikolajczyk [31] classified MSER as a region based detection strategy. The concept behind MSER is best explained with an example. Consider a grayscale image with pixel values ranging from 0 (black) to 255 (white). Now also consider some binary threshold value *n*. When the threshold is applied to the image, the pixels are divided into groups that are above and below (and equal to) the value *n*. An

example of such threshold can be seen in Figure 13. Each group can further be represented as a set of connected components, or regions. These regions are defined as extremal because every pixel within the region is brighter or darker than its surroundings. As the threshold *n* is moved from 0 to 255, these regions change size, disappear, and appear. A maximally stable extremal region is then one that has a stable size over a range of thresholds. It should be noted that the example given in Figure 13 (viewpoint change of planar surfaces with distinct edges) performs well with the MSER detector.

Figure 13 The image above shows two images of the same scene that differ on viewpoint change. The black and white images in the rows adjacent to these images show two binary threshold images. The thresholds used are 100 and 150. The green arrows demonstrate an example of a connected component that is stable over an intensity change of the surrounding area. Note that the feature is consistent over viewpoint change.

More formally this process can be described as follows. Each pixel is first grouped by intensity value. The pixels in each intensity group are then partitioned into connected components based on their adjacency to one another. The next step involves the merging of connected components between intensity levels. First, the largest component is found. If there are connected components in the above (or below) intensity level that are adjacent to the large component, then they are absorbed into it and the change in

component size is recorded. This process is continued until the component cannot grow anymore. This merged connected component is called an extremal region. The next largest component is then selected and merge continues as described above, until all extremal regions are represented. Finally, only the extremal regions that have a low (below some threshold) rate of area change are considered to be an output feature.

These regions are affine invariant in additional to being able to handle monotonic change in image intensities.

While the regions detected by this method are of variable shape, the MSER detector used in this evaluation uses elliptical approximations to these regions in a similar manner to [22].

4.2 Feature Description

The methods described in the previous section are associated with finding points or regions such that these same features are detected even after the image or scene has undergone a geometric or photometric transformation. This process reduces the number of regions and points to a relevant set that can be used for further processing. In a feature-based matching system the next step is the description phase; the conversion of the local area around a feature into an array of numeric values that can be used to uniquely describe the region. This step is known as feature description, while the numeric array associated with the feature is referred to as the feature vector. The regions that are described are commonly referred to as support regions of feature description. In this study, we focus on the two gradient distribution based techniques shown in Table 5.

Table 5 Feature description methods and implementations

Method	Implementation
Scale Invariant Feature Transform (SIFT) descriptor [17], [51]	compute_descriptors.ln (2005) [39]
Shape Context descriptor [52]	compute_descriptors.ln (2005) [39]

These types of descriptors construct their feature vectors using a histogram of local gradient magnitudes and orientations. Both methods have been shown to perform well in matching type scenarios [33].

Two pre-processing steps are usually applied to the support regions before they are converted into feature vectors: region normalization and rotation invariance. These steps are first described, after which we introduce the feature descriptors. A third common pre-processing step to attain illumination invariance has been implemented in different ways by each of the descriptor methods used. Due to this distinction, illumination invariance is discussed in conjunction with the particular algorithm.

4.2.1 Region Normalization

The inputs to the description phase are the detected features. As we've seen with the detectors described in Section 4.1, they can be points, circles, and ellipses of various sizes. Mikolajczyk *et al.* [22] found that by increasing the size of detected feature used for description, matching performance could be improved since a larger amount of content made the subsequent feature vector more distinct among other features. Thus, the regions used for description in this study are 3 times the size of their detected counterpart (a similar increase was used in [22], [33]). These support regions are then normalized to a fixed size of *NxN* pixels[22]. Therefore, the descriptor region for a point is just the *NxN* region centered around that point. With circular regions of various scales, the region must

be resized in order to fit into the *NxN* region. In the case of affine regions, they must first be transformed to a circle, and then resized to fit the *NxN* descriptor region. In this study regions are normalized to a 41x41 descriptor region, a similar size used in [33], [34].

4.2.2 Rotation Invariance

A descriptor is considered to be rotational invariant if the same feature vector is produced regardless of a rotation applied to the patch being described. Some descriptors require a pre-processing step to account for such rotation. Both SIFT and Shape Context descriptors fall into this category. To account for possible rotations of the regions, the dominant gradient orientation is calculated, and the region is rotated to align with this dominant orientation. This is achieved by first creating a gradient magnitude weighted orientation histogram for a small neighbourhood around the centre of the region. The dominant orientation is then determined by looking at the largest bin in the histogram [17].

4.2.3 SIFT Descriptor

The SIFT descriptor was developed by Lowe[17], [51] as part of his feature transform method. Given a region within an image, the magnitude and orientation (quantized into 8 orientations) of the gradients are calculated. The region is then divided into a 4x4 grid, and for each grid cell an 8 level histogram is calculated based on the orientations. The 16 (4x4) histograms are then concatenated to form a 128-element vector that is used for description. The grid size and number of orientation bins can technically be parameterized to change the size of the histogram, but the method described above is the standard usage of the descriptor. The construction on this feature vector is shown in Figure 14.

A technique for making the feature vector invariant to illumination change was also integrated into the SIFT descriptor. The 128-element vector is normalized so that its vector length is 1. This normalization makes the descriptor invariant to linear illumination because both multiplicative and additive changes to intensity values of a region do not affect the direction of the vector. SIFT also accounts for non-linear illumination change by clipping the normalized vector for values larger than 0.2. The vector is then renormalized to the unit length. The motivation behind this method is that non-linear illumination changes more likely affect gradient magnitude, rather than overall gradient orientation [17].

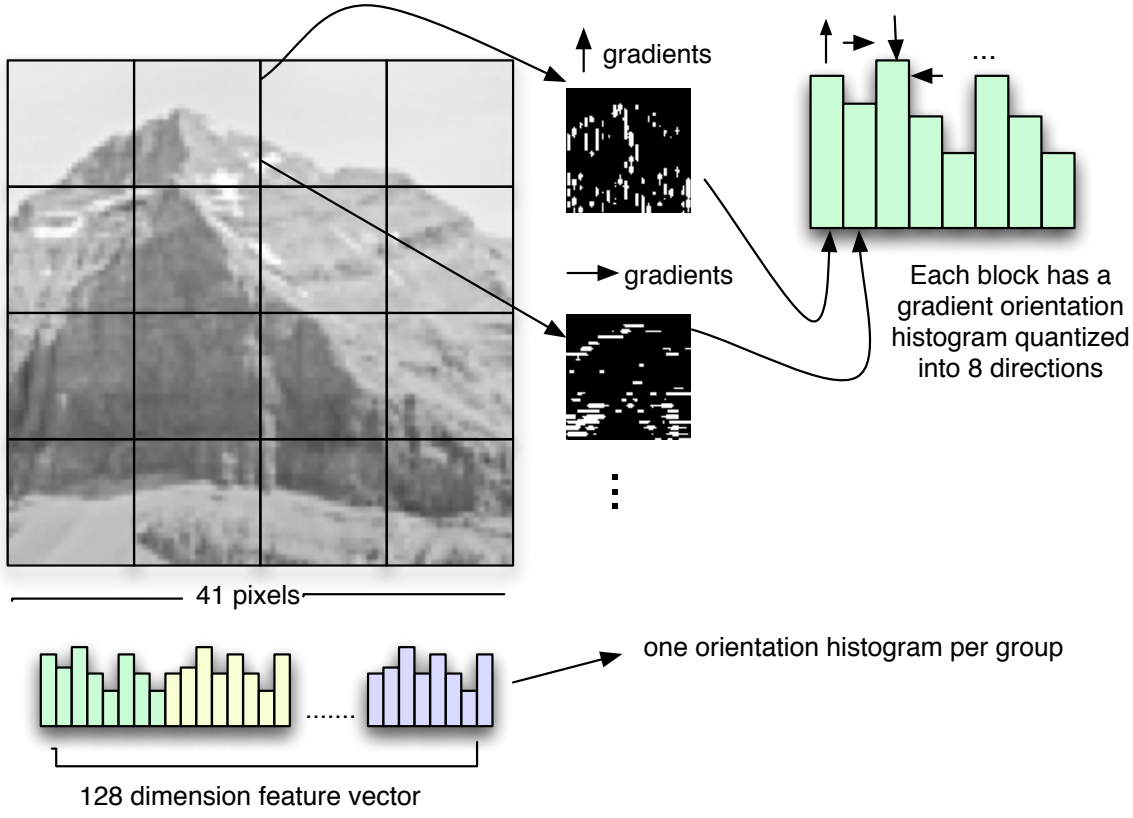


Figure 14 SIFT descriptor. A feature vector is constructed from orientation histograms located at different areas of the region to be described.

4.2.4 Shape Context

The Shape Context descriptor first processes the local region with the Canny edge detector[53]. The region is then divided into sections based on a log-polar coordinate grid. The number of edge points within each section then contributes to a 2D histogram. This histogram is then serialized creating the final feature vector description. This original Shape Context descriptor was designed particularly for finding correspondence between shapes.

Mikolajczyk [33][33] later adapted this method for describing regions from grayscale images. In particular, the 2D histogram was extended to a 3D histogram by additionally counting orientation at cell position. The gradient magnitude was also used to weigh the point contribution to the histogram. 9 log-polar bins were used and orientation was quantized into 4 directions, resulting in a 36-element descriptor. This modified version of Shape Context was used in this evaluation. Figure 15 gives an overview of how the Shape Context descriptor is created.

Illumination invariance for Shape Context is done by using the mean and standard deviation within the support region [35]

$$I'(\mathrm{x}) = I(\mathrm{x}) - \frac{mean(I(\mathrm{x}))}{std(I(\mathrm{x}))} \tag{4.13}$$

where $I(\mathrm{x})$ is the support region rather than the entire image. This method has shown good results for affine transformations in illumination [35][35].

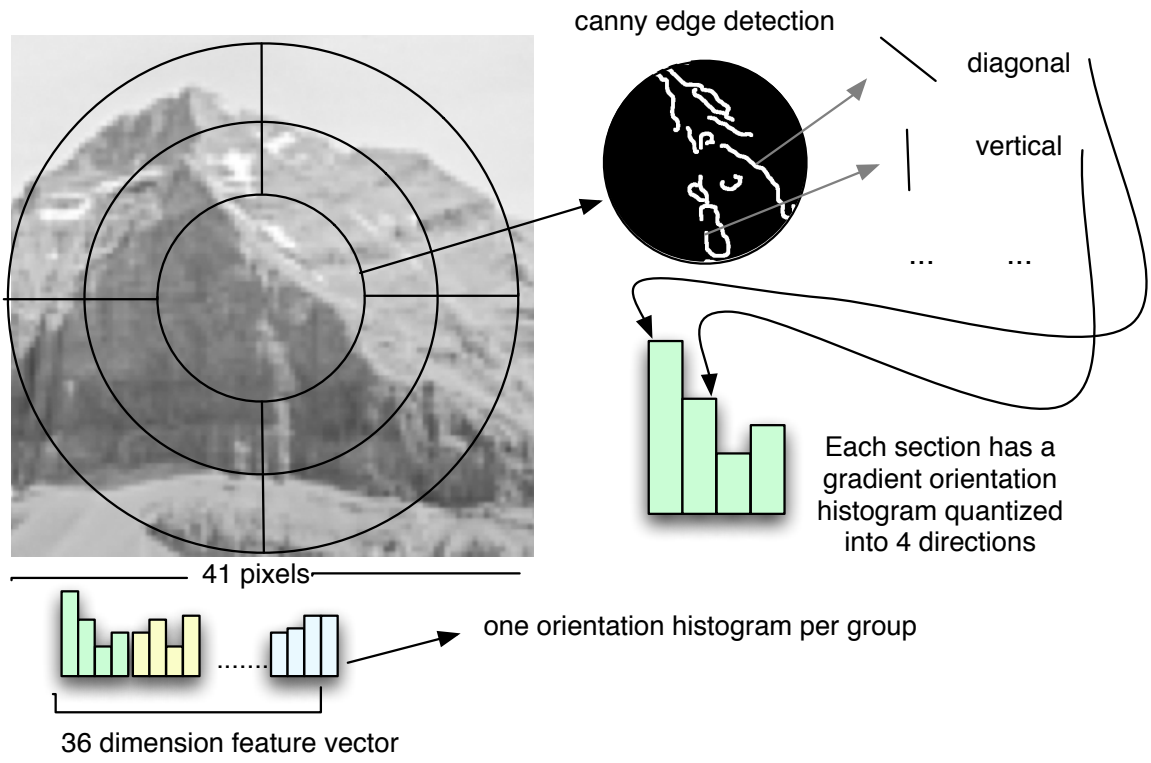


Figure 15 Shape Context descriptor. Instead of dividing the region into a grid, this descriptor uses log polar bins. A canny edge detector is then used to find edges in each location. The gradient orientation of these edges is used to formulate orientation histogram, where direction is quantized into 4 directions.

4.3 Feature Matching

The evaluation in this thesis is done in the context of the matching of features between image pairs. A feature-based matching system involves detection, description, and a matching strategy. The 3 most common matching strategies are 1) similarity threshold, 2) nearest neighbour, and 3) nearest neighbour distance ratio. Two feature vectors **A** and **B** are separated by a distance DAB. The similarity threshold defines matches based on a simple distance threshold. Thus, **A** and **B** are a match if $\mathbf{D_{AB}}$ falls below some threshold. In this circumstance, a feature vector from one image can match more than one feature

vector from another image. For example, consider the 2D descriptor space example in Figure 16. If the smaller distance threshold is used, there is only one match, but if the larger distance threshold is used, there are two because both are within the distance threshold bounds. For the nearest neighbour strategy, A is matched to B if they are neighbours in descriptor space (no other feature vector is closer to A than B), and the distance between them is also below a threshold. In this circumstance, a feature vector from one image can match only one other feature vector from another image. For example, when the larger distance threshold is used in Figure 16, only the closest red triangle is matched even though both are within the threshold bounds. The final strategy, nearest neighbour ratio distance, was introduced as part of Lowe's SIFT method for finding and matching points in a scale-invariant manner[17]. Instead of thresholding the distance between nearest neighbours, the ratio between the first nearest neighbour distance and the second nearest neighbour distance are compared to a threshold. The motivation behind this approach is that a feature vector is considered less distinct if it has many neighbours close to it. There is a greater chance that the true match is not the nearest neighbour, but rather one of the neighbours that are a bit further away. This can be further exhibited in Figure 16. In the descriptor space to the right the distance of the second neighbour has moved closer to the reference feature vector (the green triangle). As shown, the ratio for this scenario, 0.9, is higher than the other case meaning that it is not considered a match. Such a strategy inevitably returns fewer matches than the others.

The distance between feature vectors can be computed in different ways. Two of the most common distance measures are the Euclidean and Mahalanobis distances. The

Mahalanobis distance takes into account the covariance of the feature vector dimensions, while the Euclidean distance can be formulated as

$$dist(\mathrm{A},\mathrm{B}) = \sqrt{(a_1 - b_1)^2 + (a_2 - b_2)^2 + \ldots + (a_N - b_N)^2} \tag{4.14}$$

$$\mathrm{A} = (a_1, a_2, \ldots, a_N) \tag{4.15}$$

$$\mathrm{B} = (b_1, b_2, \ldots, b_N) \tag{4.16}$$

where A and B are feature vectors of size N.

For the experiments in this thesis, nearest neighbour matching with the Euclidean distance measure has been used as our matching approach.

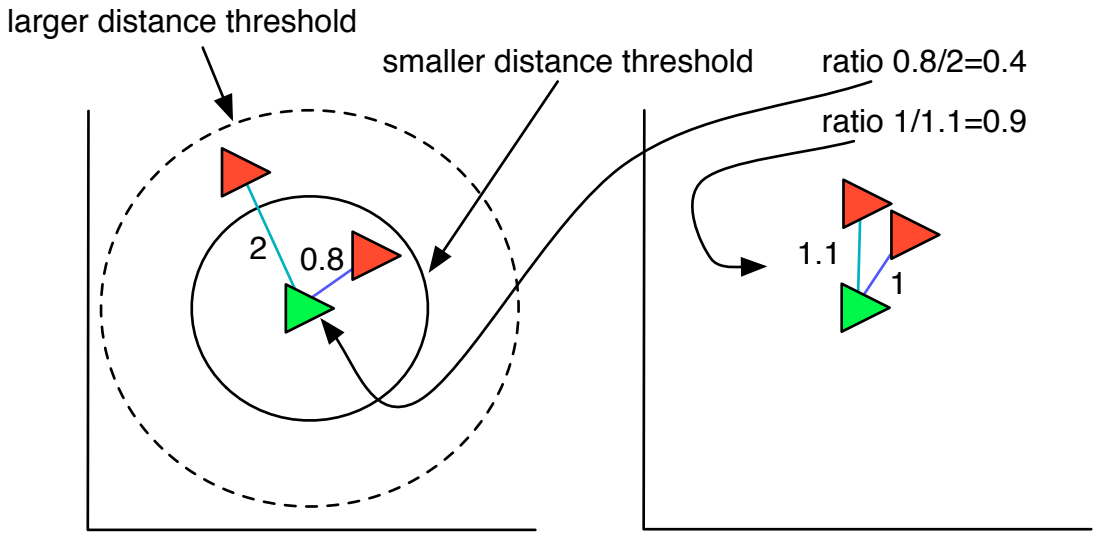


Figure 16 Matching Strategies. The green triangle is a feature vector is being matched to feature vectors from another image (red triangles). With the nearest neighbour strategy, only one match is returned, regardless of whether the smaller or larger distance threshold is used. However, if the similarity strategy is used, both red triangles are matched when the larger distance threshold is used. The final strategy, nearest neighbour distance ratio, the ratio of the 1st and the 2nd nearest neighbours are considered. The scenario in the right hand side demonstrates a case where, even though the nearest neighbour is close to the feature vector, it is rejected because of the near proximity of the 2nd nearest neighbour.

4.4 Outlier Removal and Transformation Estimation

The output of a feature-based matching system is a set of matching points. It is common for this output to include both correct matches (inliers) and false matches (outliers). Therefore, a typical subsequent step in this process is a method to remove the unwanted outliers. The most often used approach to this problem is Random Sample Conesus (RANSAC) [54]; a method that iteratively estimates the parameters of a model that describes the transformation between the images. The basic concept can be described as follows:

- A random set of matches is selected as inliers and a model is fitted to this points.
- Each remaining match is tested against this fitted model. If it fits well, it is included in the hypothetical set of inliers.
- If a sufficient number (this is a parameter to the algorithm) of matches are included in the hypothetical set of inliers, the model is re-estimated with all of the inliers. If an insufficient amount of hypothetical inliers is reached, the model is rejected.
- This process continues for a fixed amount of iterations. If a model is accepted for an iteration, then the new set of inliers are included in the current set of inliers and the model is re-estimated.

Thus, output is a set of proposed inliers in addition to a model estimating the transformation between inliers. Extensions of the classic RANSAC have also been proposed by Torr and Zisserman [55] in the form of MSAC and MLESAC. These extensions were shown to perform better than the classic method, especially in circumstances where the number of outliers greatly exceeds the number of inliers (low

precision). In our experiments, we have chosen to use the MSAC extension from the RANSAC Matlab Toolbox implementation, developed by Marco Zuliani [56].

Chapter 5 Proposed Evaluation Methodology

5.1 Objectives and Overview

The objective of this evaluation is to compare the performance of a set of detectors and descriptors under the conditions of historic repeat photography. The detection and description algorithms we have chosen to evaluate are as follows:

Detectors:

- Harris-Laplace
- Hessian-Laplace
- Difference-of-Gaussian (DoG)
- Maximally Stable Extremal Regions (MSER)

Descriptors:

- SIFT
- Shape Context

These methods have been chosen on the basis that they have previously shown good performance [22], [27], [33]. Our evaluation is carried out in the context of matching, where the detectors and descriptors are interconnected and combined with a nearest neighbour matching approach to form feature-based matching systems. That is, a total of 8 (4x2) methods are compared and evaluated. We analyze the detection and description methods in combination for two reasons: 1) we are interested in finding matches between image pairs, and neither detection nor description can perform this task independently, and 2) the performance criterion typically used to evaluate detectors is not applicable with historic repeat photographic pairs (see details in Section 5.2.2.3).

The evaluation of detectors and descriptors in combination with historic repeat photography is important because:

- Historic repeat photography encompasses important information in regards to temporal change.
- Change detection techniques have potential to aid in the analysis of these types of images, but require as a pre-processing step the registration of images.
- Registration can be done with feature-based matching systems.
- The initial registration of a repeat and historic image pair is currently performed manually.

In this chapter, we present our proposed evaluation methods (Section 5.2) and the experiments constructed based on these methods (Section 5.3).

5.2 Evaluation Methods

The evaluation methods outlined here assess the performance of feature-based matching systems (based on different detector and descriptors) on historic repeat photography. The output of a matching system, when applied to an image pair, is a set of proposed matching features (which may be correct or false). An overview of a matching system can be seen in Figure 17. We use two methods to assess the performance of a detector-descriptor at an image pair: 1) the analysis of precision and the number of correct matches, and 2) the accuracy of an estimated transformation. In the following paragraphs, we give a brief overview of the evaluation method associated with each of these types of criteria.

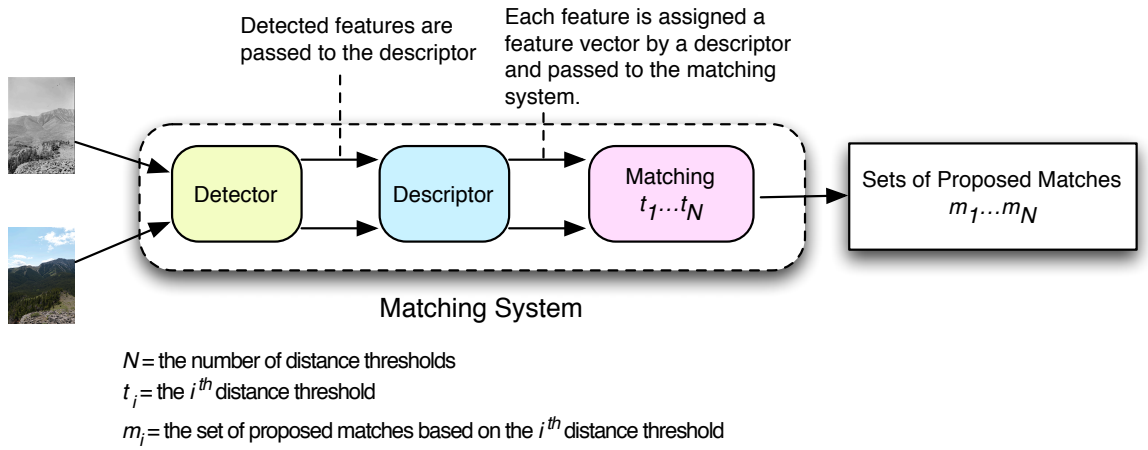


Figure 17 An overview of a feature-based matching system. A detector finds features, which are described by a descriptor, and matched using a matching strategy (nearest neighbour in this case).

Precision and correct number of matches. We assess the performance of a feature-based matching system by analyzing its ability to pass a certain minimum precision and number of correct matches requirements. For example, if the precision requirement is 30%, the number of correct matches requirement is 10 and the feature-based matching system on a specific image pair achieves 41% with 50 correct matches, then the matching system is considered to have passed the requirement. This same test can be applied across a set of image pairs, and the percentage of these image pairs that pass is called the pass rate (Section 5.2.3). In our evaluation, the detectors and descriptors that form the matching system are compared with the pass rate metric for a set of different requirements (which will differ based on the application) and different distance thresholds. A similar metric, called rank score (Section 5.2.4), also uses a minimum number of correct matches requirement, but considers the ranking of the detector-descriptors at each image pair (as opposed to the binary approach in the pass rate).

Before the pass rate and rank score can be calculated, precision and the number of correct matches must clearly be defined. The set of proposed matches can be divided into

correct matches and false matches. Precision is the percentage of correct matches in the set. A proposed match is correct if the features correspond (ground truth correspondence), or in other words, are located at approximately the same location and cover the same area (when the geometric relationship between the images is taken into account).

Estimate transformation accuracy. A repeat photograph is taken with a different camera system than the historic, which causes differences in scale, rotation, and translation to occur between images. A homography that describes this relationship is known as the ground truth transformation. This transformation can be estimated by using RANSAC (Section 4.4) in conjunction with the set of proposed matches. We consider a feature-based matching system to pass if the transformation error, defined by the difference in the scale, rotation, and translation parameters, is below a set of tolerance values. The percentage of image pairs that pass over the dataset is known as the pass rate with RANSAC. It is with this metric that we compare the detector-descriptor matching systems. Since the level of acceptable error tolerance may differ based on the needs of the application, a range of difference tolerances are used and a pass rate presented for each.

5.2.1 Ground Truth

In the overview two types of ground truth are noted: ground truth correspondences and ground truth transformations. The difference between the two is that one establishes whether two features from two images correspond, while the other defines the transformation required to align two images. The knowledge of the latter, in fact, is required to establish the former. In following subsections we define these terms for use in our evaluation method.

5.2.1.1 Ground Truth Transformations

In a similar manner to the Oxford dataset used in [22], [33][22], [33], the geometric relationship between the image pairs can be described by a homography. The creation of these homographies is independent from the evaluation method, and therefore the assumption is made that the homography is available when our evaluation method is applied.

5.2.1.2 Ground Truth Correspondences

As with other previous evaluations, we define our ground truth correspondences by specifying an error tolerance for both the location and overlap. We assume the repeat photograph has been taken in the same viewpoint, and therefore the homography that describes the geometric relationship between our image pairs is simply a 3x3 identity matrix. The error tolerances are constructed in a similar manner to [22], [23]. Consider that a local feature is represented by both its location, **x,** and its region, μ (either a circle or ellipse as defined in Chapter 4). Two features, ($\mathbf{x}_a$, μ_a) and ($\mathbf{x}'_b$, μ'_b), correspond if:

- Location error, ε_l, is less than a location error tolerance of τ_l pixels, where ε_l is the Euclidean distance between $\mathbf{x}_a$ and $\mathbf{x}'_b$,.
- The overlap error, ε_o, is less than an overlap error tolerance of τ_o percent. ε_o is defined as $\varepsilon_O = 1 - \frac{\mu_a \cap \mu'_b}{\mu_a \cup \mu'_b}$.

where $\mathbf{x}'_b = \mathbf{x}_b H$, $\mu'_b = \mu_b H$ and H is the homography between the image pair.

Feature region's area, μ, is computed by counting the pixels that lay within the region boundaries (which can be defined geometrically). An example of location and overlap error between two features is shown in Figure 18.

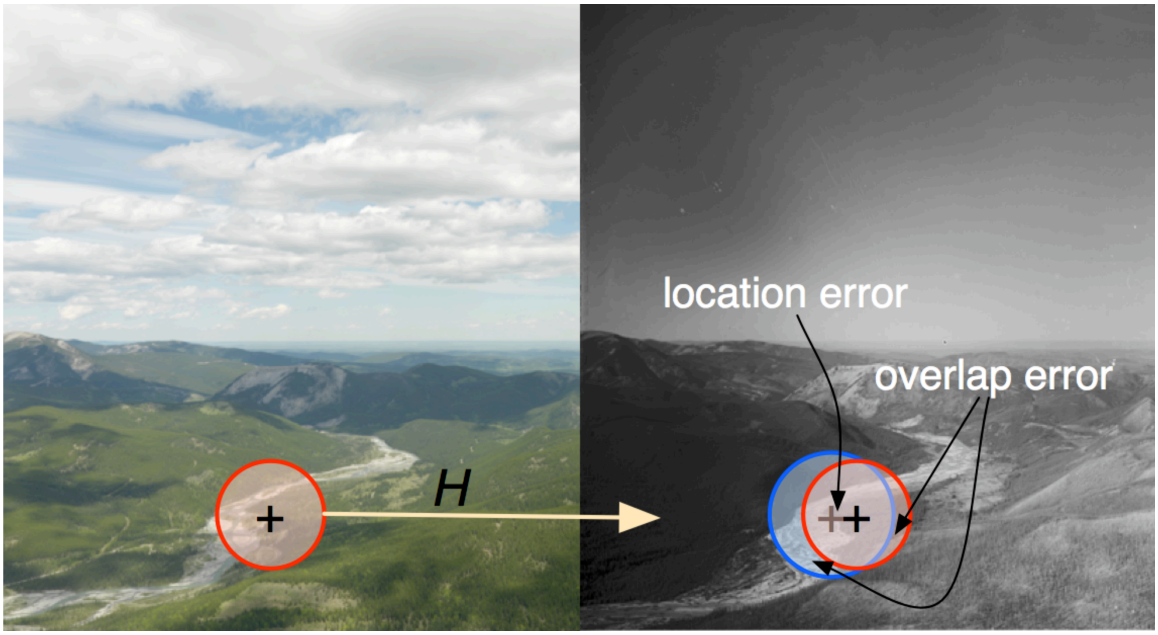


Figure 18 An example of location and overlap error. One feature is detected in both the repeat and historic image. In order to compute the location and overlap error between the features, one feature is projected onto the image canvas of the other base on the homography *H*. The distance between their centers is the location error, while the parts of their regions that do not overlap form the overlap error.

In the evaluation experiments conducted in this thesis, we have set the location error tolerance to 6 pixels and the overlap error tolerance to 50%. The selection method and rationalization for these tolerances is discussed in Section 5.3.4.

We note that these ground truth correspondences are used as a basis for labelling matches correct or false in the matching approach taken for our experiments. Specifically, if two features vectors are matched and they correspond, they are considered a correct match. Alternatively, if two features vectors are matched and they do not correspond, they are considered a false match.

5.2.2 Performance Criteria

5.2.2.1 Precision and Number of Correct Matches

A correct match occurs when features correspond and meet predefined matching criteria. The number of correct matches is an important metric because different applications require different quantities of corresponding points between images. For example, only two matching points are needed to resolve scale, rotation, and translation between a pair of images, while 8 points are needed to recover the fundamental matrix between 3D scenes taken at different viewpoints [21], while scene recognition may require hundreds to accurately represent a scene.

Precision is the number of correct matches relative to the number of total matches (both false and correct) determined by the matching system. After matching has occurred, correct matches need to be discerned from false matches. A common method used to remove false matches (the outliers) is RANSAC [54]. The success and efficiency of RANSAC based methods are closely related to the precision of the data set. At lower precision values, the probability of finding the correct set of inliers is reduced and the cost of finding them is increased. Therefore, precision, like the number of correct matches, is a suitable criterion to measure performance of feature detectors and descriptors.

Ideally, both precision and the number of correct matches are high. However, high performance in one of these metrics usually comes at the cost of the other. At lower distance thresholds there is higher precision, but less correct matches, while at higher distance thresholds there is lower precision (more false matches), but more correct matches.

A feature-based matching system produces a set of precision and number of correct matches, one for each distance threshold used (Figure 19).

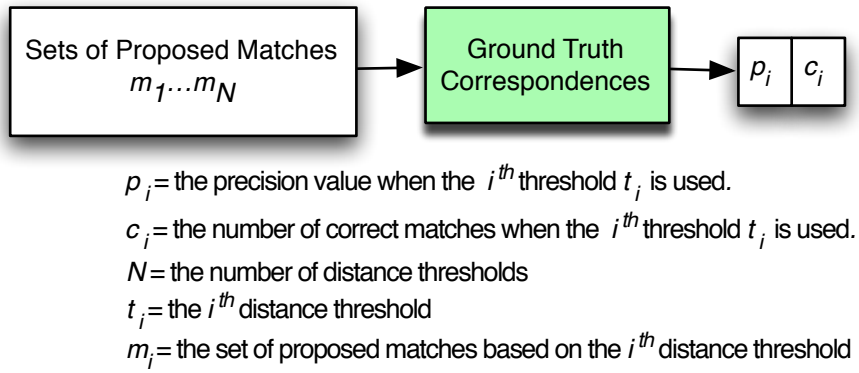


Figure 19 The precision and number of correct matches metrics are formed by combining the proposed matches with our ground truth.

5.2.2.2 Estimate Transformation Error

We also use the error between our ground truth transformation and an estimate transformation as a measure of performance. From a practical standpoint, these criteria determine how well a feature-based matching system can perform the alignment of an image pair in comparison to the alignment achieved by a human. The only requirement to produce the estimate transformation is 2 correctly matched features between images. Whether the RANSAC approach generates the correct transformation may be influenced by the precision of the match set, but is not explicitly bound by any specific precision limit.

The estimate and ground truth transformations can be described as a change in scale, rotation, and translation. The error between the estimate and ground truth transformations can then be described as the difference between these values. Formally this is defined as

$$\varepsilon_s = \left| s_{est} - s_{gt} \right|$$

$$\varepsilon_r = \left| r_{est} - r_{gt} \right| \tag{5.1}$$

$$\varepsilon_t = \left| t_{x,est} - t_{x,gt} \right| + \left| t_{y,est} - t_{y,gt} \right|$$

where s, r, t_x , and t_y represent the scale, rotation, and translation (along each axis) for the estimate (*est*) and ground truth (*gt*) transformations.

Since the estimate transformation is based on the proposed matches, a set of transformation errors exist, one for each distance threshold.

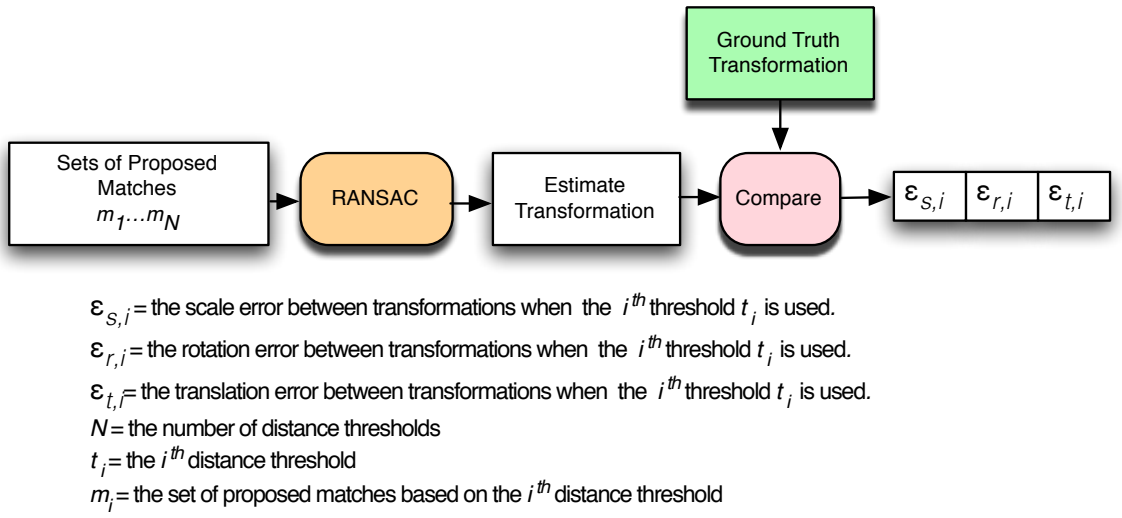


Figure 20 The transformation error of a RANSAC estimate transformation and our ground truth transformation is the difference between their respective scale, rotation, and translation changes.

5.2.2.3 Discussion

Two other popular criterions, repeatability and recall (equation 3.4), are difficult to apply to photographs that incur change.

The criterion of repeatability is desirable because it can be used independent of application and description. However, the concept of repeatability is not transferable to scenes that incur physical change. Consider an example of a chalkboard with two dots on

it. These dots could be considered features. When photos of this chalkboard are taken from different angles, and the dots are consistently detected as features, the detector is said to have good repeatability. Now consider erasing one of the dots and the detector no longer detects erased feature. Unless the change is known *a priori*, the repeatability score for the method will unjustly decrease, since the absence of detector was due to the scene changing, not the inability of detector. With historic repeat photographs, there are cases where change is expected to occur in the scene, and therefore such a situation exists in our dataset.

Using recall faces a similar difficultly. The recall value is the ratio of correct matches to potential matches. In [33] a potential match is defined by correspondence; when two features share the same point or area in a scene. However, when physical change occurs in the scene, two features may correspond, but not cover the same object. That is, they could correspond by accident and should not be considered a potential match. The recall criterion is therefore not a good candidate in our context. An example of an accidental correspondence can be seen in Figure 21.

Figure 21 An example of two features that correspond, but for different reasons. We call these "accidental correspondences". The Hessian-Laplace feature on the left is oriented on forest clear-cut, while the feature on the right is oriented on forest cover underneath cloud cover.

A further problem with the recall criterion is that it is not well suited for evaluating detectors. Recall and precision are typically used for evaluating pattern recognition algorithms where the dataset is shared among all the methods being experimented. When feature detectors are being compared in the context of matching for a specific pair of images, the dataset is the set of features and their correspondences. Here in lies the problem: the dataset is defined by the feature detector and thus it will differ from method to method. The recall value, then, does not prove to be a very informative measure of comparison. This is easily shown with an example. Consider feature detectors **A** and **B**. When detector **A** is used on the image pair, there are 10 corresponding features. When matching has been applied 20 matches are returned, and 10 of these matches are correct matches. Method **A** has achieved 50% precision and 100% recall. When detector **B** is used on the image pair, there are 100 corresponding features. When matching is applied, 80 matches are returned, and 40 of these are correct matches. Method **B** has achieved 50% precision and 40% recall. Method **A** has a much higher recall value, however, only found 10 correct matches compared to the 40 by method **B**. Clearly detector **B** has better performance. We should note that the Mikolajczyk and Schmid [33] study that used precision vs. recall was a comparison of descriptors where the same features were used with all methods (with some exception). In their case, recall is a valid criterion.

5.2.3 Pass Rate

The *pass rate* is defined as the percentage of image pairs out of the entire dataset that fall within the tolerances imposed on the precision (p_{tol}) and number of correct matches (c_{tol}). The pass rate is computed with the process shown in Figure 22. The process unfolds as follows. A detector-descriptor is applied to the image pair in the dataset,

resulting in a set of feature vectors that are processed by the matching module. Matching is performed using the nearest-neighbour approach with a distance threshold. The distance threshold applies to the feature space and specifies the tolerance for matching; that is, if the nearest neighbour falls outside the distance threshold, then no match is found. A set of N different distance thresholds $\{t_1, t_{2, \ldots,} t_N\}$ is applied. For each image pair, the precision and number of correct matches are computed for a specific distance threshold, and then compared to the tolerance values p_{tol} and c_{tol}; the image is passed or failed. The pass/fail decisions are aggregated into the pass rate over the entire dataset.

As previously mentioned, different amounts of correct matches per image pair are necessary for different applications. Therefore, the pass rate is computed for 6 different tolerances for the number of correct matches: c_{tol}= 2, 10, 25, 50, 100, and 200.

Different precision values are also required for different applications. Therefore, the pass rate is computed at 6 precision tolerances: p_{tol}= 50%, 40%, 30%, 20%, 10%, and strictly greater than 0%. The quantitative evaluation data is presented as a function of the tolerances for the distance threshold, precision and number of correct matches.

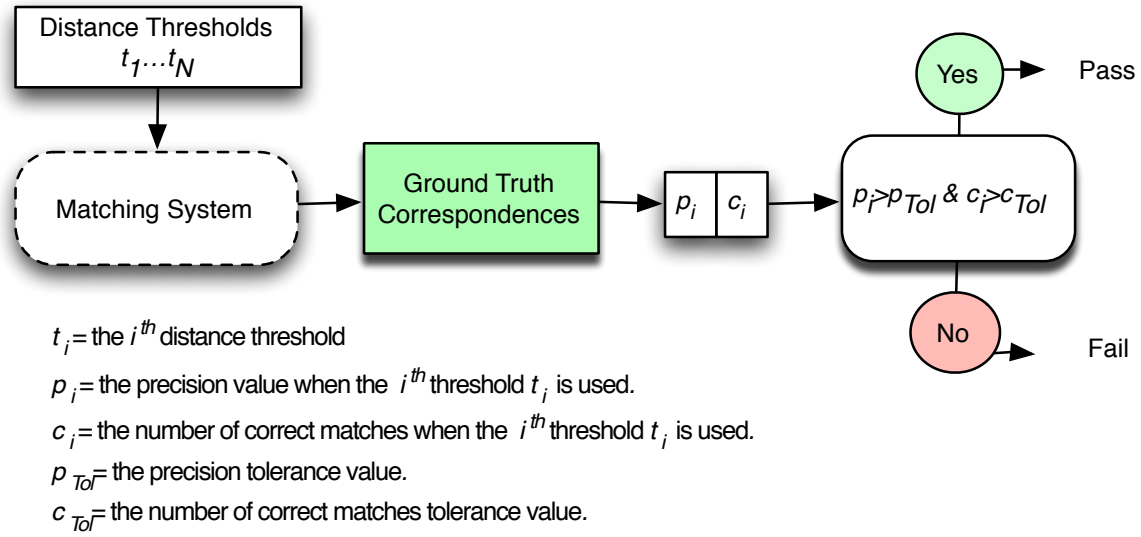


Figure 22 An overview of the computation of the passing criteria for a specific image pair. This process is repeated for the entire dataset to compute the final pass rate.

An important consideration when determining the pass rate at distance thresholds is that each descriptor uses a semantically different distance measure that is unique to its construction and dimensionality. Specifically, our experiments are conducted on two different descriptors, SIFT and Shape Context, each of which produce feature vectors with different dimensions (128 for SIFT, and 36 for Shape Context), and thus the relative distances between features in feature space are different. One way to deal with this inherent difference is to threshold not on distance threshold, but rather number of total matches (the correctness of which are unknown) returned. That is, the nearest-neighbours in feature space can be ranked from closest to furthest and threshold is applied to the ranking. For example, a total match threshold of 100 would return the closet 100 nearest-neighbours in feature space. This concept can be applied to a descriptor independent of dimensionality, and was used in Mikolajczyk and Schmid's [33] evaluation of descriptors. The problem is that such an approach does not adapt well when applied to many image pairs over a large dataset (as opposed to single image pairs like Mikolajczyk

and Schmid). The reason is that the total number of matches returned is dependent on the scene, and thus the same total matches threshold applied to two different image pairs may be associated with very different distance thresholds. For example, consider that a total match threshold of 100 is associated with a distance threshold of 200 for an image pair A, but when applied to an image pair B it is associated with a distance threshold of 400. The fact image pair was much more difficult to match is lost in the total match threshold technique. We take a different approach and adjust one distance threshold relative to the other such that each returns approximately the same amount of total matches.

The concept can be achieved in the following manner. Consider an evaluation has a set of descriptors $\{d_1, d_2, d_3, \ldots, d_n\}$ under experiment. One descriptor, d_b, is chosen as the base descriptor. Distance thresholds for the base descriptor d_b are called base thresholds t_b. We can then define a relative threshold, $t_{i,r}$, for each of the remainder descriptors d_i such that $totalMatches(d_b,t_b) \approx totalMatches(d_i,t_{i,r})$, where *totalMatches* is a function that produces the number of total matches when the matching system is constructed with descriptor d and a distance threshold t. We note that when this method is applied to a single image pair, a similar result to the one obtained by Mikolajczyk and Schmid is achieved. In this case though, the base threshold would need to be set such that 100 matches are returned (as per the previous example).

5.2.4 Rank Score

The pass rate gives a good idea of what detector-descriptor performs best over the entire dataset in different contexts, but ignores the rank of these methods at individual image pairs because the pass criteria is binary. Specifically, two detector-descriptors might always pass the criteria on the same image pairs, but one might consistently

outperform the other. Therefore, while the pass rate gives a good indication of performance at a specific precision and number of correct matches requirements, it may fail to capture one algorithms comparative advantage over the other. To overcome this problem, we introduce a simple scoring system based on rank.

We rank detectors by precision. In a similar manner to the pass rate, we examine precision with respect to a required number of correct matches and a set of distance thresholds. If a detector-descriptor does not meet the correct match criteria, it is given a rank score of 0. Otherwise, the detector-descriptor with the lowest precision value receives a rank score of 1, and the next lowest a rank score of 2. This process is repeated until all detector-descriptors have received a score. These scores are accumulated over the dataset, and then divided by the number of image pairs in the dataset to produce the final score. In the event of a tie, both methods receive the same value, and the process continues as discussed above. The final rank score of each detector-descriptor is compared to one another to determine its relative performance.

5.2.5 Pass Rate with RANSAC

Pass rate with RANSAC defines the notion of a pass based on the transformation error meeting a transformation error tolerance. Since the acceptable amount of error is undefined, we choose to examine this pass rate as a function of different error tolerance values. Specifically, 3 graphs are produced. Each graph corresponds to a change in one tolerance, with the other two tolerance values being fixed. The fixed tolerance values are 0.01 for scale (1% difference is scale factor), 1 degree for rotation, and 5 pixels for translation. The set of scale tolerances s_{tol} range from 0.001 to 0.015 (with increments of

0.001). The set of rotation tolerances r_{tol} are 0.2 to 2 degrees (with increments of 0.1). The set of translation tolerances t_{tol} ranges from 1 to 15 pixels.

5.3 Experimental Design

The evaluation methods discussed in the previous section have been applied to two sets of 73 image pairs from the MLP database. The first set consists of registered image pairs (i.e. images are already aligned). The second set consists of raw image pairs, which differ in scale, rotation, and translation. The specifics of these experiments are given in Section 5.3.1. The details of the dataset are further elaborated in Section 5.3.2. In Section 5.3.3 we explain how our ground truth transformations were formed. The selection of the location and overlap error tolerance values that defines our ground truth correspondences is described in Section 5.3.4. Finally, we discuss how the relative distance threshold value was set in our experiments (Section 5.3.5).

5.3.1 Experiments

5.3.1.1 Registered Image Pairs

The focus of the experiments conducted on the registered image pairs is theoretical, in that we make no assumption in terms of the number correct matches and precision that might be required for an application. The use of registered image pairs (as opposed the raw image pairs) is beneficial since the focus of the results is on the radiometric, acquisition sensor, and landscape changes between the images rather than geometric changes (e.g. scale) which have been studied in other evaluations [22], [27], [28], [37].

The pass rate and rank score are used to evaluate these image pairs. Five distance thresholds are considered: t_i=150, 200, 250, 300, and 350. These values were considered in accordance with Ke and Sukthankar [34]; they found that SIFT operated optimally (in

the context of image retrieval) with a distance threshold of 141, but suggested good performance may be obtained with a variety of thresholds.

5.3.1.2 Raw Image Pairs

The focus of the experiments conducted on the raw image pairs is of practical nature, in that we seek to address an actual problem within the MLP organization. The trade-off is in generality. Specifically, instead of showing the pass rate at numerous distance thresholds, we examine it at only one. Furthermore, we use the pass rate with RANSAC experiment discussed Section 5.2.5, the results of which can only extend to applications where the geometric transformation is restricted to scale, rotation, and translation. Finally, in addition to the raw image pairs, these experiments are applied to the match sets that have been filtered using the geometric filtering technique that uses *a priori* information about the MLP dataset. This filtering technique is discussed in next section. The distance threshold used in these experiments is 250, a value that provided a relatively good trade off between precision, the number of correct matches, and overall pass rate.

5.3.1.3 Geometric Filtering

The main geometric difference between raw image pairs is scale. The scale change between an image pair is a function of the field of view for each image, which is directly related to the camera system (lens and sensor size) used to capture each image. Therefore, the scale change should be constant among repeat photographic pairs where the same repeat camera and historic camera were used to capture their respective photos. In the case of the repeat photographs in the MLP dataset, the majority of these images have been taken with the Hasseblad HD3 system accompanied by a 35mm lens. The camera systems used for the historic images may vary, but it is assumed that the same camera

was used for a specific survey and year. In other words, the scale change between image pairs is known for a specific survey. We can use this information to remove false matches. Specifically, if two features (from two images A and B) at locations x_a and x_b match they must fulfil the additional spatial constraint

$$x'_a = x_a \begin{bmatrix} s & 0 \\ 0 & s \end{bmatrix}$$
$$dist(x'_a, x_b) < \varepsilon \tag{5.3}$$

where s is the scale change between an image pair with a specific camera system, *dist* is the Euclidean distance function, and x_a' is the location of x_a transformed to image B. If scale change was the only difference between the image pairs, then the images could be aligned simply by applying the scale factor to one of the images (in this case, the *dist* function would be 0 in the above example). However, images also differ by rotation and translation. Hence the need to allow for a certain amount of error ε in the distance between locations. In our experiments, we have set this error tolerance to 50 pixels (where the resolution of the images is approximately 800x600). This tolerance value was chosen based on the fact that it could handle large translation differences between image centers and still increase the precision of the proposed match set.

This filter requires a parameter that is an estimate of the scale change between an image pair. This parameter is calculated based on the average scale change of image pairs belonging to the same survey. The only exception was for the Bridgland 1913 survey (See Table 7, Appendix A for details of image pairs from this survey), where the average scale change differed between stations. The origin of this difference is unknown, but may be due to a change in the camera used to capture the historic survey. In this circumstance,

the scale estimate was based on the average scale change for a specific station, where the same camera was guaranteed to be used between image pairs.

5.3.2 Dataset

While several thousand historic images have been digitized and repeated, our evaluation is applied to only a sample of this database. There were 5 determining factors that decided if a photographic pair would be included in our sample:

1. The images must be part of a full image set. An image set can be defined as 4 images: the registered pair, and the raw pair. The expectation was that these images were located in their respective station directories: Original Masters, Repeat Masters, Original Scans, and Repeat Tiffs.
2. Each image in the image set respected its expected resolution. The expected resolutions were approximately 7216x5412 and 11800x8600 for the repeat raw and historic raw images, and 5500x3800 for the registered images.
3. Each image in the image set was stored in the Tiff format (with lossless compression).
4. The repeat image must not have been located at a fully occluded viewpoint. In the practice of repeat photography, historic scenes are retaken from the exact same position even if trees or brush block the scene depicted in the historic photograph. In these situations, detection and description will fail.
5. The repeat error was not significant. Repeat error is the error in determining the location and viewpoint of the historic photo. In these cases, large portions of the scene cannot be aligned properly, and therefore could not provide reliably ground truths.

Organizational inconsistencies within the MLP dataset significantly reduce the number of potential image pairs for evaluation. In total, 73 image pairs, from 24 stations of 6 different surveys that were dated from 1896 to 1927, met these requirements. The year and locations of the photographs cover a time span and geographical area that is representative of the MLP dataset as a whole. In comparing the size of our dataset with others, Gil *et al.* [37] in their application specific (vSLAM) evaluation of detectors and descriptors used only 4 unique scenes. Some examples of our dataset can be seen in Figure 23. All 73 image pairs are presented in Appendix A.

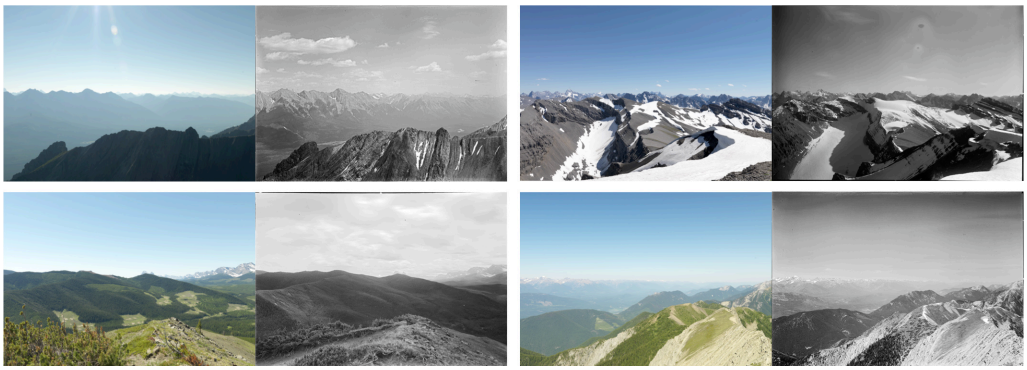


Figure 23 Examples of historic repeat photography image pairs from our dataset.

These images were downsized versions of their full resolution counterparts so that the evaluation could be executed in a manageable timeframe. Specifically, the repeat raw images were 794x596, the historic raw images were approximately 700x500, and the registered images were approximately 600x400. These resolutions are consistent with images used in other evaluations[22], [33], [37]. It should be noted that historic raw images at full resolution are much larger than their corresponding full resolution repeat raw images, while this property does not hold for their downsized counterparts. The large resolution of the historic raw files is related to the resolution at which they are scanned. The repeat raw images are taken with a larger angular field of view than the historic

image, which guarantees that historic scene fits with the repeated photograph. With this knowledge, the historic raw image could be downsized such that its longest edge is equal to the longest edge of the repeat raw image. The decreased file resolution allows for faster computations and also decreases the difference in scale between the images. After this adjustment, these images differed by an average scale factor of 0.784 (i.e. the historic image still needs to be reduced by this factor to match the scale of the repeat) with standard deviation of 0.049, an average rotation of 0.0068 degrees with standard deviation of 0.3242, a translation along the x-axis of 6.8 pixels with standard deviation 6.532 and a translation along the y-axis of 12.9 pixels with standard deviation 4.38. The exact geometric differences between each raw image pair can be seen in Appendix A.

5.3.3 Creation of Ground Truth Transformations

Our experiments comprise of 2 sets of image pairs. In this section we describe how the ground truth transformations for these sets were defined.

The registered image pairs were manually aligned by MLP lab assistants using the scale, rotation, and translation transforms in Photoshop CS4. The homography that describes their geometric relationships is simply the identity matrix.

Generating the homography for the raw repeat and historic images is slightly more complex. We chose to do so systematically, leveraging existing automatic control point generation software that performs well under the simple transformations that exist between the registered and raw versions of the images. Specifically, the raw historic and registered historic images can be described exactly by a scale, rotation, and translation, where the scale is the principle transformation. The same idea can be applied between the raw repeat and the registered repeat images, though the scale does not change between

these images (the registered is a cropped version of the raw). By using an automatic control point generator (we use panomatic, which is based on SURF+RANSAC [30], [54]) we can find reliable matching points between the raw and registered versions of the images. As long as we have at least 2 pairs of matching points, we can resolve a transformation and thus create a homography, H_{raw2reg}, for both historic and repeat images. We can then transitively form the homography of the raw images:

$$H_{raw2raw} = H_{Repeat\_raw2reg} * H_I * inv(H_{Historic\_raw2reg}) \quad (5.4)$$

where H denotes a homography and the subscripts denote the relationship being described by the homography. The identity matrix H_I is included for the sake of completeness; it is the homography between the registered images, and thus acts as a bridge between the raw2reg transformations. Figure 24 visually depicts this relationship.

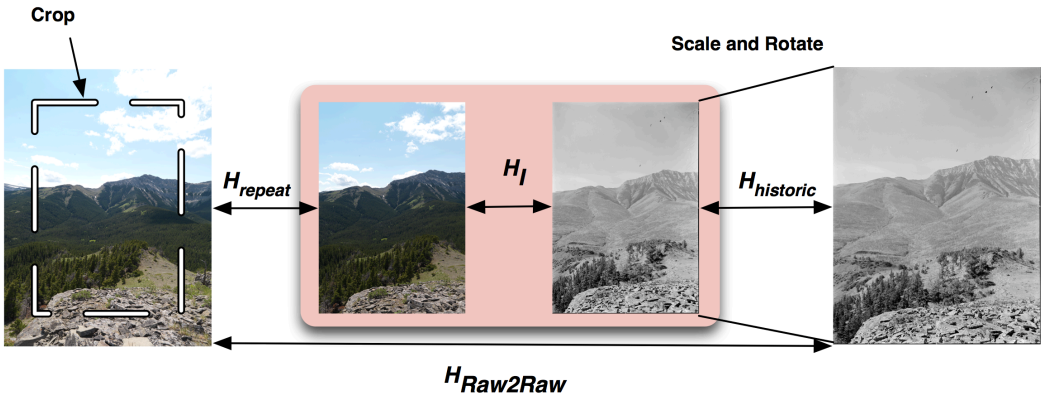


Figure 24 The relationship between the registered and raw image pairs. The homography between the raw image pairs in calculated based on the intermediate transformations $H_{historic}$, H_I, and H_{repeat}.

5.3.4 Location and Overlap Error Tolerance Selection

The ground truth correspondence defined in our evaluation method requires location and overlap error tolerance values. The tolerance values used in previous work have been 40% for overlap tolerance in the context of evaluating detectors [22], [23], 50% in

the context of evaluating descriptors [33], and 1.5 pixels [20], [23] for location tolerance. In this section, we discuss why these tolerances were chosen, how we arrived at tolerances that were appropriate for our dataset.

Mikolajczyk and Schmid [23] chose a 40% overlap error tolerance based on the scale space sampling of the Harris-Laplace and Harris-Affine detectors (the experiment was centered around these detectors). The overlap error tolerance was increased to 50% when descriptors were also involved in the evaluation based on observations that two features could still be matched with this amount of error. This property still holds with our dataset, and therefore, we adopt this tolerance value as well.

The location error tolerance of 1.5 pixels used by Mikolajczyk and Schmid [20], [23] was chosen based on evidence that, at this tolerance, there was a low probability that two points would be accidentally within the error tolerance (i.e. the error between points was due to detector accuracy error rather than detecting different points close by). With the repeat photographic pairs in our dataset, a homography cannot exactly represent the geometric relationship due to potential repeat error and distortion. Repeat error occurs when differences occur between the camera position of the historic and repeat image. Distortion is caused by the lens used to capture an image, therefore, unless the same lens is used in both the repeat and historic photograph, or the distortion is fully rectified in both, there will be differences in point locations in some areas of the image pairs. An example of repeat error/distortion (it is difficult to discern which of these properties is the main contributor) can be seen in Figure 25. Three separate areas of the image pair are highlighted to demonstrate the effect of repeat error/distortion. Notice that the error can

be different for different parts of the image. To accommodate for repeat error/distortion, a less restrictive location tolerance is used.

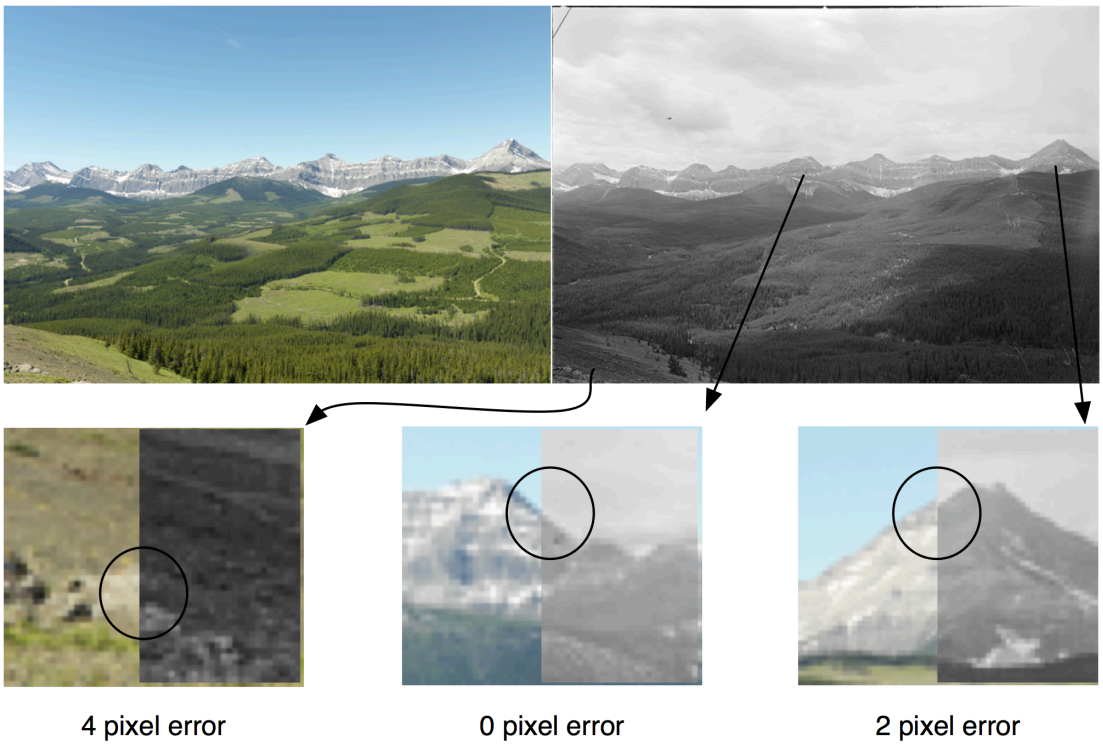


Figure 25 Repeat Error / Distortion. Certain areas within a repeat pair may align perfectly, while others may not.

To determine the location error tolerance, we use a similar method to the one used by Mikolajczyk and Schmid [33] for defining overlap tolerance. We examine the number of correct matches that are identified at intervals of location tolerance values 1, 2, 4, 6, 8, 10, while fixing the overlap tolerance to 20%. All detectors, paired with the SIFT descriptor, are used with nearest neighbour matching and a relatively low distance threshold of 150. By setting the distance threshold and overlap tolerance low, we ensure reliable matches and can assess the location error between these matches. We restrict this experiment to only the SIFT descriptor because, in this context, we are analyzing the location error of features, which does not change based on the descriptor used. Figure 26

shows the delta number of correct matches and corresponding features over the entire dataset (registered image pairs) that occur at location tolerances intervals. Note that the data point on a curve located at a location error tolerance value is related to the number of correct matches that occur between that value and its lower adjacent tolerance value (i.e. the change in number of correct matches that occurs from one tolerance to the other). For example, the number of correct matches at the interval denoted by 4 on the x-axis is 55 (Hessian-Laplace), and is associated with the number of correct matches that occur between a location error tolerance between 2 and 4 pixels. For the top graph in Figure 26, most correct matches occur very close to each other, with a location error between 0 and 1 pixel, however, we see that correct matches can still be found up until a location tolerance of 4 pixels. The bottom graph in Figure 26 shows the total number of corresponding features with respect to location tolerance. We see that correspondences can still occur up until a 6 pixel tolerance value. Given that these correspondences may be matched with a higher distance threshold, we have chosen to set the location tolerance to 6 pixels for our experiments. This experiment was further supplemented by manually examining the distance between samples of point locations between registered images pair.

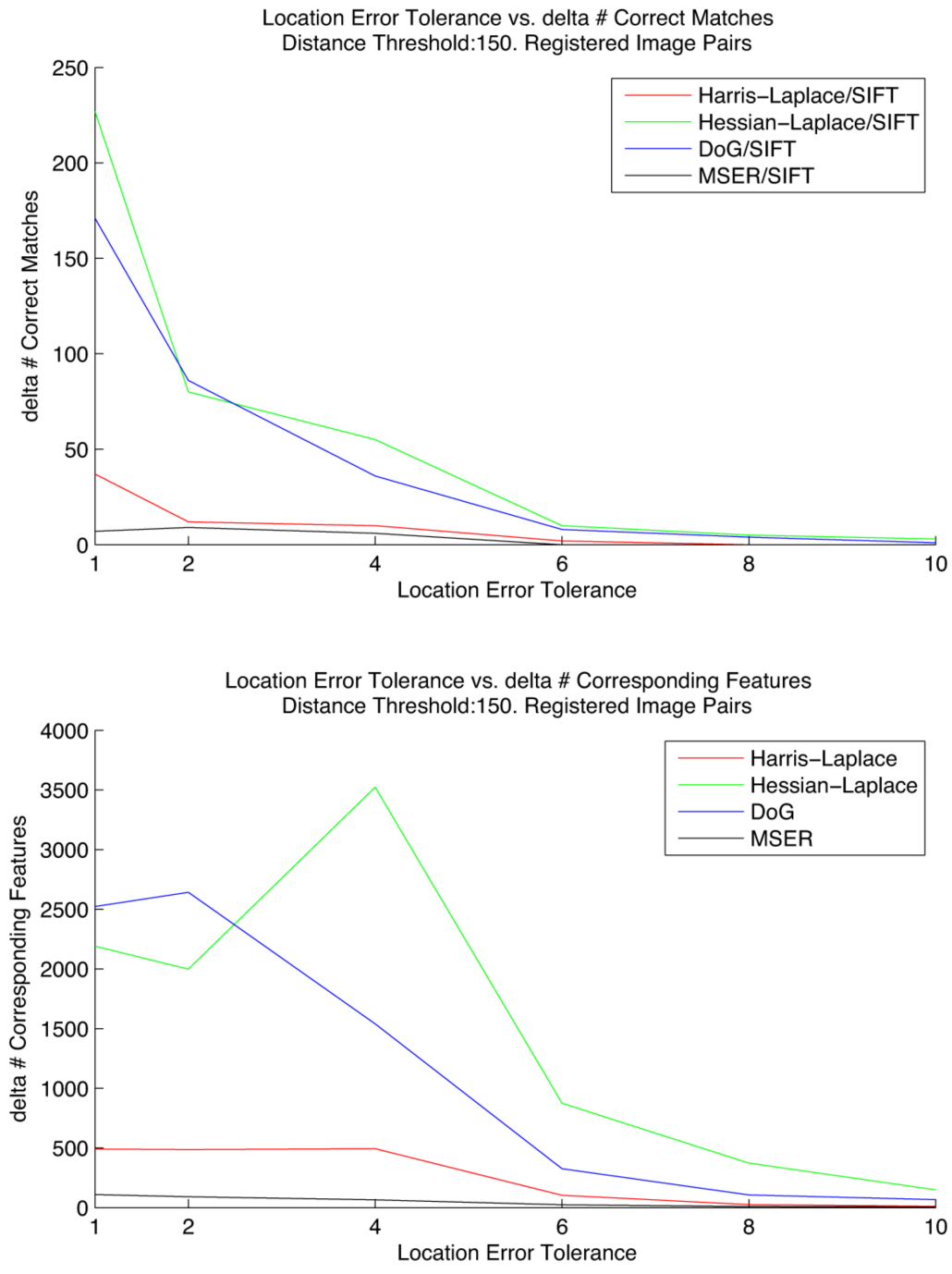


Figure 26 The plots show the change in (delta) number of correct matches and corresponding features over the dataset as the location error tolerances is increased. The graphs show the number of correct matches/corresponding feature that occur at intervals of an increased location error threshold, demonstrating that very few strong (at low threshold) correct matches and correspondences occur above a 6 pixel threshold.

5.3.5 Relative Threshold Selection

In our evaluation method we outlined a method to adjust the distance thresholds of descriptors relative to a base distance threshold threshold (Section 5.2.3) in order to properly compare their results. In our experiments, we use SIFT as the base descriptor and determine the relative threshold for Shape Context. Through observation, we noticed that the relative threshold of Shape Context is approximately a 50 distance unit decrease of a SIFT base threshold. This behaviour is exhibited in Figure 27. Each graph is associated with one base threshold t_b, while the x-axis of each graph is associated with a candidate relative threshold, which is an offset {-100, -50, 0, 50, 100} of the base threshold. The data point on the curve is the average difference between the base threshold and the candidate relative threshold over the dataset. Formally this can be described by

$$\frac{1}{M}\sum_{1}^{M}\left|totalMatches_m(SIFT,t_b) - totalMatches_m(SC,t_b + offset)\right| \tag{5.2}$$

where the subscript *m* denotes the total matches function at a specific image pair and *M* is the number of pairs in the dataset. The candidate relative threshold with the lowest average difference is chosen as the relative threshold since it is at this threshold that the total number of matches returned by the descriptors is the most similar.

Examining Figure 27 again, we clearly see that the best relative threshold is achieved at an offset of -50 for all the base thresholds sampled. This is especially so at higher base thresholds. We should note that while the method we used to determine relative thresholds works for Shape Context and SIFT, it may not apply to other descriptors where the relationship of features space may not be as linear as it is here. The concept of base and relative threshold, however, is applicable to other descriptors.

For a matter of the convenience, we only state the SIFT distance threshold (the base threshold) throughout this thesis.

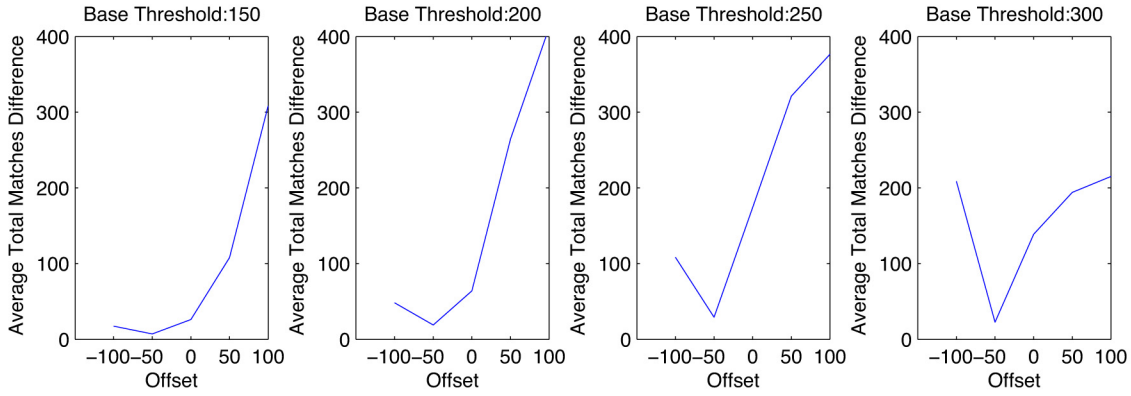


Figure 27 Relative Thresholds of Shape Context. At each SIFT base threshold, the relative threshold for the Shape Context descriptor is computed by finding the threshold with the average number of total matches that is most similar to base descriptor and threshold. In this case, the relative threshold is achieved at a -50 offset to the base threshold.

Chapter 6 Experimental Results

6.1 Registered Image Pairs

This section gives the results for all the experiments based on the registered image pairs in our dataset. The results are given for two experiments based on the pass rate and rank score discussed in Chapter 5. These experiments serve two purposes. The first allows the reader to easily identify the best performing detector-descriptor for different application requirements for precision and quantity of correct matches. The second examines the feasibility of using a feature-based matching system to find matching points between historic repeat photography image pairs.

6.1.1 Pass Rate

The pass rate results are presented in Figure 28. The characteristics of these plots can be explained as follows. Ideally, one aims at obtaining feature matching with high precision and a high number of correct matches. However, there is a trade-off between precision and correct number of matches. At low distance thresholds, the precision is expected to be better, since there is a stricter requirement on the similarity measure in the feature space. This stricter requirement also restricts the number of possible matches that are returned. Therefore, when moving the distance threshold higher, it is expected that more correct matches are found at the cost of lower precision.

The trade-off between precision and correct match quantity introduces some distinct trends in the pass rate results. If one examines a specific detector-descriptor curve over a change in distance threshold the slope increases. This phenomenon occurs because the image pairs that previously passed at higher precision tolerances for lower distance thresholds cannot sustain that precision with at higher distance thresholds, which drops

the right hand side of the curve. Furthermore, as the distance threshold increases, so does the pass rate at the lowest precision requirement (greater than 0%), because more image pairs obtain at least 2 correct matches, when at lower distance thresholds they could not. The increase in distance threshold also causes more correct matches, which causes an increase in pass rate at higher number of correct matches tolerance.

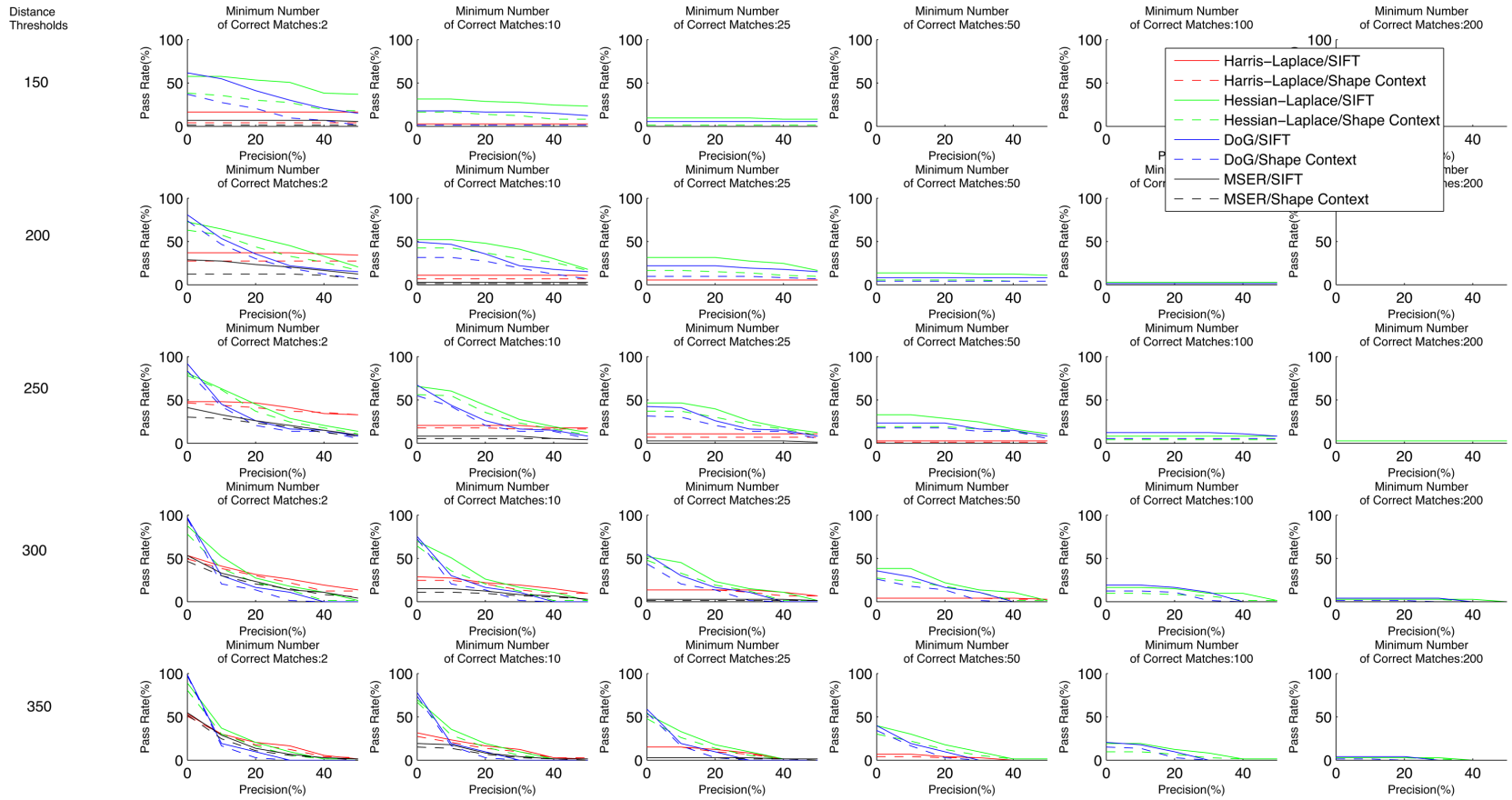


Figure 28 Pass Rate Results. A graph represents the pass rate for each detector-descriptor based on precision and correct number of matches requirements. Each row is associated with a different threshold used to produce the pass rate results. Each column is associated with a requirement number of correct matches. The x-axis of each graph is associated with the precision requirement

6.1.1.1 Discussion of Pass Rate Results.

Descriptor Pass Rate Performance. When comparing the pass rate results of the descriptors evaluated, the SIFT descriptor always outperformed Shape Context when paired with the same detector. As the distance threshold increased, the difference in pass rate performance between these descriptors also decreased, though SIFT continued to perform marginally better in all cases. The main contributing factor to overall pass rate performance was the choice of the feature detector.

Detector Pass Rate Performance. The overall best performing detector is the Hessian-Laplace followed by the DoG. At the lowest precision requirement, the DoG does show better performance; the difference being that DoG finds more image pairs with at least some correct matches than Hessian-Laplace, but with worse precision. These detectors are able to detect corresponding features for every image pair in the dataset, and 2 or more correct matches for a large majority (90-99%) of these image pairs. In the most difficult requirements, where precision of 50% and 200 correct matches are required with only 3% of the image pairs pass.

The Harris-Laplace and MSER have a subset of image pairs for which the precision requirement is stable (i.e. less significant slope in the graphs), but suffer from the fact that their features cannot be matched for a large portion of the dataset and only with a low number of correct matches. This is partially due to the absence of any corresponding features being detected between the image pairs. In the case of the Harris-Laplace 10% of the images pair had no corresponding features, while with MSER 6.4% of the image pairs had no correspondence.

6.1.2 Rank Score

The rank score experiment demonstrates an alternative method to examining the precision and number of correct matches criteria over a dataset. Instead of focusing on the ability of an image pair to pass two requirements, one requirement is fixed, while the other is used to rank the detector-descriptor performance. The ranking of the detector-descriptors at each image pair in the dataset is aggregated into the final rank score. Our rank score results are shown in Figure 29, where graphs show the rank score of precision at different minimum number of correct match requirements.

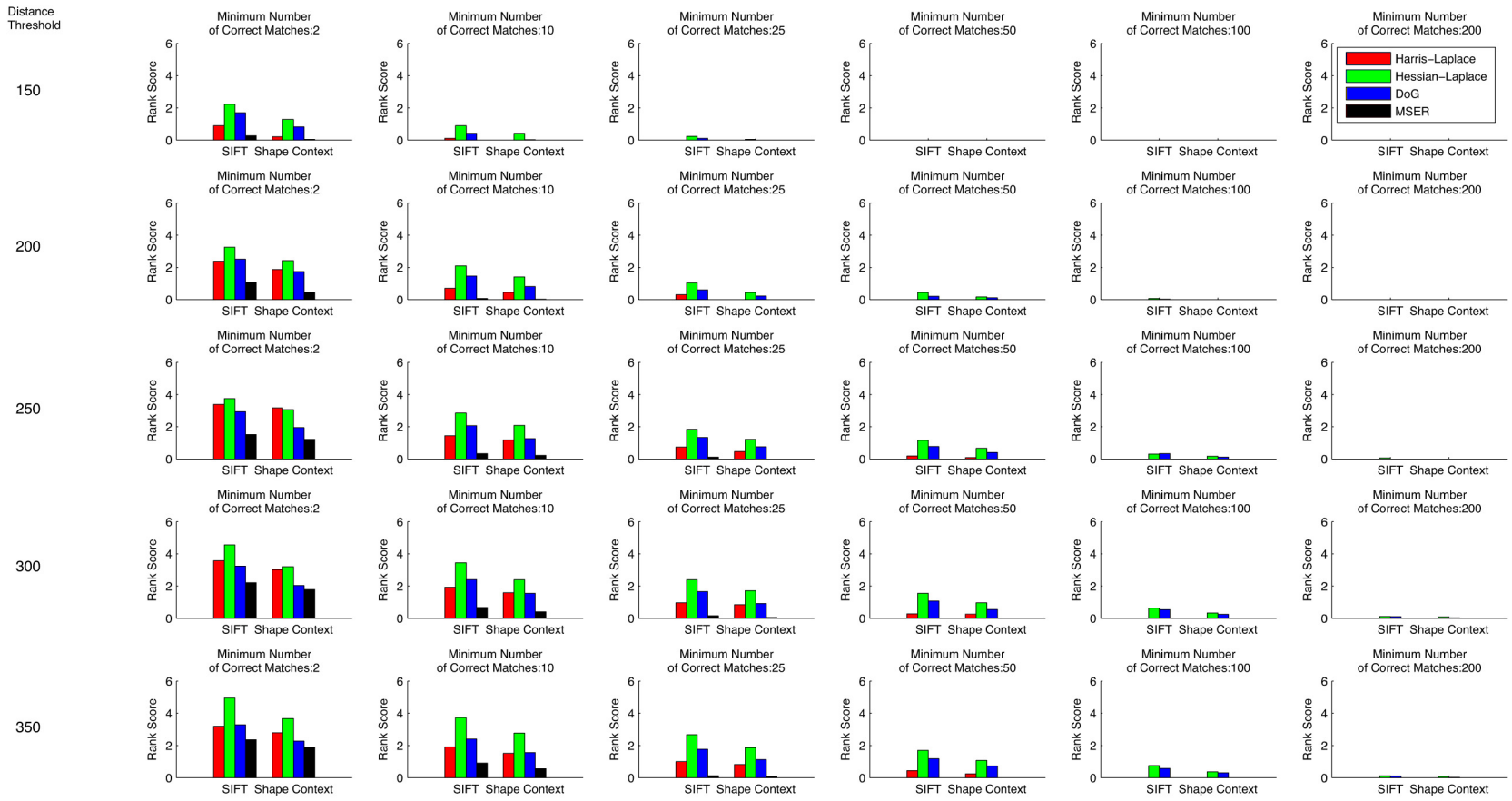


Figure 29 Rank Score Results. Each graph shows the rank score for the detector-descriptors under a certain minimum number of correct matches requirement. The precision value is used to rank the methods. Each row corresponds to a different distance threshold used to define matches.

6.1.2.1 Discussion of Rank Score Results

Descriptor Rank Score Performance. As in the pass rate experiment, the SIFT descriptor always outperforms Shape Context when paired with the same detector.

Detector Rank Score Performance. The Hessian-Laplace detector is the best-ranked detector for all number of correct matches tolerances. In cases where a minimum of 10 correct matches or more are required, the DoG has the second best rank score. In the case where only 2 correct matches are required, however, there is considerably more competition from the Harris-Laplace detector. Specifically, Harris-Laplace marginally outranks DoG at distance thresholds of 250 or greater, even though the DoG has an approximately 40% higher pass rate at these thresholds. The reason for this is that Harris-Laplace almost always outranks DoG when both detectors find some correct matches between image pairs, while DoG accumulates a score in a large portion of pairs where Harris-Laplace does not, resulting in a fairly even rank between the two.

6.1.3 General Discussion

Comparison with Other Evaluation Studies. The direct comparison of our results to other evaluations is not straightforward, since our evaluation is performed on pairs of images where an uncontrolled amount of physical change has occurred under different performance criteria (i.e. no other evaluation uses the notion of pass rate). Some similarities exist though, and they are worth mentioning, especially when trying to express the difficulty of feature-based matching with historic repeat photography.

In the Mikolajczyk and Schmid study [22] on affine feature detectors several transformations were studied that were not subject to geometric change. While the primary criterion used in this study was the repeatability rate, the number of correct

matches was given (based on the SIFT descriptor), giving a point of reference. The most difficult case of jpeg compression occurred when the compression parameter was set to 2% (i.e. a significant and noticeable amount). The Hessian and Harris-Affine methods still were able to find over 400 matches, while MSER achieved approximately 100. In the most difficult cases of change for blur and illumination more than 100 correct matches found for these detectors. In comparison, only 20% of the images in our dataset could attain more than 100 correct matches (given any amount of precision).

An evaluation that gave results for precision and correct matches can be found in the Mikolajczyk *et al.* performance evaluation on descriptors [33]. The most difficult test case was a viewpoint change of 50^{o} on a planar scene. Their matching experiment with the SIFT descriptor obtained a 44% precision with 177 correct matches. In comparison, at best only 38% (with Hessian-Laplace/SIFT) of our dataset was passed with minimum requirements of 40% precision and 2 correct matches. The situation in worsened if we require a comparable amount of correct matches achieved in the Mikolajczyk test. The Mikolajczyk result is, in fact, much closer to some of our best performing image pairs; Figure 30 shows example of an image pair that achieved a 47% precision with 274 correct matches.

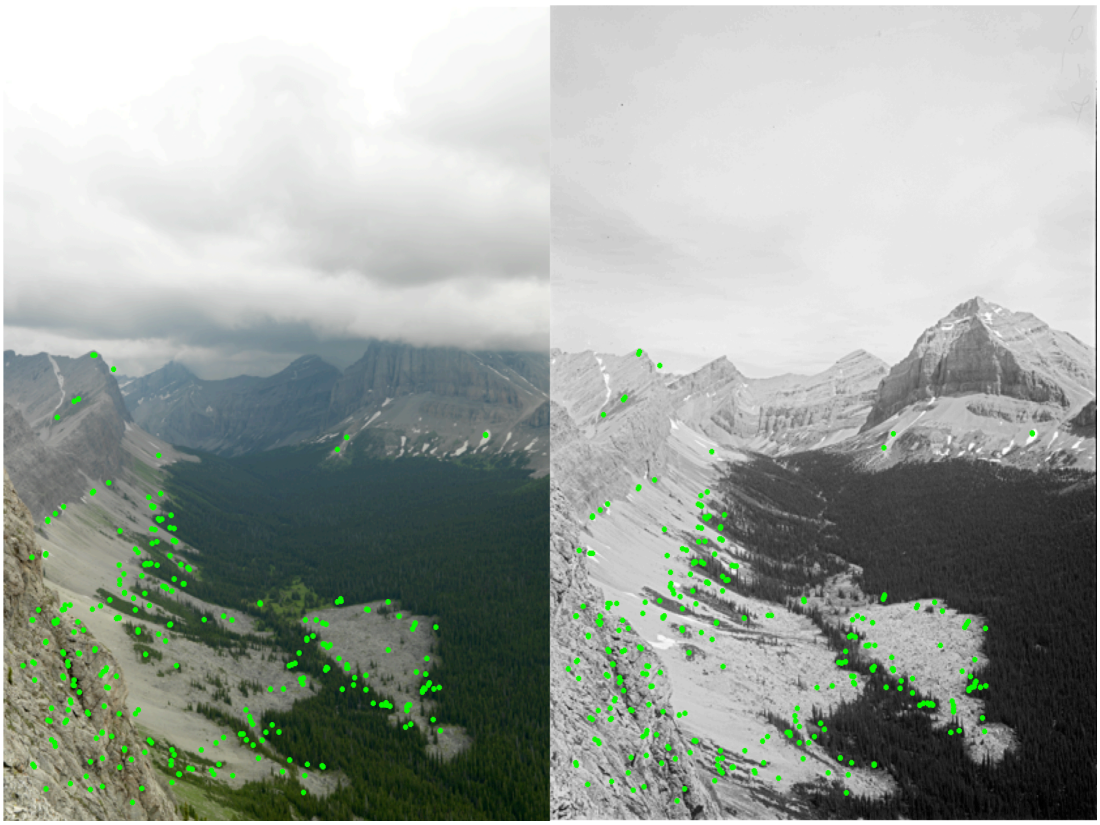


Figure 30 Image pair that performed relatively well, attaining over 200 matches with a precision of 47% (Hessian-Laplace). The green dots represent correct matches.

The dataset used by Valgren and Lilienthal [29] in their evaluation of SIFT and SURF on outdoor images with seasonal changes is perhaps a better point of comparison. However, SIFT was still able to achieve an average precision of 89% with 248 correct matches. This suggests that the additional complexities of historic repeat photography, the difference in camera, unknown calibrations, and photo degradation, make it a much more difficult problem.

Considerations for the Quantity of Features Detected. Both Hessian-Laplace and DoG are blob detectors based on the second derivatives, suggesting that blob features may be better than corner like features (such as detected by Harris-Laplace) for use with repeat photography. It should be noted that both the Hessian-Laplace and DoG found, on average, significantly more features than both Harris-Laplace and MSER. The Harris-

Laplace follows a similar structure to the Hessian-Laplace and DoG that allows the increase and decrease of the detected features through the adjustment of a parameter (in the case of the Harris-Laplace, it is the threshold applied to the *cornerness* measure). Therefore, one could argue that the Harris-Laplace may achieve better performance by adjusting this parameter so that more features were detected. While the optimal parameter adjustment has not been studied for the Harris-Laplace and Hessian-Laplace, an evaluation by Moreels and Perona [27] did take this into consideration for their affine invariant counterparts (the Harris-Affine and Hessian-Affine [23]). Conveniently, these detectors use the same regions, but differ in that they apply affine adaptation to the region shape. They initially evaluated these detectors (which included the DoG) at numerous parameter settings so that each produced a various number of features. The subsequently evaluation results were presented for the parameter (for each detector) that produced the best results. Both the Hessian-Affine and DoG detectors performed best when detecting more than double the amount of features than the Harris-Affine (the exact parameters used with the detectors was note given). In our case, however, the Hessian-Laplace produces 4 times more and the DoG has 8 times more than the Harris-Laplace, suggesting that better results might obtained from the Harris-Laplace by reducing the *cornerness* threshold.

The Failure of MSER Detection in Historic Repeat Photography. The MSER performed the worst, even though the study in [22] showed that it performs well for viewpoint and affine illumination change for both textured and structured planar scenes. This detector is based on the stability of connected components created when incrementing a binary threshold on the images. When viewpoint is the only

transformation on a planar scene, the same connected components are stable because there are no occlusions (see Figure 13 in Chapter 4). It is only the shape of the stable connected components that changes and needs to be adapted for matching. When affine illumination is applied to an image, most connected components persist, but at different levels of stability. However, in the case of complex and mixed changes of a scene, the nature (the shape and stability) of the connected components drastically changes. This can be seen in Figure 31. This example shows a significant change in illumination (direction) and atmospheric conditions (haze). Each image sequence has had a binary threshold set at a progressively higher value. The textured details of the foreground mountain in the historic image are fairly stable connected components. However, with the change in illumination direction and atmosphere, these details are no longer shown as connected components. In general, when a landscape scene that has an atmospheric presence, the stability of potential MSER components are reduced because there is a gradual change in contrast relative to distance. This can be seen in both the historic and repeat image in Figure 31. The reduction in stability decreases the amount of features detected by MSER. It should be noted that the example in Figure 31 achieves a pass with the Hessian-Laplace and DoG methods.

Figure 31 An example of how MSER fails in circumstances of non-affine illumination and atmospheric change. Clearly the stability and persistence of MSER regions are more difficult to achieve in historic repeat photography compared to a viewpoint change of a planar scene.

Considerations for Feature-Based Batching in Historic Repeat Photography. One aspect of a repeat photograph pair that limits any feature-based matching system from achieving numerous matching points is the fact that some of these images contain very few stable features, and thus, the ability to find any correct matches at all is an achievement in itself. An example of such a landscape is shown in Figure 32. Only 2 correct matches were found near the horizon. The number of correct matches, then, are limited by the specific scene. The precision, however, is not. We've noticed that to achieve at least some correct matches over the entire dataset the distance threshold needs to be set relatively high (over 200) compared to the optimal operating threshold of 141

found by Ke and Sukthankar [34]. This inevitably produces numerous false matches (poor precision). A portion of the false matches that arise in our data is due to the rotation invariance of the descriptors. Before the region of a feature is given to the descriptor it is rotated with respect to the orientation of its largest gradient magnitude. When it is known beforehand that an image pair does not differ in rotation, this can result in the matching of areas with different orientations. An example of such a false match is given in Figure 33. In repeat photography, the need for rotation invariance is negligible because the photographer naturally orientates the camera with the historic image, and therefore better results may be obtained by not performing the rotation invariance step before description. The benefit of this has been seen in other applications that do not need to consider the rotation of the scene [29], [37]. Another element that contributes to the high amount of false matches is the matching between features of considerably different scales. If scale-invariance is not required, or the scale change between the images is approximately known, then matches that do not agree on the *a priori* information may be removed.

In terms of detection only (as opposed to a full matching system), we note that the saliency of stable (visually persistent features) regions of the image change between a historic and a repeat image, and thus in the general case of repeat photography, detectors should be used with a loose restriction on their saliency parameter (e.g. as described earlier, lowering the *cornerness* threshold of the Harris detector).

Figure 32 An image pair in our dataset where few correct matches were found. Notice that the majority of the landscape has changed, leaving very few stable features to be detected and matched. The circles represent the scale of the local features. In this case (DoG/SIFT, distance threshold = 250), 334 false matches existed in the proposed match set (not shown).

Figure 33 A false match that occurred because of the rotation invariance of the descriptor.

6.2 Raw Image Pairs

In this section we present the results for the pass rate and pass rate with RANSAC experiments on the raw image pairs. The raw image pairs differ in scale and field-of-view (the amount of the scene captured), encompassing a practical problem faced by the MLP organization; how can historic and repeat images be aligned without the need to hire and train an employee to perform the task manually? The pass rate experiment provides an extension of the evaluation in the previous section. Specifically, we examine how much the performance degrades in the more practical, yet more challenging, matching conditions. Furthermore, an alternative approach to our previous passing criteria is given

in the RANSAC experiments, where a pass is based on how close the RANSAC estimated transformation is compared to the ground truth transformation. These experiments also consider a case where geometric constraints are used to improve precision in the match sets.

6.2.1 Pass Rate

The pass rate for the registered, raw, and raw with geometric filtering (spatial constraint) image pairs can be seen in Figure 34. The results are presented for a distance threshold of 250 and can be interpreted in a similar manner explained in Section 6.1.1 (pass rate for registered image pair results).

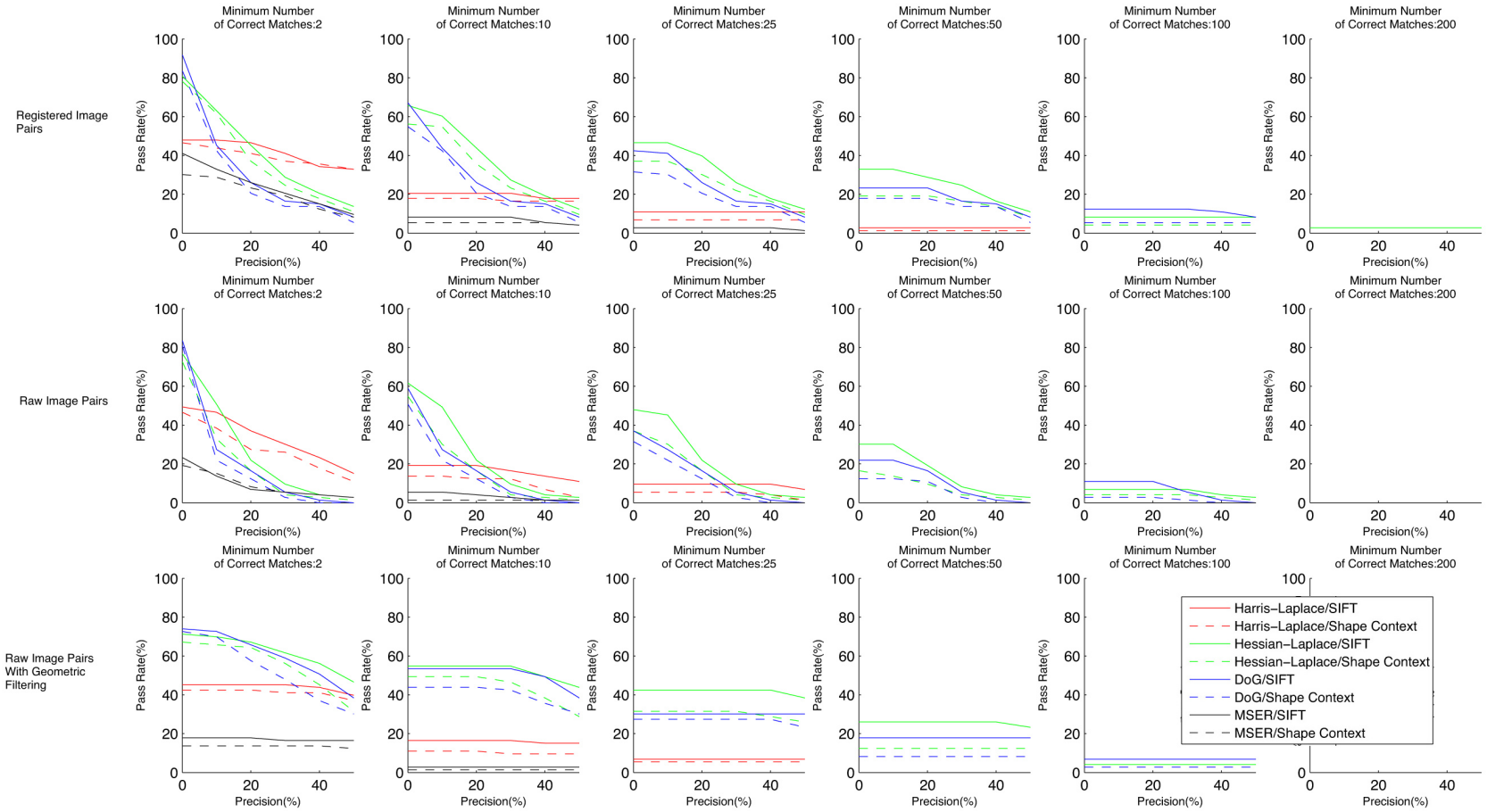


Figure 34 Pass Rate results for registered, raw, and raw with geometric filtering image pairs. The distance threshold was fixed to 250 for this experiment. Registered results are the same as the corresponding row in Figure 28, but have been included here for the purpose of comparison.

6.2.1.1 Discussion of Pass Rate Results

Comparison of Registered and Raw Pass Rates. When comparing the registered results to their raw counterpart, there is a shift downward for each respective pass rate curve. This shift indicates an overall decrease in precision, which can be due to either a decrease in number of correct matches, an increase in the number of false matches, or a combination of both. The main contributing factor is a significant increase in number of false matches, which doubled in frequency, while the quantity of correct matches decreases by 12%. This increase in false matches is caused by the introduction of non-common area between the image pairs rather than differences in scale and resolution between the images. The decrease in correct matches is expected, and is caused by the geometric differences between the images. A similar reduction in correct matches due to scale change was found by Mikolajczyk *et al.* [22].

Spatial Constraints. The pass rate for the raw image pairs with geometric filtering constraints is presented in Figure 34 (third row). When comparing with the non-filtered results, there is a clear and significant increase in overall precision, with the greatest improvement coming at a precision tolerance of 50% and a number of correct matches tolerances of 2, where the Hessian-Laplace/SIFT method went from a 1.3% pass rate to a 46.6% pass rate. The total amount of correct matches over the entire dataset (with all detector-descriptor) was reduced by 16% compared to the non-filtered matching, a consequence of some image pairs not meeting the assumption that the center of the raw images are approximately the same. The total amount of false matches was reduced by

93% compared to the non-filtered matching, further demonstrating the benefit that this simple filtering technique had on precision.

6.2.2 Pass Rate with RANSAC

The pass rate with RANSAC graphs for the raw and raw with geometric filtering image pairs can be seen in Figure 35. The top row denotes the RANSAC results for raw image pairs, while the bottom row denotes RANSAC results for the raw pairs with geometric filtering. Each graph shows the change in pass rate as a change in one of the transformation tolerances (with the other two tolerance values are fixed). The fixed tolerances are: scale = 0.01%, rotation = 0.5 degrees, and translation = 5 pixels.

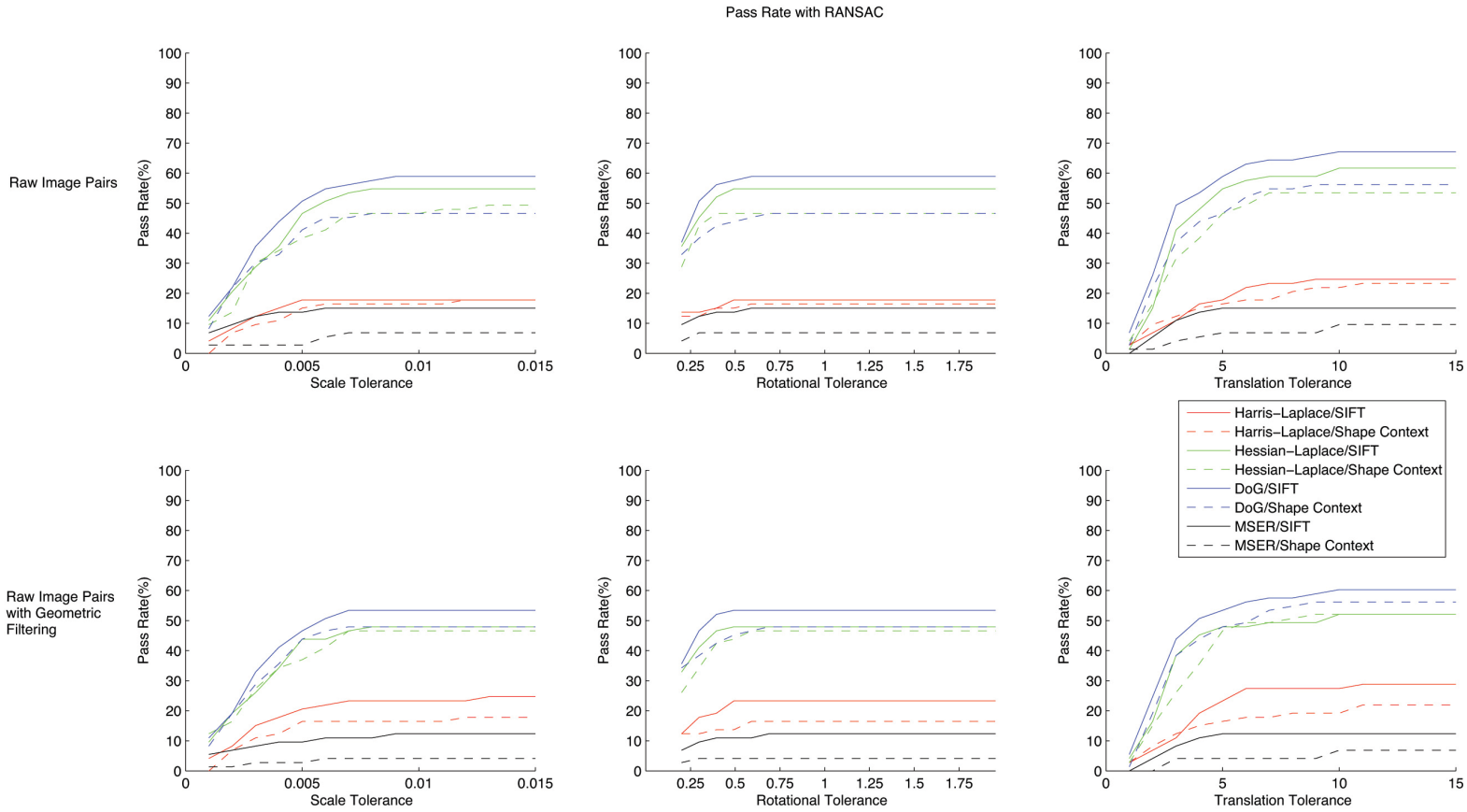


Figure 35 Pass Rate with RANSAC Results. The top row denotes the RANSAC results for raw image pairs, while the bottom row denotes RANSAC results for the raw pairs with geometric filtering. Each graph shows the change in pass rate as a change in one of the transformation tolerances (with the other two tolerance values are fixed). The fixed tolerances are: scale = 0.01%, rotation = 0.5 degrees, and translation = 5 pixels

6.2.2.1 Discussion of Pass Rate with RANSAC Results

Stable Transformation Error Tolerances. We have based our passing criteria on transformation tolerances for scale, rotation, and translation. By varying these tolerances, the pass rate increases or decreases. A stable tolerance, then, is when an increase in transformation tolerances does not show a significant increase in pass rate. Based on the results shown in Figure 35, an approximate stable tolerance is reached for a scale tolerance of 0.01, a rotation tolerance of 0.5 degrees, and a translation tolerance of 10 pixels (a summation of the translation error in both the x and y direction). The stable tolerance for scale and rotation is insignificant; at such errors a difference is barely noticeable. The translation tolerance, however, is significant enough for further discussion. We found that there are two main reasons why this translation error arises. The pair of images shown in Figure 30 demonstrates the first reason. While there are significant correct matches found with relatively high precision, the matches clearly favour the right side of the image pair, covering a large amount of depth in the scene. Naturally, the right side of the image pair ends up better aligned than the left side. This isn't a problem if there is no repeat error between the images, but in this case a slight viewpoint change does exist. Interestingly enough, the MLP lab assistant chose to align the parts of the images in the exact opposite side where the matches were found (the mountain peak occluded by clouds). The result is a translation error of 5.2 pixels. The second reason for a larger stable translation tolerance is due to landscape change. Figure 36 demonstrates the issues that can occur when matches are located near changing structure. This image pair was well-aligned, with very little, if any, repeat error. With

close inspection of the split views, we can see a rise in tree canopy height and an encroachment of trees into the valley meadow. As it happens, these are also areas where features were matched. These changes in landscape cause the location of the matches to shift as well. When this occurs at several feature locations over the image, the estimated transformation by RANSAC is inherently more prone to error. The translation error between these images is 6.2 pixels. However, in contrast to the previous example, the translation does not make the alignment better at any specific location within the image, and therefore may be considered an insufficient registration.

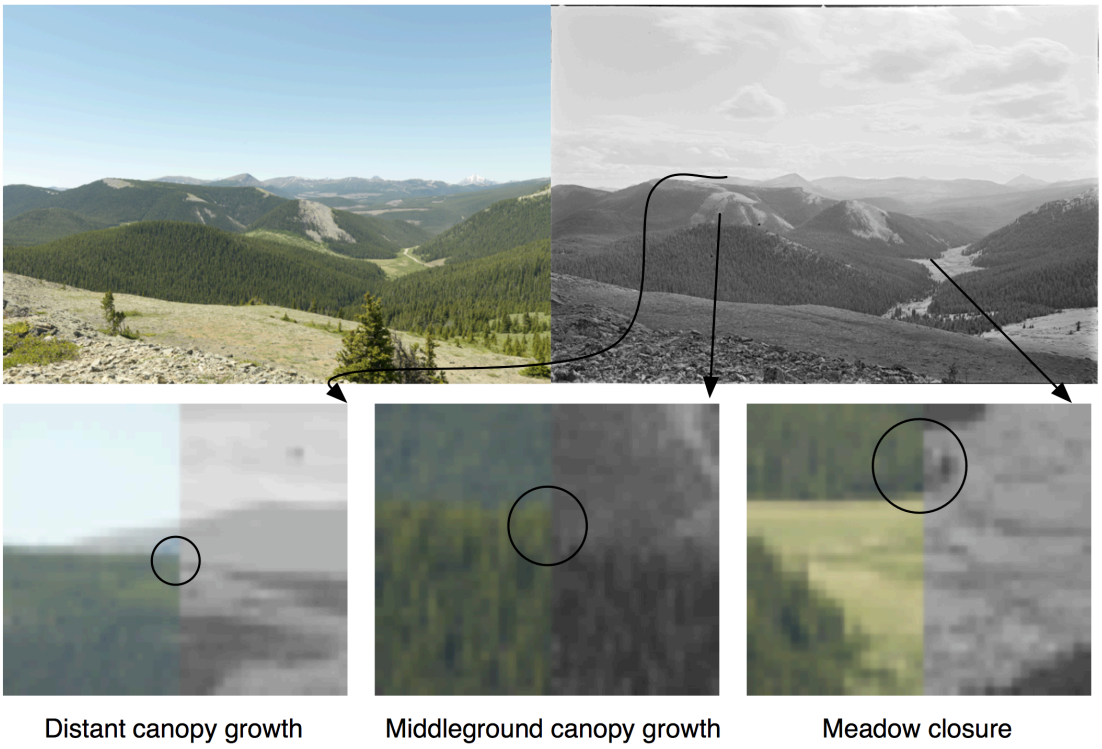


Figure 36 An example of a well-aligned image pair with landscape change. Tree growth and meadow closure cause movement in the contours of the image, effecting localization accuracy of features, which may lead to poor automatic alignment.

Spatial Constraints. The geometric spatial constraints removed matches from the match set which did not meet a constraint that was based on an expected scale change

between image pairs. As was demonstrated in Section 6.2.1, this constraint significantly increased overall precision. However, as we have previously mentioned, in the context of the registration of these image pairs, low precision proved to be less influential than expected, and therefore the subsequent increase in overall precision has little effect on the RANSAC pass rate. In fact, the RANSAC pass rate in this circumstance was slightly lower, a phenomena that can be attributed to a certain image pair having a greater than expected translation (specifically, image pair 65 listed in Appendix A). The filtering of the false matches does, however, increase the speed at which RANSAC can execute. With the current implementation of filtering, the extra cost of obtaining the filtered match set did not outweigh the speed benefits in RANSAC, and therefore we do not recommend the proposed geometric spatial constraint.

6.2.3 General Discussion

Comparison of Pass Rate with RANSAC and Pass Rate Results. Only two matching points between an image pair are required to resolve a scale, rotation, and translation. With this in mind, when we compare the RANSAC pass rates with the results from the previously used pass rate with the minimum number of correct matches tolerance set at 2 (Figure 34), we can see a clear difference in the highest results of each experiment; the maximum RANSAC pass rate is 67.1%, while the maximum pass rate is 80.8%. When examining the cases that passed our RANSAC experiment (at the stable tolerance values), we found good estimations of the transformation were obtained when precision was as low as 1%. The number of correct matches, on the other hand, were found to be more important for a successful transformation estimation. While technically only two matches are needed to estimate the transformation, more are needed for a more reliable

result. For example, when we consider all the detectors-descriptors and their performance at each image pair in our dataset, there are a total of 332 instances where less than 6 correct matches are found. Of these 332 instances, only 17 are passed with the RANSAC criteria.

The best performing detector is DoG paired with SIFT, with a RANSAC pass rate of 67.1%. Even when the translation tolerance is lowered from the stable tolerance of 10 pixels to 3 pixels, 50% of the dataset still pass and should be able to be registered reasonably well. As a reminder, these results are limited to the context of this application.

6.3 Recommendations

Based on the results from the registered and raw image pair experiments two detectors and one descriptor have shown give the best results. For applications where the only geometric difference between the images is scale, rotation, and translation, the transformation can be estimated regardless of low precision in the result. The practical problem of image alignment within the MLP is one such application, and therefore, in this case we recommend the DoG detector paired with the SIFT descriptor. In general, higher precision may be required for other applications (e.g. viewpoint estimation). In such circumstances, we recommend the Hessian-Laplace detector paired with the SIFT descriptor for feature-based matching.

6.3 Threats to Validity

The main metric of comparison in this study is a percentage of passing image pairs on our sample dataset. The sample dataset used in this evaluation comprised of images from geographical locations and dates that spanned an area and temporal range representative of the MLP database. Ideally, the image pairs selected in our sample would have been

chosen at random, however, due to organizational inconsistencies in the MLP database, the choice of image pairs of the sample was limited. This may affect the validity of the pass rate percentage values, however, the comparative results still hold since the detector-descriptors are being compared under the same conditions. We note that it is inherently difficult to qualify what is representative of historic repeat photography as a whole. The MLP database itself is composed of only a small percentage of the tens of thousands of survey images that are currently archived at Library and Archives Canada. As more of these historic images are repeated, the characteristics of the population may change.

A further threat to validity is based on the fact that the historic repeat photography examined in this evaluation are specific to landscape scenes. A dataset composed of repeated urban scenes may yield different results since the structure of the scene is different (i.e. urban scenes tend to have much more straight lines than landscape). Each scene, does, however, share common characteristics that make detection and description a difficult task.

Chapter 7 Conclusions and Future Work

7.1 Conclusions

This thesis evaluated a set of detectors and descriptors in the context of matching features between historic repeat photographic pairs. In doing so, we have proposed an evaluation method for this type of imagery and presented the first quantitative results relating to the feasibility feature-based matching with historic repeat photography, where images may differ on a variety of factors such as, weather, illumination, atmospheric, acquisition device, and landscape change.

The dataset used included 73 unique image pairs from the Mountain Legacy Project. In comparison to other evaluations, only one [27] has used more unique scenes (Moreels and Perona used 100 different objects).

Our theoretical results were based on the registered image pairs, and showed the characteristic of precision and number of correct matches obtained over the dataset. Specifically, the results showed that the task of finding matching features is much more difficult compared to other geometric and photometric transformations that have been evaluated. In the pass rate and rank score experiments, the SIFT descriptor outperformed Shape Context when paired with the same detector. For the detectors evaluated, the Hessian-Laplace was shown to perform best in most situations, while the DoG detector was able to find at least some correct matches in more image pairs than the Hessian-Laplace, albeit at the sacrifice of precision. Both these detectors were able to detect corresponding features for almost every pair in the dataset, which is a considerable feat considering that these detectors have not been designed to handle such nonlinear and uncontrolled differences. The description and matching of these correspondences proved

a difficult task; the precision at which correct matches were obtained was low (below 20%) for the most image pairs in the dataset. We suggest that precision would likely improve if the rotation invariance of the descriptors were removed. The results further demonstrate that applications that require large amounts of matched features (greater than 50) or high precision are not well-suited to historic repeat photography with even the best performing detector descriptor in our evaluation.

We also examined the feasibility of using a feature-based matching system to align the raw image pairs in our dataset. Surprisingly, we found that low precision was less of a contributing factor than a low number of correct matches for a failure in this experiment. Specifically, passes (based on the pass rate with RANSAC experiment) were obtained with a precision value as low as 1%, while very few image pairs that had less than 5 correct matches were passed. This resulted in the less precise DoG/SIFT method performing the best. The pass rate with RANSAC was based on a transformation error tolerance between an estimated transformation and a human created ground truth transformation. We found that adjusting the translation tolerance affected the pass rate results the most, exposing some potential problems with using feature matching in repeat photography; changes in landscape can cause slight shifts in feature location which results in transformation estimations that are more prone to error, and matches may not be located on areas of the image that a human observer would consider most important for alignment, which causes a translation movement with the presence of repeat error. Whether these feature-based methods are sufficient is dependent on the accuracy required (which may differ by application). With that being said, DoG/SIFT was still able to pass 50% of the dataset with relatively low translation tolerances (scale tolerance = 0.01,

rotation tolerance = 0.5 degrees, and translation tolerance = 3 pixels). The use of *a priori* information, while increasing precision, had no benefit to the RANSAC pass rate results, since precision was not a major contributing factor to a pass in this sense.

The genesis of this work was the automation or semi-automation of change detection, a topic that has great potential benefit to research related to historic repeat photography. In reviewing literature on change detection, a couple things became clear: a pre-processing step of image registration is typically performed before the application of a change detection algorithm, and before we can examine the ability of computer vision methods to detect change, it may be advantageous to see if we can first detect similarities. In this thesis we've made some initial and incremental strides towards these goals, showing that indeed some similarities can be discerned between the pairs, and furthermore, registration based on simple transformations can successfully performed. This research is really only the tip-of-an-iceberg for an application that has seen little work in computer vision, and therefore we hope for the continued progress and interest in this field. The following section presents some ideas and possibilities in future work.

7.2 Future Work

One aspect of the automatic registration that is performed in this thesis is that it is compared to a ground truth created with scale, rotation, and translation transformations. With the existence of repeat error between images, these transformations do not suffice, meaning that the sub-pixel registration that might be required for a change detection algorithm is not achieved. To get this accuracy, one needs many matching points spread over the objects of the images. With these points, a nonlinear transformation could be applied to the image for better registration. Therefore, future work could focus on the

distribution of the features across the scene and the use of nonlinear transformation methods for registration. Furthermore, more correct matches with higher precision may be achieved if the registered image pairs are used with non-scale invariant detectors (like the Harris) and matches are only searched for in close proximity.

Another application of feature detector and description for the MLP data comes in the form of tracking boundary movements. The potential problem of moving landscape features such a tree growth, tree lines, and meadow closures for registration could be leveraged to help visualize boundary change between image pairs. Specifically, an image pair would first be registered, and then feature matched again. Stable features would see no change in location between images, but unstable features, on the other hand, would show movement, which could be overlaid on the image with an arrow.

Practically, both these methods would most likely require some expert intervention, such as the manually culling of bad matches, or in the case of registration, the adding of more match points across the images.

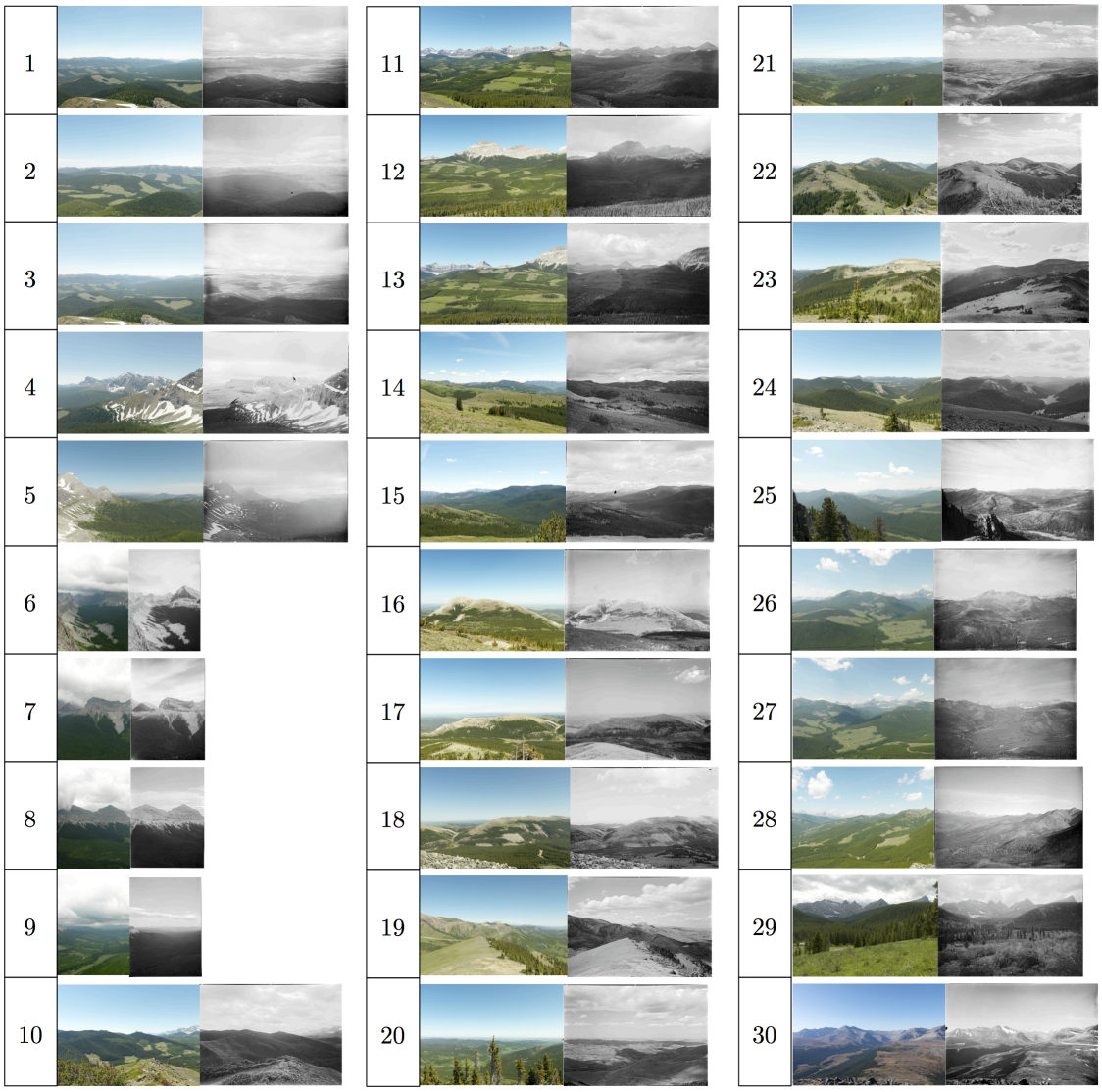
The feature-based matching systems evaluated in this thesis were based on different detectors and descriptors, while fixing the matching approach used. Therefore, it may also interesting to evaluate another matching strategy, especially the nearest neighbour ratio distance which is an integral part of the SIFT method developed by Lowe [17]. Given the indistinctiveness of the described features (evident through low precision and the need for high distance thresholds), we assume that this would have an overall negative effect on the pass rate, as many correct matches would likely be removed. Nonetheless, such an experiment would yield interesting results, especially since this type of matching is in common use.

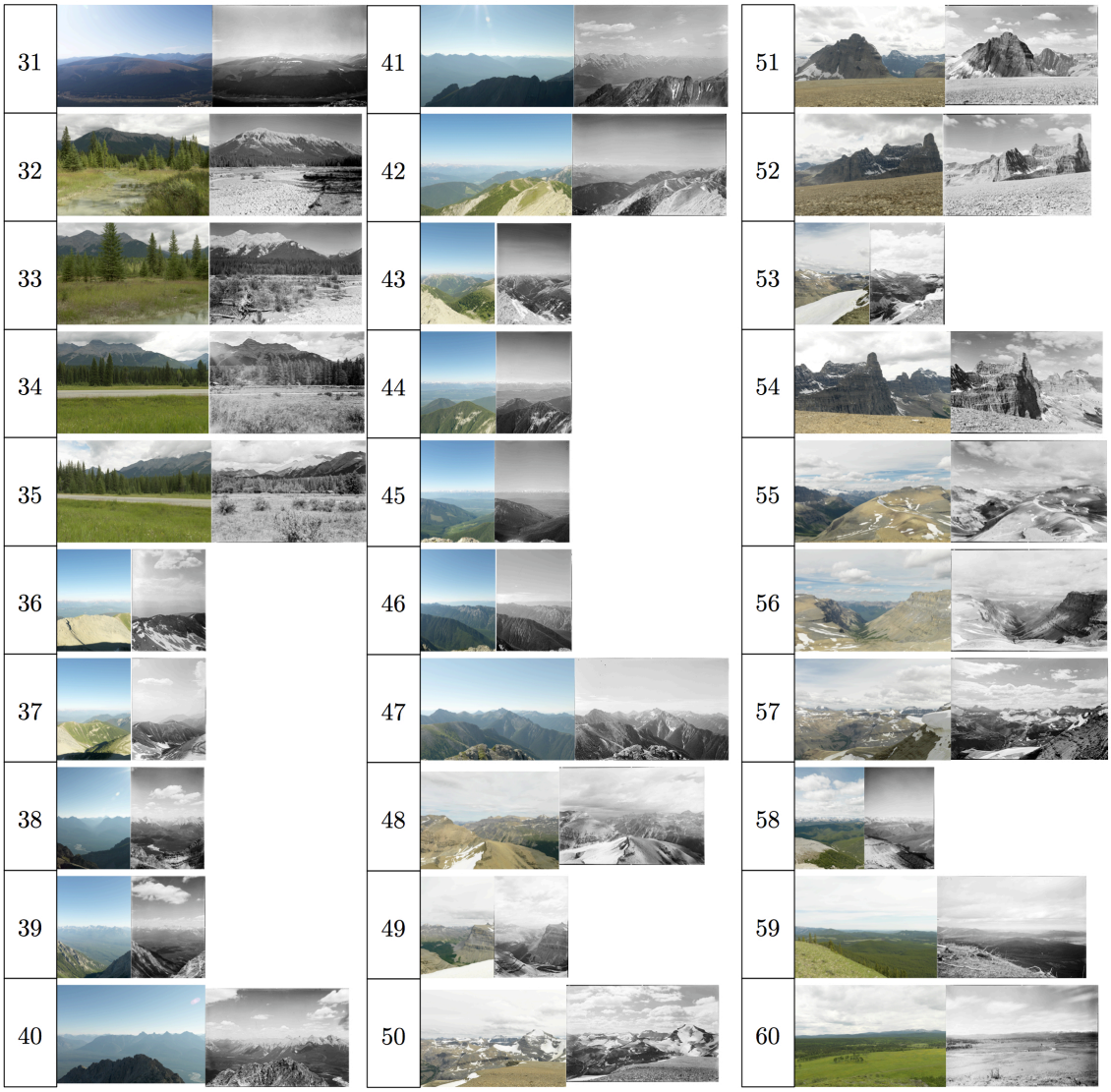
Finally, this thesis only examined a handful of detectors and only a couple descriptors. It therefore may be interesting to extent this evaluation to some other methods, especially SURF [30], which has shown to perform as well or better in some circumstances than the methods used here.

Appendix A
Dataset

This appendix provides two tables with exact information pertaining to the images used as the data set for this thesis. Table 6 shows the registered image pairs and index numbers that can be used to reference the image information in Table 7. Table 7 lists the survey, station, repeat file name, historic file name, and the ground truth transformations that are required to register the raw image pairs.

Table 6 Dataset image pairs.




31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

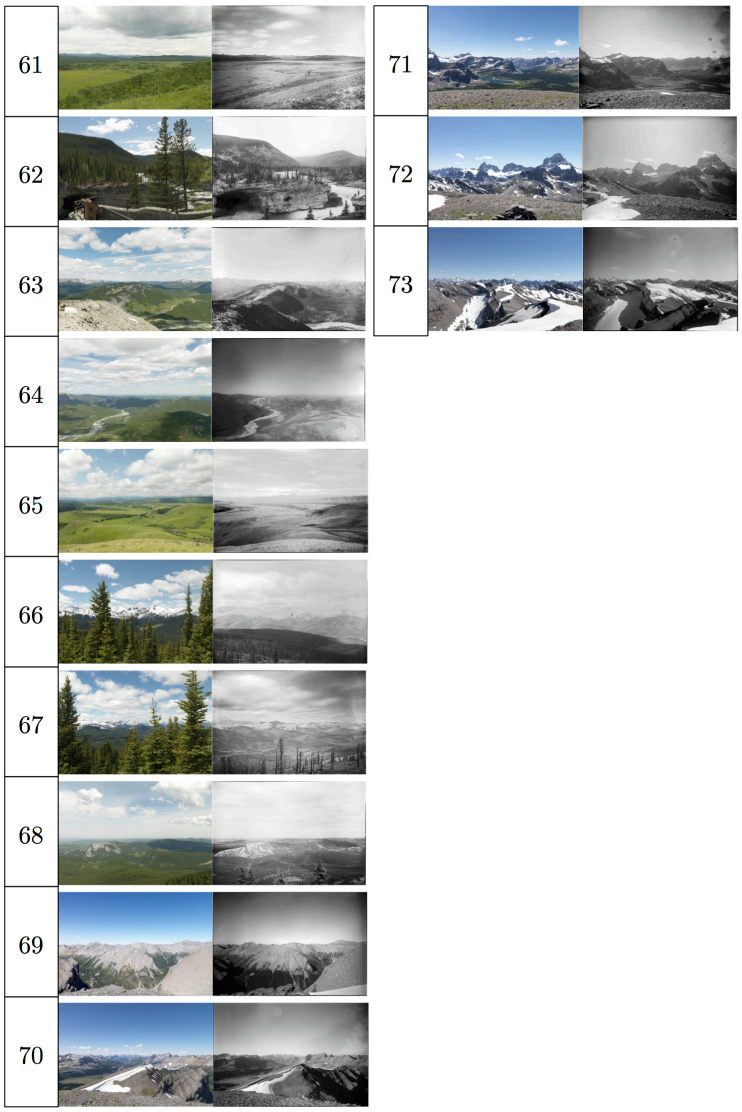

61
62
63
64
65
66
67
68
69
70
71
72
73

Table 7 Dataset reference information.

Index	Surveys	Stations	Repeat File Name	Original File Name	Scale Change	Rotation Change	Centre Offset X	Centre Offset Y
1	'Bridgland_1913-14_Crowsnest Forest Reserve - Waterton Lakes N.P.'	'Stn. 12 Boundary No. 2A'	'B 1850s'	'BRI1913_B13-83s'	84.32%	-7.14E-03	8.4	14.3
2	'Bridgland_1913-14_Crowsnest Forest Reserve - Waterton Lakes N.P.'	'Stn. 12 Boundary No. 2A'	'B 1852s'	'BRI1913_B13-84s'	84.08%	3.10E-03	14.2	15.7
3	'Bridgland_1913-14_Crowsnest Forest Reserve - Waterton Lakes N.P.'	'Stn. 12 Boundary No. 2A'	'B 1854s'	'BRI1913_B13-86s'	83.43%	4.31E-03	8.2	14.8
4	'Bridgland_1913-14_Crowsnest Forest Reserve - Waterton Lakes N.P.'	'Stn. 12 Boundary No. 2A'	'B 1857s'	'BRI1913_B13-82s'	84.40%	-8.09E-03	8.9	14.8
5	'Bridgland_1913-14_Crowsnest Forest Reserve - Waterton Lakes N.P.'	'Stn. 12 Boundary No. 2A'	'B 1859s'	'BRI1913_B13-85s'	84.71%	-2.93E-03	11.4	16.0
6	'Bridgland_1913-14_Crowsnest Forest Reserve - Waterton Lakes N.P.'	'Stn. 27 Dutch Creek Head No. 1'	'B_1941s'	'BRI1913_B13-157s'	79.77%	-3.30E-03	10.6	5.0
7	'Bridgland_1913-14_Crowsnest Forest Reserve - Waterton Lakes N.P.'	'Stn. 27 Dutch Creek Head No. 1'	'B_1943s'	'BRI1913_B13-158s'	80.60%	6.30E-03	8.7	4.4
8	'Bridgland_1913-14_Crowsnest Forest Reserve - Waterton Lakes N.P.'	'Stn. 27 Dutch Creek Head No. 1'	'B_1945s'	'BRI1913_B13-159s'	78.92%	3.48E-03	6.7	7.2
9	'Bridgland_1913-14_Crowsnest Forest Reserve - Waterton Lakes N.P.'	'Stn. 27 Dutch Creek Head No. 1'	'B_1950s'	'BRI1913_B13-160s'	79.55%	1.62E-04	5.4	3.4
10	'Bridgland_1913-14_Crowsnest Forest Reserve - Waterton Lakes N.P.'	'Stn. 43 Grassy Ridge'	'B 1756s'	'BRI1913_B13-259s'	82.68%	-3.46E-03	10.6	15.1
11	'Bridgland_1913-14_Crowsnest Forest Reserve - Waterton Lakes N.P.'	'Stn. 43 Grassy Ridge'	'B 1757s'	'BRI1913_B13-260s'	82.91%	-5.03E-03	7.7	16.6
12	'Bridgland_1913-14_Crowsnest Forest Reserve - Waterton Lakes N.P.'	'Stn. 43 Grassy Ridge'	'B 1759s'	'BRI1913_B13-261s'	81.32%	3.49E-03	10.6	13.0
13	'Bridgland_1913-14_Crowsnest Forest Reserve - Waterton Lakes N.P.'	'Stn. 43 Grassy Ridge'	'B 1763s'	'BRI1913_B13-263s'	82.19%	3.41E-03	7.7	17.5
14	'Bridgland_1913-14_Crowsnest Forest Reserve - Waterton Lakes N.P.'	'Stn. 43 Grassy Ridge'	'B 1765s'	'BRI1913_B13-257s'	82.56%	7.66E-03	7.2	15.6
15	'Bridgland_1913-14_Crowsnest Forest Reserve - Waterton Lakes N.P.'	'Stn. 43 Grassy Ridge'	'B 1767s'	'BRI1913_B13-258s'	82.31%	1.59E-02	7.2	15.8
16	'Bridgland_1913-14_Crowsnest Forest Reserve - Waterton Lakes N.P.'	'Stn. 54 Sentinel Pass West No. 2'	'B 1834s'	'BRI1913_B13-310s'	82.08%	-9.99E-04	9.0	15.7
17	'Bridgland_1913-14_Crowsnest Forest Reserve - Waterton Lakes N.P.'	'Stn. 54 Sentinel Pass West No. 2'	'B 1835s'	'BRI1913_B13-311s'	83.46%	-6.16E-03	8.6	11.0
18	'Bridgland_1913-14_Crowsnest Forest Reserve - Waterton Lakes N.P.'	'Stn. 54 Sentinel Pass West No. 2'	'B 1837s'	'BRI1913_B13-309s'	82.29%	2.54E-02	8.5	9.4
19	'Bridgland_1913-14_Crowsnest Forest Reserve - Waterton Lakes N.P.'	'Stn. 58 Willow Creek No. 2'	'B 1807s'	'BRI1913_B13-331s'	83.29%	-2.52E-03	8.9	13.6

Index	Surveys	Stations	Repeat File Name	Original File Name	Scale Change	Rotation Change	Centre Offset X	Centre Offset Y
20	'Bridgland_1913-14_Crowsnest Forest Reserve - Waterton Lakes N.P.'	'Stn. 58 Willow Creek No. 2'	'B 1810s'	'BRI1913_B13-332s'	83.01%	7.10E-03	9.0	12.9
21	'Bridgland_1913-14_Crowsnest Forest Reserve - Waterton Lakes N.P.'	'Stn. 58 Willow Creek No. 2'	'B 1811s'	'BRI1913_B13-333s'	83.62%	1.71E-03	2.6	14.5
22	'Bridgland_1913-14_Crowsnest Forest Reserve - Waterton Lakes N.P.'	'Stn. 59 Willow Creek No. 3'	'B 1813s'	'BRI1913_B13-334s'	83.15%	-3.92E-03	7.1	15.9
23	'Bridgland_1913-14_Crowsnest Forest Reserve - Waterton Lakes N.P.'	'Stn. 59 Willow Creek No. 3'	'B 1815s'	'BRI1913_B13-336s'	82.23%	5.90E-04	6.9	16.5
24	'Bridgland_1913-14_Crowsnest Forest Reserve - Waterton Lakes N.P.'	'Stn. 59 Willow Creek No. 3'	'B 1818s'	'BRI1913_B13-335s'	82.46%	9.89E-04	6.9	12.9
25	'Bridgland_1913-14_Crowsnest Forest Reserve - Waterton Lakes N.P.'	'Stn. 6 Bolton No. 1'	'B 1952s'	'BRI1913_B13-43s'	83.90%	-6.13E-03	6.0	13.7
26	'Bridgland_1913-14_Crowsnest Forest Reserve - Waterton Lakes N.P.'	'Stn. 6 Bolton No. 1'	'B 1953s'	'BRI1913_B13-42s'	83.67%	5.21E-03	5.9	14.7
27	'Bridgland_1913-14_Crowsnest Forest Reserve - Waterton Lakes N.P.'	'Stn. 6 Bolton No. 1'	'B 1955s'	'BRI1913_B13-44s'	84.34%	-1.21E-02	9.1	13.7
28	'Bridgland_1913-14_Crowsnest Forest Reserve - Waterton Lakes N.P.'	'Stn. 6 Bolton No. 1'	'B 1957s'	'BRI1913_B13-45s'	83.70%	5.43E-03	7.4	13.7
29	'Bridgland_1913-14_Crowsnest Forest Reserve - Waterton Lakes N.P.'	'Stn. 8 Near Falls in Oldman Valley'	'B 1749s'	'BRI1913_B13-54s'	83.74%	3.24E-05	12.0	8.5
30	'Bridgland_1917-20_Bow River & Clearwater Forest Reserves'	'Stn. 235'	'HB1_A_04627s'	'BRI1919_B19-79s'	75.14%	6.93E-03	10.4	6.5
31	'Bridgland_1917-20_Bow River & Clearwater Forest Reserves'	'Stn. 235'	'HB1_A_04630s'	'BRI1919_B19-78s'	75.23%	-1.52E-03	2.8	2.6
32	'Bridgland_1922-23_Kootenay & Columbia Valleys'	'Stn. 30'	'B 2277s'	'BRI1922_B22-483s'	72.13%	-3.25E-03	10.4	12.4
33	'Bridgland_1922-23_Kootenay & Columbia Valleys'	'Stn. 30'	'B 2278s'	'BRI1922_B22-482s'	71.37%	-5.68E-03	9.4	11.8
34	'Bridgland_1922-23_Kootenay & Columbia Valleys'	'Stn. 31'	'B 2390s'	'BRI1922_B22-480s'	70.35%	-6.70E-04	9.9	7.6
35	'Bridgland_1922-23_Kootenay & Columbia Valleys'	'Stn. 31'	'B 2391s'	'BRI1922_B22-479s'	70.17%	4.02E-03	5.3	5.6
36	'Bridgland_1922-23_Kootenay & Columbia Valleys'	'Stn. 34'	'B 2457s'	'BRI1922_B22-215s'	69.96%	1.69E-03	7.4	16.5
37	'Bridgland_1922-23_Kootenay & Columbia Valleys'	'Stn. 34'	'B 2458s'	'BRI1922_B22-216s'	70.80%	5.64E-03	8.0	16.1
38	'Bridgland_1922-23_Kootenay & Columbia Valleys'	'Stn. 34'	'B 2460s'	'BRI1922_B22-213s'	70.41%	4.29E-03	4.7	17.7
39	'Bridgland_1922-23_Kootenay & Columbia Valleys'	'Stn. 34'	'B 2461s'	'BRI1922_B22-211s'	70.38%	-8.89E-04	4.5	16.5
40	'Bridgland_1922-23_Kootenay & Columbia Valleys'	'Stn. 34'	'B 2463s'	'BRI1922_B22-212s'	69.79%	6.47E-04	6.6	14.0
41	'Bridgland_1922-23_Kootenay & Columbia Valleys'	'Stn. 34'	'B 2464s'	'BRI1922_B22-	70.61%	-2.41E-03	11.1	15.3

Index	Surveys	Stations	Repeat File Name	Original File Name	Scale Change	Rotation Change	Centre Offset X	Centre Offset Y
				214s'				
42	'Bridgland_1922-23_Kootenay & Columbia Valleys'	'Stn. 35'	'B 2467s'	'BRI1922_B22-209s'	70.78%	-2.75E-03	5.5	13.9
43	'Bridgland_1922-23_Kootenay & Columbia Valleys'	'Stn. 35'	'B 2469s'	'BRI1922_B22-210s'	70.35%	1.58E-03	7.5	14.7
44	'Bridgland_1922-23_Kootenay & Columbia Valleys'	'Stn. 35'	'B 2470s'	'BRI1922_B22-208s'	70.14%	-7.62E-04	8.4	14.6
45	'Bridgland_1922-23_Kootenay & Columbia Valleys'	'Stn. 35'	'B 2472s'	'BRI1922_B22-207s'	70.65%	-1.03E-03	10.4	14.0
46	'Bridgland_1922-23_Kootenay & Columbia Valleys'	'Stn. 35'	'B 2473s'	'BRI1922_B22-206s'	70.54%	9.72E-04	8.2	13.7
47	'Bridgland_1922-23_Kootenay & Columbia Valleys'	'Stn. 35'	'B 2474s'	'BRI1922_B22-205s'	70.91%	-3.29E-03	6.6	15.7
48	'Bridgland_1927-28_Brazeau Forest Reserve & Jasper N.P.'	'Stn. 426'	'B 2329s'	'BRI1927_117s'	76.30%	5.38E-03	9.6	13.7
49	'Bridgland_1927-28_Brazeau Forest Reserve & Jasper N.P.'	'Stn. 426'	'B 2330s'	'BRI1927_118s'	76.31%	-4.55E-03	3.9	17.8
50	'Bridgland_1927-28_Brazeau Forest Reserve & Jasper N.P.'	'Stn. 426'	'B 2331s'	'BRI1927_114s'	75.86%	8.27E-04	8.0	13.7
51	'Bridgland_1927-28_Brazeau Forest Reserve & Jasper N.P.'	'Stn. 426'	'B 2333s'	'BRI1927_115s'	76.38%	5.10E-03	7.8	12.7
52	'Bridgland_1927-28_Brazeau Forest Reserve & Jasper N.P.'	'Stn. 426'	'B 2335s'	'BRI1927_116s'	76.19%	-3.66E-03	6.9	17.0
53	'Bridgland_1927-28_Brazeau Forest Reserve & Jasper N.P.'	'Stn. 426'	'B 2336s'	'BRI1927_119s'	76.43%	1.23E-02	10.4	12.6
54	'Bridgland_1927-28_Brazeau Forest Reserve & Jasper N.P.'	'Stn. 426'	'B 2338s'	'BRI1927_110s'	76.08%	1.71E-03	14.1	12.5
55	'Bridgland_1927-28_Brazeau Forest Reserve & Jasper N.P.'	'Stn. 426'	'B 2341s'	'BRI1927_111s'	76.72%	1.78E-03	11.2	13.2
56	'Bridgland_1927-28_Brazeau Forest Reserve & Jasper N.P.'	'Stn. 426'	'B 2342s'	'BRI1927_112s'	76.92%	4.16E-03	0.6	15.2
57	'Bridgland_1927-28_Brazeau Forest Reserve & Jasper N.P.'	'Stn. 426'	'B 2343s'	'BRI1927_113s'	76.25%	7.20E-03	3.4	15.9
58	'Wheeler_1895-1900_Canadian Irrigation Survey'	'Moose Mt Centre'	'B 1672s'	'WHE1896_W96-12-3s'	80.13%	7.29E-03	13.9	-2.9
59	'Wheeler_1895-1900_Canadian Irrigation Survey'	'Nichi'	'B 1690s'	'WHE1896_W96-23-1s'	79.10%	9.42E-03	11.5	18.3
60	'Wheeler_1895-1900_Canadian Irrigation Survey'	'Barnes Ridge'	'B 1744s'	'WHE1897_W97-10-7s'	80.19%	2.80E-03	9.8	16.1
61	'Wheeler_1895-1900_Canadian Irrigation Survey'	'Barnes Ridge'	'B 1745s'	'WHE1897_W97-11-7s'	80.74%	-2.19E-03	12.1	14.2
62	'Wheeler_1895-1900_Canadian Irrigation Survey'	'Elbow Falls'	'B 1540s'	'WHE1897_W97-13-11s'	79.06%	-3.54E-03	6.2	20.5

Index	Surveys	Stations	Repeat File Name	Original File Name	Scale Change	Rotation Change	Centre Offset X	Centre Offset Y
63	'Wheeler_1895-1900_Canadian Irrigation Survey'	'Forget-me-not-ridge'	'B 1668s'	'WHE1897_W97-22-8s'	80.01%	-2.95E-03	4.1	14.3
64	'Wheeler_1895-1900_Canadian Irrigation Survey'	'Forget-me-not-ridge'	'B 1670s'	'WHE1897_W97-23-8s'	81.69%	1.78E-03	7.2	15.7
65	'Wheeler_1895-1900_Canadian Irrigation Survey'	'Ing"s Flats-Triangulation'	'B 2171s'	'WHE1897_W97-1-19s'	80.25%	4.84E-03	-42.6	17.1
66	'Wheeler_1895-1900_Canadian Irrigation Survey'	'Jumping Pound North'	'B 1603s'	'WHE1897_W97-23-9s'	81.43%	7.00E-03	7.5	17.3
67	'Wheeler_1895-1900_Canadian Irrigation Survey'	'Jumping Pound North'	'B 1610s'	'WHE1897_W97-24-9s'	81.05%	1.96E-03	10.6	15.0
68	'Wheeler_1895-1900_Canadian Irrigation Survey'	'Powderface Ridge'	'B 1718s'	'WHE1897_W97-20-9s'	80.33%	-4.20E-03	7.2	14.8
69	'Wheeler_1913-24_Interprovincial Boundary Survey'	'Divide Hill'	'HB1_A0005270s'	'WHE1913_W13-292s'	80.44%	-3.67E-03	6.0	7.4
70	'Wheeler_1913-24_Interprovincial Boundary Survey'	'Divide Hill'	'HB1_A0005272s'	'WHE1913_W13-291s'	80.34%	-2.37E-03	1.1	5.7
71	'Wheeler_1913-24_Interprovincial Boundary Survey'	'Divide Hill'	'HB1_A0005273s'	'WHE1913_W13-290s'	80.24%	2.45E-03	8.2	4.6
72	'Wheeler_1913-24_Interprovincial Boundary Survey'	'Divide Hill'	'HB1_A0005274s'	'WHE1913_W13-289s'	79.13%	1.63E-03	6.2	7.2
73	'Wheeler_1913-24_Interprovincial Boundary Survey'	'Divide Hill'	'HB1_A0005276s'	'WHE1913_W13-288s'	79.26%	4.34E-03	9.7	9.6

Bibliography

[1] D. Levere, B. Yochelson, and B. Abbott, *New York Changing: Revisiting Berenice Abbott's New York*. Princeton Architectural Press, 2005.

[2] E. McNutty, *Boston Then and Now*. Thunder Bay Press, 1999.

[3] M. Klett, E. Manchester, and J. Verburg, *Second view: the rephotographic survey project*. University of New Mexico Press (Albuquerque), 1984.

[4] W. Fox, M. Klett, K. Banjakian, B. Wolfe, T. Ueshina, and M. Marshall, *Third View, Second Sights: a Rephotographic Survey of the American West*. Musuem of New Mexico Press, 2004.

[5] I. S. MacLaren, E. Higgs, and G. E. M. Zezulka-Mailloux, *Mapper of Mountains: MP Bridgland in the Canadian Rockies 1902-1930*. Edmonton: University of Alberta Press, 2005, p. 295.

[6] E. Higgs, *Nature by design: people, natural process, and ecological restoration*. MIT Press, 2003.

[7] W. Roush, "A substantial upward shift of the alpine treeline ecotone in the southern Canadian Rocky Mountains," *MSc Thesis*, pp. 1–175, Dec. 2009.

[8] J. Rhemtulla, R. Hall, E. Higgs, and S. Macdonald, "Eighty years of change: vegetation in the montane ecoregion of Jasper National Park, Alberta, Canada," *Canadian Journal of Forest Research-Revue Canadienne De Recherche Forestiere*, vol. 32, no. 11, pp. 2010–2021, 2002.

[9] D. LU, P. MAUSEL, E. Brondizio, and E. MORAN, "Change detection techniques," *International Journal of Remote Sensing*, vol. 25, no. 12, pp. 2365–2401, Jun. 2004.

[10] A. Singh, "Review Article Digital change detection techniques using remotely-sensed data," *International Journal of Remote Sensing*, vol. 10, no. 6, pp. 989–1003, 1989.

[11] C. Petit and E. F. Lambin, "Integration of multi-source remote sensing data for land cover change detection," *International Journal of Geographical Information ...*, 2001.

[12] J. Morgan and S. Gergel, "Aerial photography: a rapidly evolving tool for ecological management," *BioScience*, 2010.

[13] M. Bridgland, "Photographic Surveying," pp. 1–52, Apr. 1924.

[14] M. Sonka, V. Hlavac, and R. Boyle, *Image processing, analysis, and machine vision*, Third Edition. Thompson Learning, 2008.

[15] J. Wu and R. J. Hobbs, *Key topics in landscape ecology*. Cambridge University Press, 2007.

[16] S. Bae, A. Agarwala, and F. durand, "Computational rephotography," *Transactions on Graphics (TOG*, vol. 29, no. 3, Jun. 2010.

[17] D. G. Lowe, "Distinctive Image Features from Scale-Invariant Keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.

[18] N. Snavely, S. M. Seitz, and R. Szeliski, *Photo tourism: exploring photo collections in 3D*, vol. 25, no. 3. ACM SIGGRAPH 2006 Papers, 2006, pp. 835–846.

[19] G. Schindler, F. Dellaert, and S. B. Kang, "Inferring Temporal Order of Images From 3D Structure," in *Computer Vision and Pattern Recognition, 2007. CVPR '07. IEEE Conference on*, 2007, pp. 1–7.

[20] C. Schmid, R. Mohr, and C. Bauckhage, "Evaluation of Interest Point Detectors," *International Journal of Computer Vision*, vol. 37, no. 2, pp. 151–172, 2000.

[21] R. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*, Second. Cambridge University Press, ISBN: 0521540518, 2004.

[22] K. Mikolajczyk et al., "A Comparison of Affine Region Detectors," *International Journal of Computer Vision*, vol. 65, no. 1, pp. 43–72, 2005.

[23] K. Mikolajczyk and C. Schmid, "Scale & Affine Invariant Interest Point Detectors," pp. 1–24, Jul. 2004.

[24] C. Schmid, R. Mohr, and C. Bauckhage, "Evaluation of Interest Point Detectors," pp. 1–22, May. 2000.

[25] T. Kadir, A. Zisserman, and M. Brady, "An Affine Invariant Salient Region Detector," in *Computer Vision - ECCV 2004*, vol. 3021, no. 18, T. Pajdla and J. Matas, Eds. Berlin, Heidelberg: Springer Berlin / Heidelberg, 2004, pp. 228–241.

[26] A. Haja, B. Jahne, and S. Abraham, "Localization accuracy of region detectors," in *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, 2008, pp. 1–8.

[27] P. Moreels and P. Perona, "Evaluation of features detectors and descriptors based on 3D objects," *International Journal of Computer Vision*, vol. 73, no. 3, pp. 263–284, 2007.

[28] F. Fraundorfer and H. Bischof, "A novel performance evaluation method of local

detectors on non-planar scenes," in *Computer Vision and Pattern Recognition, 2004. CVPR 2004. Proceedings of the 2004 IEEE Computer Society Conference on*, 2005, pp. 33–33.

[29] C. Valgren and A. J. Lilienthal, "SIFT, SURF & seasons: Appearance-based long-term localization in outdoor environments," in *Robotics and Autonomous Systems*, 2010, vol. 58, no. 2, pp. 149–156.

[30] H. Bay, T. Tuytelaars, and L. Van Gool, "SURF: Speeded Up Robust Features," in *Computer Vision – ECCV 2006*, vol. 3951, A. Leonardis, H. Bischof, and A. Pinz, Eds. Springer Berlin / Heidelberg, 2006, pp. 404–417.

[31] T. Tuytelaars and K. Mikolajczyk, *Local Invariant Feature Detectors*. Now Pub, 2008, p. 124.

[32] N. Sebe, Q. Tian, E. Loupias, M. S. Lew, and T. S. Huang, "Evaluation of salient point techniques," *Image and Vision Computing*, vol. 21, no. 13, pp. 1087–1095, Dec. 2003.

[33] K. Mikolajczyk and C. Schmid, "A performance evaluation of local descriptors," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 27, no. 10, pp. 1615–1630, 2005.

[34] Y. Ke and R. Sukthankar, "PCA-SIFT: A more distinctive representation for local image descriptors," *Computer Vision and Pattern Recognition, 2004.*, vol. 2, 2004.

[35] K. Mikolajczyk and C. Schmid, "A performance evaluation of local descriptors," in *Computer Vision and Pattern Recognition, 2003. Proceedings. 2003 IEEE Computer Society Conference on*, 2003, vol. 2.

[36] K. Mikolajczyk and C. Schmid, "Indexing based on scale invariant interest points," in *Computer Vision, 2001. ICCV 2001. Proceedings. Eighth IEEE International Conference on*, 2001, vol. 1, pp. 525–531.

[37] A. Gil, O. Mozos, M. Ballesta, and O. Reinoso, "A comparative evaluation of interest point detectors and local descriptors for visual SLAM," *Machine Vision and Applications*, vol. 21, no. 6, pp. 905–920, Jan. 2010.

[38] G. Carneiro and A. Jepson, "Phase-Based Local Features," in *Lecture Notes in Computer Science*, vol. 2350, no. 19, A. Heyden, G. Sparr, M. Nielsen, and P. Johansen, Eds. Berlin, Heidelberg: Springer Berlin / Heidelberg SN -, 2002, pp. 282–296.

[39] *robots.ox.ac.uk*. [Online]. Available: http://www.robots.ox.ac.uk. [Accessed: 02-May-2011].

[40] A. Vedaldi, Ed. *vlfeat.org*, *vlfeat.org*. [Online]. Available: http://www.vlfeat.org.

[Accessed: 02-May-2011].

[41] J. Matas, O. Chum, M. Urban, and T. Pajdla, “Robust wide-baseline stereo from maximally stable extremal regions,” *Image and Vision Computing*, vol. 22, no. 10, pp. 761–767, Jul. 2002.

[42] “Featurespace,” *featurespace.org*.

[43] T. Lindeberg, “Scale-space theory: A basic tool for analysing structures at different scales,” pp. 1–50, Feb. 2011.

[44] J. Koenderink, “Representation of local geometry in the visual system,” *Biological cybernetics*, 1987.

[45] A. Baumberg, “Reliable feature matching across widely separated views,” in *Computer Vision and Pattern Recognition, 2000. Proceedings. IEEE Conference on*, 2000, vol. 1, pp. 774–781.

[46] Y. Dufournaud, C. Schmid, and R. Horaud, “Matching images with different resolutions,” in *Computer Vision and Pattern Recognition, 2000. Proceedings. IEEE Conference on*, 2000, vol. 1, pp. 612–618.

[47] T. Lindeberg, “Feature Detection with Automatic Scale Selection,” *International Journal of Computer Vision*, vol. 30, no. 2, pp. 79–116, Jan. 1998.

[48] K. Mikolajczyk, “Detection of local features invariant to affine transformations,” *PhD Thesis*, 2002.

[49] J. L. Crowley and A. C. Parker, “A Representation for Shape Based on Peaks and Ridges in the Difference of Low-Pass Transform,” *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, no. 2, pp. 156–170, 1984.

[50] S. Grossberg, E. Mingolla, and D. Todorovic, “A neural network architecture for preattentive vision,” *Biomedical Engineering, IEEE Transactions on*, vol. 36, no. 1, pp. 65–84, 1989.

[51] D. G. Lowe, “Object recognition from local scale-invariant features,” in *Computer Vision, 1999. The Proceedings of the Seventh IEEE International Conference on*, 1999, vol. 2, pp. 1150–1157 vol.2.

[52] S. Belongie, J. Malik, and J. Puzicha, “Shape matching and object recognition using shape contexts,” *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 24, no. 4, pp. 509–522, 2002.

[53] J. Canny, “A Computational Approach to Edge Detection,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 8, no. 6, pp. 679–698, 1986.

[54] M. A. Fischler and R. C. Bolles, “Random sample consensus: a paradigm for

model fitting with applications to image analysis and automated cartography," *Commun. ACM*, vol. 24, pp. 381–395, Jun. 1981.

[55] P. Torr and A. Zisserman, "MLESAC: A new robust estimator with application to estimating image geometry," *Computer Vision and Image Understanding*, vol. 78, no. 1, pp. 138–156, 2000.

[56] M. Zuliani, "RANSAC for Dummies," *With examples using the RANSAC toolbox for Matlab*