Feature detection in RGB and NIR images

Damien Firmenich
EPFL
damien.firmenich@epfl.ch
Supervisor: Matthew Brown

Abstract

This project intends to evaluate feature detection performance and stability amongst common feature detectors: Harris Corner Detector, Difference of Gaussians (DoG), and Maximally Stable Extremal Regions (MSERs), when using RGB as well as near-infrared (NIR) images to perform registration. To improve the feature detection efficiency, the standard detectors will be extended to a multispectral (RGB + NIR) version, which shows a significant improvement over single-band detection. Matching detected features between RGB and NIR images is also more compromised because of the different signal response, so a method is proposed to improve the matching performance, which makes use of the contrast inversion between the two types of images occurring in specific regions.

1 Introduction

Feature detection has been used for many purposes such as image registration, object and scene recognition, video tracking. Often, the use of visible light only (RGB) is sufficient for good feature detection performances, but in specific cases it fails to deliver optimal results. For example, a panoramic photo sequence containing heavy haze may be impossible to stitch using only visible light. The use of near-infrared images is aimed to improve detection in those cases, and completely replace visible light when needed. For example, Microsoft's recent Kinect system uses NIR light to compute a depth map of a scene. Near-infrared imaging can be done using a standard silicon based digital camera, and thus can be used rapidly and cheaply in a passive way.

The introduction of combined RGB and NIR feature detection methods has many applications, mainly the use of multi-sensor cameras such as jAi's camera (Figure 1) which captures both color and near-infrared at the same time. With multispectral images, there is extended possibilities when used for photography, leading to easier panoramic stitching, or object tracking, where near-infrared markers can be placed on an object and tracked without it being visible by the human eye.

Figure 1: jAi multispectral camera

Features detected in color images can also be matched to those detected in near-infrared images to perform registration within them. Project Thalis aims to solve the problem of automated plane landing using ground images to compute the appropriate location and direction of the plane.

2 Standard methods

Image registration requires a combination of various techniques in order to achieve accurate and robust registration. This project studies feature-based image registration which makes use of feature extraction algorithms combined with matching and motion estimate methods.

The three main steps involved in image registration are feature detection, which finds good features in the images, feature matching, which makes correspondences within the detected features, and motion estimate, which finds a transformation model to produce accurate alignment of the images.

2.1 Feature Detection

Harris Corner Detector The Harris operator [2] provides a fast and simple way to accurately detect corners. A corner is a good feature to extract because it has an orientation.

The basic concept behind corner detection is that the overall intensity of a small window placed on a

corner has large intensity variation when shifted in every direction. If the variation occurs only in one direction, the window is placed on an edge, and if there is little or no variation, then the surface is flat, see Figure 2. A corner can thus be detected by the smallest shift that produces a large variation.

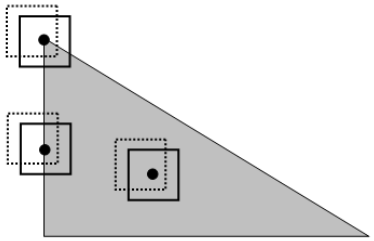


Figure 2: Corner detection

This concept can be modeled by the autocorrelation matrix of a window patch:

$$A = \begin{pmatrix} I_x^2 & I_x I_y \\ I_y I_x & I_y^2 \end{pmatrix}$$

with each components computed using a circular weighted window such as a Gaussian. A strength score is then assigned to each coordinates using a measure of the principal curvatures of this matrix, which are proportional to its eigenvalues $\lambda_{1,2}$. The score is defined by the following formula:

$$s = \lambda_1 \lambda_2 - k\left(\lambda_1 + \lambda_2\right)^2 = \det(A) - k \cdot \text{trace}^2(A)$$

in which the determinant and trace are used to efficiently compare the eigenvalues. The factor k is an arbitrary number, usually set to 0.04. Corners are determined by an 8-way local maxima of the strengths.

Difference of Gaussians The blob detection approach used by the Difference of Gaussians (DoG) method, as described by [4] for the Scale Invariant Feature Transform (SIFT) algorithm, detects stable keypoints locations in scale-space by convolving the image with a difference-of-Gaussians function:

$$D(x, y, \sigma) = (G(x, y, k\sigma) - G(x, y, \sigma)) * I(x, y)$$

which is equivalent as subtracting two images taken at different scale (see Figure 3). This function is a close approximation of the scale-normalized Laplacian of Gaussian $\sigma^2 \nabla^2 G$ from [3] which, by finding its local minima and extrema, produces very stable image features [7]. In the case of the DoG, the local extremas are found by comparing a value in scale-space to its 26 neighbors (3x3x3 region in adjacent scales, excluding the current value).

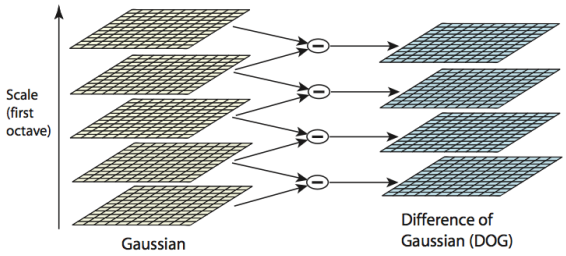


Figure 3: Difference of Gaussians (illustration from [4])

The extremas found within the DoG function are filtered to avoid features with low contrast (using a simple threshold) or features that lies on an edge (using ratios of the DoG's Hessian matrix eigenvalues) which can ben unstable across different images.

MSERs The Maximally stable extremal regions (MSERs) method, proposed by [5], is especially useful for stereo matching because of its affine and scale invariance. It detects regions in the image which have a globally stable intensity, i.e. a region with low variation of intensity within its boundaries.

This concept can be seen as thresholding the intensity level of the image. Extremal regions are connected components that have an intensity higher or lower than the threshold, and stable regions are those whose size doesn't change greatly across multiple threshold levels. A measure is thus assigned to each MSER, defining its stability: $v\left(R, \Delta\right) = \frac{|R_{+\Delta}| - |R_{-\Delta}|}{|R|}$.

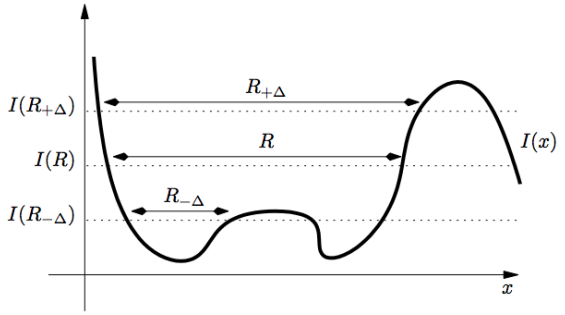


Figure 4: MSERs level thresholding (illustration from [11])

The main advantages of using MSERs for blob detection are scale and affine invariance, stability and efficiency. The algorithm, unlike DoG, doesn't involve blurring the image for multi-scale detection, which is key for efficiency.

SIFT descriptor The previously detected features need to be uniquely identifiable in order to make correspondence between matching features of multiple images. The SIFT descriptor uses a method similar to how the biological visual cortex represents features. It evaluates image gradients within a region around the detected keypoint to make a histogram of orientations and magnitudes (see Figure 5).

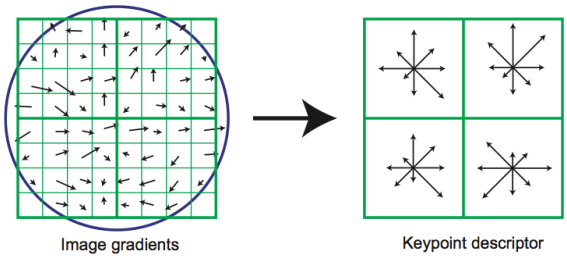


Figure 5: SIFT descriptor histogram (illustration from [4])

To achieve scale and rotation invariance, the descriptor need to be scaled and rotated according to keypoints properties. The SIFT descriptor uses frames defined by 4 parameters (x, y, σ, θ) which determines its location (x, y), its scale σ and its orientation θ.

2.2 Matching

Extracted features in one image need to be matched against the features in another image, and descriptors computed above are used for this purpose. As described by [4], matching can be done by simply finding the closest neighbor of a descriptor. False positives and false negatives need to be discarded for accurate motion estimation, so the minimum distance found between two descriptors is compared to the distance of the second-closest neighbor, and the match is kept only if their ratio is lower than 0.8.

2.3 Motion Estimate

Once the matches are found within the feature sets, a transformation model needs to be found to produce accurate alignment. A common and robust solution for this problem is the Random Sample Consensus (RANSAC).

It starts by selecting a random sample of correspondences used to compute a transformation estimate. Then all the other matches are verified, and a set of inliers is retrieved, defined by the matches whose distance after transformation is lower than an error margin. The RANSAC algorithm reproduces these operations for a fixed number of iterations, and keeps the transformation estimate that has the largest number of inliers.

The probability of finding a good transformation model depends on the efficiency of the extraction and matching method. The number of iterations needed might vary from one set of matches to another but, by estimating the success probability of the matching and the size of the sample, an estimation of this number can be computed to have a good probability of success.

3 Multispectral feature detection

The combined use of near-infrared (NIR) and RGB images for registration requires some modification of the previously described techniques. The main issue to address is the need to use multiple band to extract features. In this project, some images are composed of RGB and NIR bands, so the feature extraction methods need to take into account the information contained in both layers. The new extended methods are described in the following.

Implementation of those methods were done partly using the VLFeat library [12].

Harris corner extension The implementation of the Harris operator used in this project differs on some levels from the original. To solve the problem of multispectral detection, the autocorrelation matrix is computed for each band, and the Harris corner strengths are calculated on the sum of the matrices.

$$A = \sum_i A_i = \begin{pmatrix} \sum_i I_{ix}^2 & \sum_i I_{ix} I_{iy} \\ \sum_i I_{iy} I_{ix} & \sum_i I_{iy}^2 \end{pmatrix}$$

This way of combining the two channels ensures that the most relevant features will be detected, either in NIR or RGB, depending on the strength of the corner.

Another modification of the Harris operator was done in order to achieve scale-invariance detection. A scale-space approach similar to the one used for the Difference of Gaussians was applied, where the strengths are computed at varying scale t_i defined by $t_i = 1.6 \cdot 2^{\frac{i}{3}}$ for $i = 0, \ldots, n-1$ with n being the number of scales. The resulting strengths are contained within a 3-dimensional scale-space and the local maxima are located using a 3x3x3 region.

The strength measure was also modified using a variation from [8] to avoid using an arbitrary k factor: $s = \frac{\det A}{\text{trace} A}$.

Difference of Gaussians extension The multispectral extension of the DoG method was done by computing the differences for each band, and by using the magnitude of the vector defined by each of the differences as the final DoG value for the current location in scale-space:

$$D(x, y, \sigma) = \left\| \begin{matrix} D_1(x, y, \sigma) \\ \vdots \\ D_n(x, y, \sigma) \end{matrix} \right\|$$

with $D_i(x, y, \sigma) = (G(x, y, k\sigma) - G(x, y, \sigma)) * I_i(x, y)$ the value for the band i.

By taking a vector of those differences, we ensure that the resulting DoG won't be affected by negative values, because only its magnitude is kept.

MSERs extension The MSERs method was extended in a straightforward manner, as the stable regions detected in every band is kept for matching. One possible alternative might be to evaluate the stability score for each region and only take those above a certain threshold, or by taking the local maxima between the bands.

To assign a scale and an orientation to the region, as needed by the SIFT descriptor, the properties of the covariance matrix defining the region's ellipse were used. The angle of the matrix's principal component, discretised in 36 bins, defines the region orientation, and the corresponding eigenvalue, discretised in n bins (n being the number of scales), is stored as its scale. The discretisation allows similar regions to have exactly the same properties to improve matching efficiency. As described below, this orientation assignment is similar to the one used by the SIFT descriptor.

SIFT descriptor extension After detecting the features across multiple bands, the description method need to be extended to a multispectral image too.

As mentioned above, the SIFT descriptor uses frames to compute a histogram for each feature. The frames automatically contains the location (x, y) and the scale σ information from the detection step, and its orientation is defined by the maximums values in a 36-bins histogram of the gradients orientations weighted by their magnitude in a region around the keypoint. For a multispectral extension of the frames construction, the histogram orientations are evaluated across every bands in a 3-dimensional region, so the dominant orientations within all the bands are kept.

Using the frame properties, the descriptor histogram is computed for each band and the final descriptor will be a concatenation of those. The matching will only compute the distances between them, so the way the histograms are concatenated doesn't affect the performance, as long as it stays consistent across images.

4 Improvements

Near-infrared (NIR) light as a different response than visible light because of its different wavelength range. Visible light has wavelengths that range from 350nm to 700nm, but NIR light goes from 700nm to approx. 1100nm (depending on the applications). The light distribution captured with a NIR capable device has different properties. Some components visible to the human eye can be discarded in NIR such as colors, and other less visible elements can appear very clearly. These differences can be used for image registration to improve matching efficiency. More specifically, the information present in NIR images can complement RGB images to have a greater chance of feature matching success.

Gradient direction invariant SIFT One of the main difference we can notice from an NIR image compared to an RGB is the change of contrast between scene elements. Some of them appear dull in RGB but very bright in NIR. An example of this phenomenon is foliage, which is clear as snow in an NIR image, probably from the photosynthesis process as [6] describes. A consequence of this is that some contrasts in the image can be inverted (foliage against gray structure, see Figure 6), causing reversal of image gradients directions.

The histogram in Figure 7a shows how often this phenomenon occurs. The dot product of pixel gradients from various pairs of RGB-NIR images were computed , and the histogram shows the amount of occurrences for each result, with left pixels having inverted gradients. Occurrences are shown in Figure 7b, where we can see that they are located mainly around foliage.

This particular aspect can cause mismatch when trying to register an RGB to an NIR image because the SIFT descriptor depends heavily on gradients orientations to describe features. Although only a minority of gradients are reversed, it is worth investigating how this phenomenon alters the matching efficiency, and if some improvement could correct the loss of efficiency.

A technique to avoid this kind of inconsistencies would be to equalize the gradient of the image so their angle θ ranges the upper half circle, by rotating the reversed gradients by 180° (Figure 8).

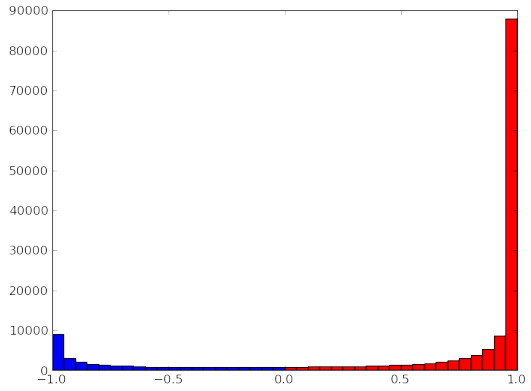
(a) Luminance channel

(b) NIR channel

(c) Luminance channel

(d) NIR channel

Figure 6: Inverted contrast in foliage scenes



(a) Reversed gradients occurrences distribution

(b) Locations of reversed gradients

Figure 7: Reversed gradients

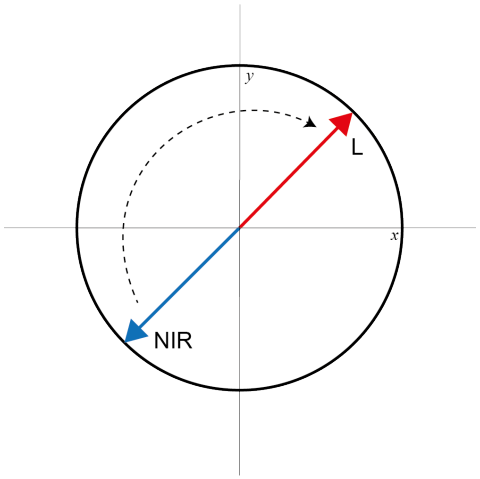


Figure 8: Gradient reversal

$$\theta = \begin{cases} \theta & \text{if } \theta \in [0, \pi[\\ \theta - \pi & \text{if } \theta \in [\pi, 2\pi[\end{cases}$$

This technique allows the SIFT descriptor to be invariant to gradient direction, causing the features lying on an inverted contrast region to be correctly matched. It does however affect the description of normal features (i.e. without contrast inversion) since correct gradients may be reversed, so it can lead to ambiguous matching. But the results of experimentations shows that the global matching performance is still better with equalization.

Visible light correlation Even though the NIR wavelengths size are close to visible light's, there is low correlation between color and NIR response, i.e. as described in [1], we have almost no information on the NIR response given the RGB response. The shape of an object is distinguishable from an NIR signal mainly from the high frequencies which are preserved (see Figure 9). Thus it is hard to adapt the detector based on the color distribution only. It depends more on the material of the object, so a possible way to predict the NIR response and adapt the detectors would be to experiment with different type of material and investigate on how the response differs from a visible light signal.

5 Evaluation

In order to evaluate the performance of previously proposed extensions and improvements of feature extraction methods, an evaluation method has been developed. The main goal is to measure the match-

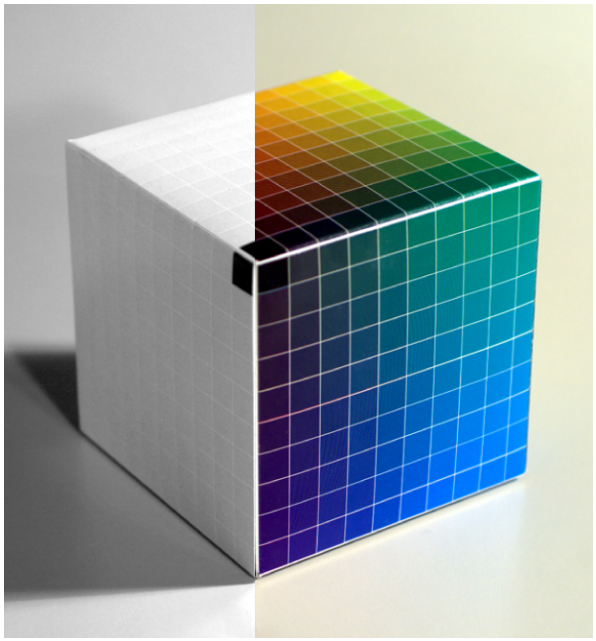
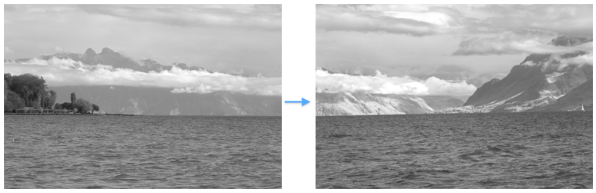
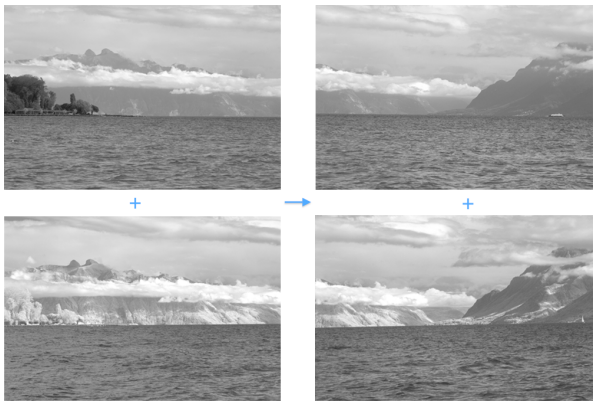


Figure 9: Left: Cube showing low correlation between NIR (left) and visible light (right) (illustration from [1])



(a) RGB (L) to NIR



(b) Multispectral to multispectral

Figure 10: Test cases

ing performance of those methods in different cases.

For each of the feature detection method, namely Harris, DoG and MSERs, two cases were investigated in this project:

- RGB (luminance channel only) to NIR registration (Figure 10a)
- multispectral to multispectral (combined luminance and NIR) registration (Figure 10b)

Comparisons will be made with both RGB to RGB and NIR to NIR to see how the improvements perform, and which advantages they give us.

5.1 Picture database

The acquisition of RGB and NIR images was done using a standard D-SLR camera. The sensors of digital cameras are capable of capturing NIR signal by default, and thus has a filter to prevent it from RGB capture. The camera used was modified in order to be able to capture the whole spectrum. RGB and NIR images were taken using an NIR-blocking and visible-blocking filter respectively.

Various types of scenes were photographed, and the resulting database contains sets of landscape panoramic sequences, captured both in RGB and NIR. Pictures in the sequence were taken with approx. 30% of overlapping area between them. This type of sequence allows to evaluate each cases independently, namely RGB to RGB, RGB to NIR, NIR to NIR, and multispectral to multispectral matching, and thus allows a good range of testing possibilities.

The following evaluations will be made on 4 sets of panoramic sequences, as previewed in Figure 11.

5.2 Method

The stability of feature extraction methods is evaluated by examining how the matching performs between two overlapping images using their sets of extracted features. In this project, the following workflow was used.

Ground truth To have accurate measures, the tests were computed using ground truth registration transform. This transform was found either by using the standard techniques of image registration as described in Section 2 (for most of the cases) or by manually selecting control points and computing the transformation from them (for a minor quantity of failed cases).

Feature selection The measurements were done for a varying number of detected features, so they have to be selected in a specific manner in order to have a consistent evaluation.

For each of the measurements, a fixed amount of features were selected, and then increased for the next measurement. The problem is to find a way to select a fixed number of features based on

(a) Mountain

(b) City

(c) Lake

(d) Forest

Figure 11: Panoramic scenes

their properties. For those evaluations, the features were selected by largest scale by sorting them in decreasing scale order, and selecting the first n of them, with n going from 100 to 2000.

Selecting features by increasing scale gives us an interesting point of view, as the matching is first evaluated for large scale feature, i.e. generic features that can be detected from a small image size, and then evaluated for more detailed and specific features.

Stability measurement Once the features are selected, their stability needs to be tested. First, matches are found between the two subsets of features (from both images). Then the correct matches are computed using the previously found registration transformation by finding which of the matches are consistent within an error of 3 pixels. The evaluated measure is the ratio between the correct matches and the matches lying in the overlapping area of both images, which gives the percentage of correct matches in this area.

$$r = \frac{\#\text{ of correct matches}}{\#\text{ of matches in overlap area}}$$

The matches are only taken in the overlap area of the two images because the evaluation has to be consistent for all types of registration. The overlap area isn't constant, so the ratio would be much lower for images with only 30% of overlap compared to those with 90%. We are only interest in potentially correct matches, and matches not lying in the overlap area are obviously false.

This evaluation method allows the measurement of the stability of feature detection and feature description methods. The testing suite includes extended feature detection methods (Harris corners, DoG, MSERs), both standard and improved versions, and the two mentioned registration cases (luminance to NIR, multispectral to multispectral).

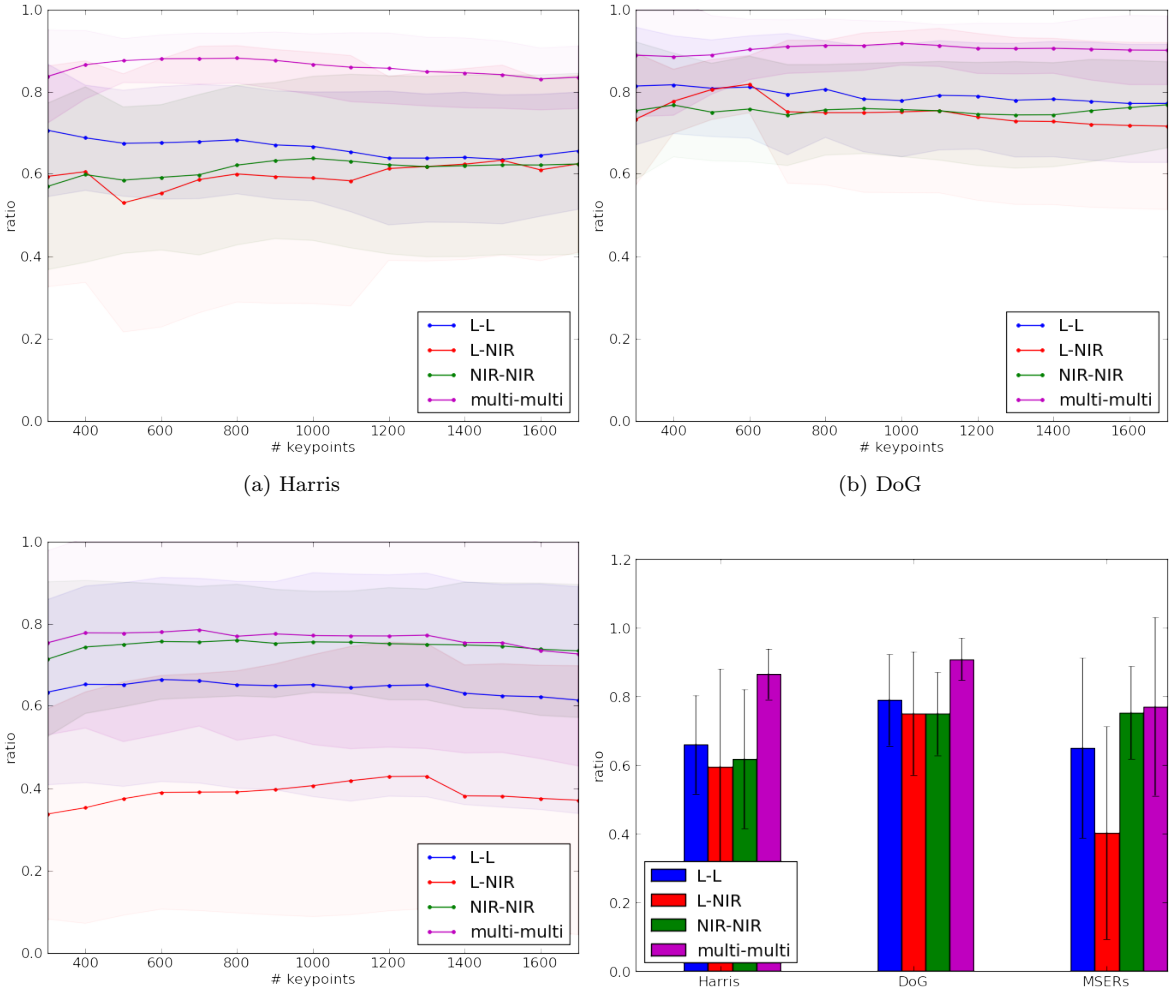
6 Experimental Results

6.1 Multispectral detection

In order to measure the stability of the extended multispectral (combined luminance and NIR channels, L+NIR) feature detectors, a comparison with single band (L, NIR) were made using the previously described evaluation framework.

The graphs in Figure 12 shows an evaluation for the three detectors (Harris, DoG, MSERs) for the multispectral extensions and for standard methods (the histogram values were computed using a range of 500-1500 features). Registration using luminance channel only has a performance comparable to NIR-NIR registration, except for the MSERs detector. However, it performs globally better than L-NIR registration which has obviously the worst recognition rate. By using the multispectral versions of the detectors, we are able to significantly improve the efficiency compared to L-L.

Multispectral MSERs improved the performance compared to L-L and L-NIR, but it wasn't significant compared to NIR-NIR. The technique was to keep all the detected regions for matching, so it may not be as optimized as the two other ones. It seems that the real improvement comes from a combination of both L and NIR channels.



(a) Harris (b) DoG

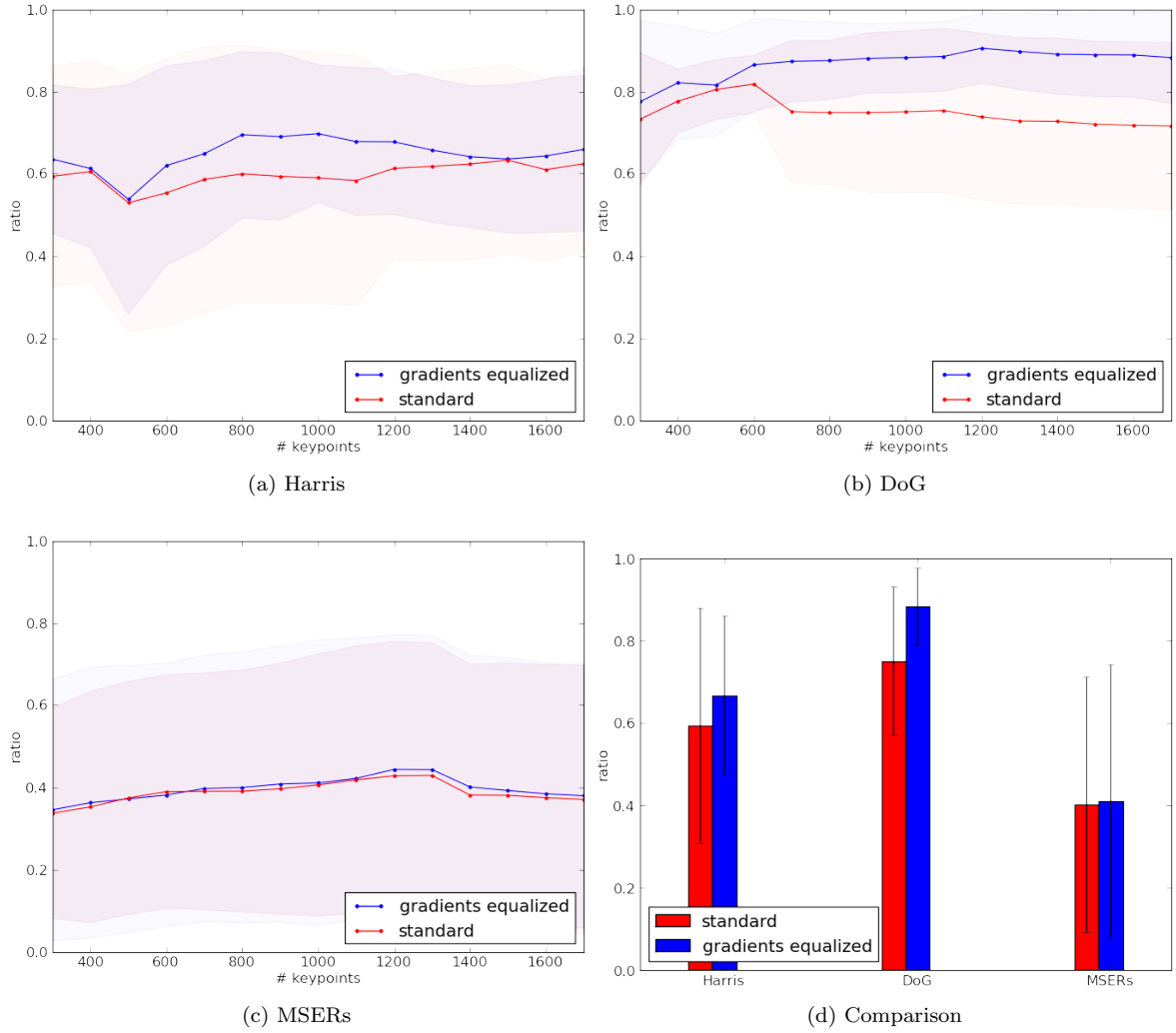
(c) MSERs (d) Comparison

Figure 12: Multispectral extension compared to standard methods

Using the Difference of Gaussians approach leads to better results overall. It has low difference of efficiency between each of the three cases because of the standard stability of the DoG keypoints. There is however an improvement when using the multispectral approach, especially for scenes with lower details in RGB than in NIR such as hazy landscapes. The Harris detector isn't efficient when matching L-NIR compared to L-L, but has a large improvement when the multispectral extension was used.

The low overall performance of MSERs for L-NIR is probably due to the fact that the features detected with this detector are described by the SIFT descriptor. It uses the gradients orientations around the feature to compute a histogram and, as MSERs detected features are essentially at the center of areas with low varying intensity, SIFT may not be able to compute a unique descriptor. The information around the feature is not as rich as for the Harris and DoG features, which are essentially composed of edges or corners, or areas with large variation on the sides.

Those results were obtained using combined L and NIR images, and the efficiency of the extended extraction methods using those images relies on the actual registration degree between the two channels. The feature detection step needs some degree of region similarities between them and, mainly for small scale features, if a region doesn't contained aligned potential features, those won't be detected. So, depending on the detection scale, it is good to have nicely registered channels. Some of the images used for those evaluation contains minor registration issues because of a parallax effect due to small change of viewpoint between the RGB and NIR capture, causing altered performance. However, considering those issues, the multispectral de-



(a) Harris (b) DoG

(c) MSERs (d) Comparison

Figure 13: Gradient direction invariant SIFT compared to standard descriptor for L-to-NIR matching

tectors performs better than standard versions because it detects mainly features unaffected by the parallax.

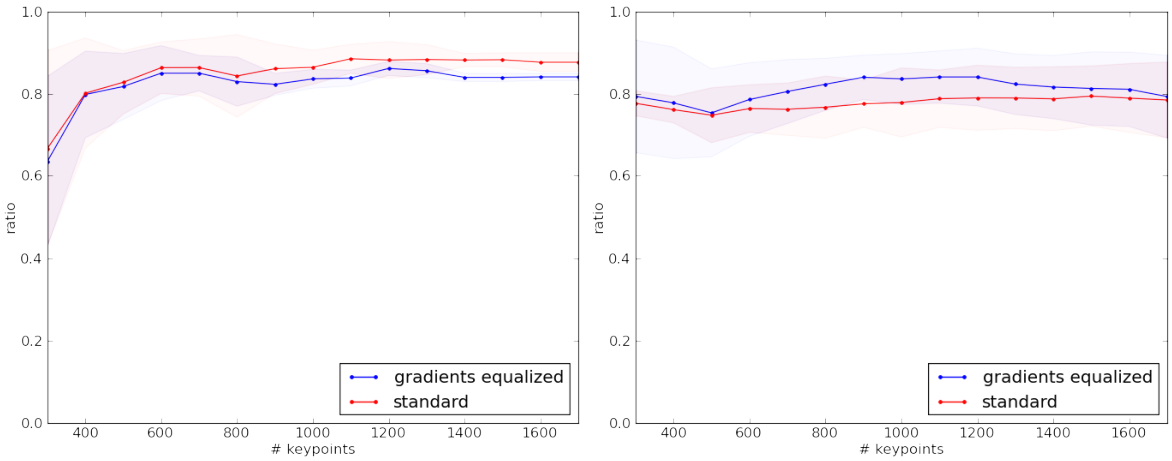
One way to resolve this issue is by using a special camera (e.g. jAi cameras), equipped with two sensors, which captures both RGB and NIR signals at the same time and thus produces perfectly registered images. Using those images would lead to higher performances, without needing to register them before.

6.2 Gradient direction invariant descriptor

As described in Section 4, one possible improvement for the stability of the detectors when matching RGB (luminance channel, L) to NIR images was to equalize the gradients orientation to discard inconsistencies when inverted contrasts occurs. For this purpose, evaluation was made for L-to-NIR matching for the three detectors, with and without gradient equalization.

Figure 13 shows the results for all three detectors separately (the histogram values were computed using a range of 500-1500 features). We can see an improvement when the gradient direction invariant SIFT descriptor is used. The Harris detector has an improvement for a mid-range number of selected features, and converges to the same performance as the standard detector when more features are used. The improvement for the DoG detector is noticeable from approx. 600 features, and rises from there. This result proves that the inverted contrast occurrences does affects the stability of feature extraction, and can be partly improved by discarding those inconsistencies using this method.

The MSERs detector shows no improvement from the standard method, and this may be due



(a) Mountain sequence

(b) City sequence

Figure 14: Comparison of Gradient direction invariant SIFT in normal conditions

to the SIFT descriptor not being optimized for this detector, as explained above.

Efficiency at normal conditions It is important to note that by equalizing the gradients, the description of normal features is affected. This can cause the matching to perform a little worse than usual, because two non-similar feature can be seen as similar if gradients are reversed, and the matching will discard both of them. This means that the gradient direction invariant SIFT must be used only if there are significant amount of inverted contrast occurrences.

For example, in Figure 14a, matching was performed with DoG for the mountain sequence where the occurrences of inverted contrast are very limited (foliage is scarce), and we can see that the stability of features are in fact decreased when using gradient equalization. However, in Figure 14, the same evaluation was done, but only on the city sequence, where there is much more foliage (and thus, inverted contrasts), and the result shows an improvement when using gradient equalization.

7 Open problems and future work

Some unresolved issue encountered in this project was the detection of inverted contrast between RGB and NIR. An approach proposed in this project was to use the gradient direction invariant SIFT whenever RGB and NIR images have to be registerd. But if the gradients can be equalized locally, only for the occurrences of inverted contrasts, the matching performance would be greater when registering those type of images.

As described above there is a risk of parallax-related registration imprecisions when preparing the RGB and NIR images for the combined registration. It occurs when the exposures weren't taken with a tripod, or when there was a signification time difference between them, altering the registration performance. Further testing can be done using a multi-sensor camera as well, avoiding those imprecisions.

Future work might also include finding improvements for RGB to NIR registration based on other response differences, possibly related to the material of objects in the scene. Different objects could be detected using an object/scene recognition technique, and the feature detection could be optimized per object. This could also help to decide wether gradient equalization would help, in a region-specific manner rather than global. Some modifications of the multispectral extensions could also be made, based on the same channel differences, to possible improve their efficiency, again in a more specific manner.

8 Conclusion

This project involved the investigation of near-infrared and RGB feature detection with the evaluation of three standard methods (Harris, DoG, MSERs) when confronted to RGB/NIR image registration. The results showed an obvious loss of performance when registering RGB to NIR images. An improvement was introduced, which makes use of the dissimilarities between RGB and NIR im-

ages, in this case the contrast inversion occurrences. The evaluation showed higher results, but only when this phenomenon occurs frequently enough. New feature detectors based on the standard detection methods were proposed, which makes use of both luminance and NIR channels. The evaluation showed that there is indeed a way to combine the two channels in an efficient way, as the multispectral methods showed higher results.

References

[1] Clément Fredembach. Near-infrared imaging, http://ivrg.epfl.ch/research/topics/nir.html.

[2] Chris Harris and Mike Stephens. A combined corner and edge detector. 1988.

[3] Tony Lindeberg. Scale-space theory: A basic tool for analysing structures at different scales. *Journal of Applied Statistics*, 1994.

[4] David G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision,*, 2004.

[5] J. Matas, O. Chum, M. Urban, and T. Pajdla. Robust wide baseline stereo from maximally stable extremal regions. *BMVC*, 2002.

[6] R. J. Mayer. The near-infrared fluorescence of green leaves. *Infrared Physics*, 1965.

[7] Krystian Mikolajczyk. *Detection of local features invariant to affine transformations.* PhD thesis, Institut National Polytechnique de Grenoble, 2002.

[8] J. Alison Noble. Finding corners. *Image Vision Comput.*, 6:121–128, May 1988.

[9] Lindsay I Smith. *A tutorial on Principal Components Analysis*, 2002.

[10] Richard Szeliski. *Image Alignment and Stitching: A Tutorial*, 2006.

[11] Andrea Vedaldi. *An implementation of Multi-Dimensional Maximally Stable Extremal Regions*, 2007.

[12] Andrea Vedaldi and Brian Fulkerson. Vlfeat, www.vlfeat.org.