A NEW INVARIANT IMAGE DESCRIPTOR FOR INDEXING AND RECOGNITION: FIRST RESULTS

BY

EVANTHIA MAVRIDOU

THESIS

Submitted in partial fulfilment of the requirements
for the degree of Master of Science in Informatics
at
University of Crete, Greece 2011
and
University Joseph Fourier, Grenoble 2011

Supervisor

James Crowley, Director: PRIMA-INRIA Rhône Alpes

Grenoble, June 2011

© 2011 by Evanthia Mavridou. All rights reserved.

UNIVERSITY OF CRETE

UNIVERSITY JOSEPH FOURIER

PRIMA-INRIA Rhône Alpes

A NEW INVARIANT IMAGE DESCRIPTOR FOR INDEXING AND RECOGNITION: FIRST RESULTS

A thesis submitted by

EVANTHIA MAVRIDOU

in partial fulfilment of the requirements for the degree of Master of Science in Informatics

Supervisor: James Crowley, Director: PRIMA-INRIA Rhône Alpes

Author: Evanthia Mavridou

Accepted by: Ioannis Stylianou

UNIVERSITY OF CRETE

UNIVERSITY JOSEPH FOURIER

PRIMA-INRIA Rhône Alpes

A NEW INVARIANT IMAGE DESCRIPTOR FOR INDEXING AND RECOGNITION: FIRST RESULTS

A thesis submitted by

EVANTHIA MAVRIDOU

in partial fulfilment of the requirements for the degree of Master of Science in Informatics

Supervisor: James Crowley, Director: PRIMA-INRIA Rhône Alpes

Committee:

Περίληψη

Η περιγραφή του περιεχομένου εικόνων ειναι η διαδικασία η οποία συλλέγει την πληροφορία που υπάρχει στα pixels της εικόνας και την εκφράζει με τέτοιο τρόπο ώστε να μπορεί να χρησιμοποιηθεί για ταίριασμα, ανίχνευση και αναγνώριση αντικειμένων σε εικόνες. Η περιγραφή είναι μια βασική διαδικασία σε όλους τους τομείς της υπολογιστικής όρασης. Τα αποτελέσματα στη διαδικασία της όρασης μπορούν να βελτιωθούν σημαντικά εαν η μέθοδος της περιγραφής μπορει να παραμένει αναλλοίωτη σε αλλαγές του περιεχομένου της εικόνας. Σε αυτην την εργασία, προτείνεται μια νέα μέθοδος περιγραφής εικόνων που παραμένει αναλλόιωτη σε αλλαγές της εικόνας και βασίζεται σε Λαπλασιανά προφίλς και ακτινικό μετασχηματισμό Fourier. Το Λαπλασιανό προφίλ ενός pixel ειναι ενα διάνυσμα Λαπλασιανών τιμών σε διαφορετικές κλιμακώσεις τις εικόνας στις ανάλογες συντεταγμένες. Η χρήση του ακτινικού μετασχηματισμού Fourier σε μικρές γειτονιές γύρω από το Λαπλασιανό προφίλ προσφέρει σε ένα περιγραφέα την ικανότητα να παραμένει αναλλοίωτος στη μετατόπιση, περιστροφή και κλιμάκωση της εικόνας. Ο υπολογισμός των Λαπλασιανών τιμών σε διαφορετικές κλιμακώσεις της εικόνας είναι εύκολα εφικτός με τη χρήση του αλγορίθμου Γκαουσιανή Πυραμίδα Μισής Οκτάβας. Το μέγεθος του περιγραφέα μπορεί να μεταβάλλεται με τη χρήση διαφορετικού αριθμού επίπεδων από την πυραμίδα. Τα πειράματα αποδεικνύουν την ικανότητα της προτεινόμενης μεθοδου να παραμενει αμεταβλητη (στην μετατόπιση, την περιστροφή και τη κλιμάκωση) και να περιγράφει με παρόμοιο τρόπο τα ίδια αντίκειμενα σε διαφορετικές εικόνες. Τα πειράματα δείχνουν επίσης ότι όσο μεγαλύτερο ειναι το μέγεθος του Λαπλασιανού προφίλ που χρησιμοποιείται για την κατασκευή ενός περιγραφέα, άρα όσο μεγαλύτερο το μέγεθος του περιγραφέα, τόσο καλύτερα αποτελέσματα παίρνουμε. Η νέα μέθοδος περιγραφής εικόνων φαίνεται οτι είναι κατάλληλη για την περίπτωση που πρέπει να γίνει ανίχνευση και αναγνώριση διαφημισ-τικών λογότυπων σε εικόνες η εντοπισμός και αναγνώριση διαφημιστικών λογότυπων σε ακολουθίες εικόνων που λαμβάνονται απο κάμερες σε κινητά τηλέφωνα.

Λέξεις - Κλειδιά:

περιγραφή περιεχομένου εικόνων, γκαουσσιανή πυραμίδα μισής οκτάβας, λαπλασιανό προφίλ, ακτινικός μετασχηματισμός Fourier, αμετάβλητη στις αλλαγές, αναγνώριση

Résumé

Le processus de description du contenu d'une image consiste à extraire l'information représentée par l'ensemble des pixels de l'image et l'exprime de sorte qu'elle soit utilisable pour l'appariement, la détection and la reconnaissance. La description est une procedure de base dans tous les domaines de la vision par ordinateur. Dans les processus de vision, les résultats peuvent être significativement améliorés si le modèle de description est invariant par rapport variations du contenu de l'image. Dans ce projet, une nouvelle méthode de description invariante est proposée; elle est basée sur des profils de Laplaciens et la transformée de Fourier radiale. Le profil de Laplaciens d'un pixel d'une image est une séquence des valeurs de Laplaciens à différentes échelles de l'image à ce même pixel. L'ajout de la transformée de Fourier Radiale aux valeurs de Laplaciens dans un petit voisinage autour du profil de Laplaciens forme un descripteur invariant en translation, rotation et homothétie. Le calcul des valeurs de Laplaciens à différentes échelles est facilement réalisable en utilisant la Pyramide Gaussienne Demi Octave. La taille du descripteur peut varier en changeant le nombre de niveaux utilisés dans la pyramide. Les experiences attestent la capacité de la méthode proposée à être invariante (par rapport translation, rotation et homothétie) et à être discriminante. Elles montrent aussi que, plus le profil de Laplacien est grand, donc plus la taille du descripteur est grande, et meilleurs seront les résultats. La nouvelle méthode de description apparait être adaptée pour détecter et reconnaître des logos publicitaires sur des images ou encore le suivi de logos publicitaires dans des séquences d'images prises par des caméras de téléphone portable.

Mot - clés:

description de l'image, pyramide gaussienne demi octave, le profil laplacien, transformée de Fourier radiale, invariance, la reconnaissance

Abstract

Description of the image content extracts the information represented by the set of image pixels and expresses it in a form that is useful for matching, detection and recognition. Description is a basic procedure for all areas of computer vision. Results in vision processes can be significantly improved if the description method is invariant to changes in image contents. In this project, a new invariant image description method is proposed based on Laplacian profiles and Radial Fourier transform. The Laplacian profile of an image pixel is a sequence of Laplacian values of corresponding coordinates to different scales of the image. Adding the Radial Fourier transforms of the Laplacian values of small pixel neighbourhoods around the Laplacian profile corresponding coordinates of different scales formulates an invariant descriptor to translation, rotation and scaling. The computation of Laplacian values at different scales is easily achieved by the use of the Half-Octave Gaussian Pyramid. The size of the descriptor can be varied by changing the number of pyramid levels used. Experiments attest to the capabilities of the new proposed method towards invariance (to translation, rotation and scaling) and discriminative power and as well as show that the longest the Laplacian profile is, so the longest the size of the descriptor, the better the results are. The new description method appears suitable for detecting and recognizing publicity logos on images or, furthermore, tracking publicity logos on sequences of images taken from cameras on cell phones.

Keywords:

image description, half-octave gaussian pyramid, laplacian profile, radial Fourier transform, invariance, recognition

Dedicated to all Computer Vision Master students that I met during the time of my Master.

Acknowledgments

I would like to thank my colleagues in PRIMA team for their friendship and assistance. Special thanks to Prof. James L. Crowley and my colleagues Harsimrat Singh Sandhawalia, Claudine Combe and John-Alexander Ruiz-Hernandez.

Contents

List of Figures

List of symbols

$SIFT$	Scale Invariant Feature Transform
DoG	Difference of Gaussians
LoG	Laplacian of Gaussian
HOG	Histogram of Oriented Gradient
$RDFT$	Radial Discrete Fourier Transform
SVM	Support Vector Machine(s)
$RMSE$	Root Mean Square Error
DR	Detection Rate
ER	Error Rate
$FPPW$	False Positive Per Window

Chapter 1

Introduction

1.1 Introduction

1.1.1 Computer vision vision

To develop a technology that provides machines with the ability to see with performance found in biological systems is a very difficult and challenging problem that when solved is expected to have an enormous impact on human quality of life, leading to a technology for more "intelligent"machines that will be able to evaluate and further interact with their environment by visual stimuli. Computer vision is a mixture of artificial intelligence and machine learning focused on obtaining information from images. Image processing, the predecessor and base of computer vision, has existed since the 1960's but the first big steps towards artificial vision were taken in the 1980's, when methods for interpreting image content by mathematical calculations on pixels were introduced. Computer vision is an extended field with a number of subdomains exploring a variety of different issues from image representation to camera integration.

Description in computer vision is the extraction of information from images. The description of image content has two major factors: the invariance to image transformations and the discrimination ability on existing objects or structures in images. Algorithms for image description have been introduced over the previous decade. In the 1990's, methods used in computer vision were mainly geometry-based, such as Multi Camera reconstruction for 3-D scenes representation and mathematical matrices that expressed projective geometric relationships such as the Fundamental Matrix and the Trifocal Tensor. In the 2000's, appearance-based methods were introduced, such as Scale-Invariant Feature Transform (SIFT) [31] and Histogram of Oriented Gradients (HOG) [19]. Appearance in an image can be interpreted as the mathematical representation of attributes like shape, color, lightning, direction or size. The success of appearance-based methods have led researchers to focus on them. Image content description based on appearance can be both invariant and discriminative, without the one factor burdening the other, because it considers objects or structures as a unity and aims to identify these unities in images. This project proposes a new method that attempts to

capture appearance and perform efficiently concerning both invariance and discrimination.

1.1.2 Setting the Problem

Contributing to a field as vast as computer vision, requires centring on several smaller and more particular problems and working on their solution. The motivation for this project lies in the need for an efficient method to detect and recognize publicity logos from real-world images as well as tracking and recognition of publicity logos on sequences of real-world images (e.g. videos). For the rest of this document, the use of the term detection in images implies also tracking in sequences of images. Logos can exist in countless different positions and directions in real-world images and, nevertheless, the variety of them in shape, size, color combination and structure is endless. This project is the first step towards developing a new high performance image descriptor that will be able to perform on images belonging to complex objects. To achieve this goal, we seek a highly discriminative invariant image descriptor.

An image descriptor is a vector which represents an image or a part of an image in a certain manner, including Laplacian profiles and Radial Fourier transform, capturing meaningful information on pixel neighbourhoods. In this first approach, invariance exploration is limited only to translation, rotation and scaling for images. The starting point of the descriptor construction is the representation of images to a color space that can reveal all the important information of the image content. The LC_1C_2 color space was chosen as it provides an effective image representation. The Half-Octave Gaussian pyramid is used to calculate Laplacian values in different scales and provide the Laplacian profiles to corresponding image positions as well as Laplacian values of neighbourhoods around the Laplacian profiles, imposing invariance to different scales and positions. Radial Fourier transforms computed on Laplacian values of neighbourhoods around the Laplacian profiles provided magnitude and phase information on local appearance that enhanced the invariance to rotation. Laplacian profiles and Radial Fourier information were stacked in vectors to formulate descriptors covering the whole image. The normalization of the vectors with the L_2 formula added to the improvement of the descriptors. Finally, all the descriptors of an image were further combined in a large global descriptor vector that describes the image.

1.1.3 Experimental validation

In this project we have performed first experiments to evaluate invariance to translation, rotation and scaling and as well discrimination power. These initial experiments have provided the confirmation that the new method can work in practice. Both for the tests of invariance and discrimination, results indicated that both invariance and discrimination improve as the sizes of descriptor vectors are increased by the addition

of information from the Half-Octave Gaussian pyramid levels. The more pyramid levels used, the more invariant and discriminative the descriptors become.

Invariance is measured by the Root Mean Square Error (RMSE) [49]. Translation invariance experiments revealed a periodic noise in the descriptor. The RMSE plot, for a fixed interval, repeats in the exact same way. A probable explanation for this repeating trend is the Radial Discrete Fourier transform used in place of the Radial Fourier Transform (due to images being discrete signals) in the construction of the descriptors. Rotation invariance RMSE plot showed invariance to symmetry. Scaling invariance experiments showed expected results. While the scaling becomes increases, the RMSE increases. In general, the results are not perfect but they are good enough in the sense that they confirm invariance.

Discrimination has also been demonstrated. Experiments were conducted using the INRIA Person Dataset collected by Dalal [17], an image dataset widely used and suitable for demonstrating discrimination. Global image descriptors from 2416 images containing humans and 4832 images not containing humans resulted in different classification models for the Support Vector Machines (SVM) method [11]. The models were used to identify several patches taken from an image as positive (containing humans) or negative (non containing humans). This is a common method for locating humans present in images. Results showed that the majority of humans were located in the testset images. Therefore, we can be optimistic that discrimination can be obtained while maintaining invariance to translation, rotation and scaling.

1.1.4 Chapter sequence

The present chapter introduces the reader to common methods of image description and summarizes the initiatives and objectives of this project. Chapter Two describe the most important currents in the existing state of the art of image description by discussing three widely used techniques and concluding through the comparison of their attributes that the proposed method should perform dense description. Chapter Three establishes the particular problem of publicity logo detection and recognition in real world images, to which this project contributes, and sets the demands for an efficient new method on it. Chapter Four describes in detail the proposed method. Representing the image in LC_1C_2 color space increases the probability for success as this color space provides meaningful details on the image content. Extracting Laplacian profiles and collecting Radial Fourier information on small circular neighbourhoods around them formulate the basic descriptor vector structure on each of the three color space components. Normalization adds to the improvement of performance. Finally, the SVM method is used for creating classification models used in discrimination experiments. In Chapter Five, the method is tested for evaluating its invariance abilities towards translation, rotation and changes of scale. Invariance is demonstrated in all three cases

but the RMSE plots also reveal issues that need further exploration. In Chapter Six, tests continue on the evaluation of the discriminative power of the proposed method. Discrimination exists in the proposed method and performance improves depending on the quantity of information used from the Half-Octave Gaussian pyramid. The final Chapter reviews the conclusions reached through the experimental validation for the new method and collocates further improvements based either on the undesirable outcomes revealed during testing or on totally new ideas that move one step ahead in the construction of the proposed descriptors.

Chapter 2

Invariant Image Descriptors

2.1 Introduction

This Chapter reviews three widely used methods of the current state of the art for image description. The methods are firstly reviewed for using one image channel (grayscale) and then additional information is given on their extension to three image channels (color). At the end, there is a discussion on the disadvantages of this methods in relation to the problem of publicity logo detection and recognition. This provides a basis for the theoretical approach of the new image description method later explained in this report.

2.2 Scale Invariant Feature Transform (SIFT)

In 1999, David Lowe presented Scale Invariant Feature Transform (SIFT) [31]. SIFT transforms an image into a large collection of local descriptors vectors, each of which is invariant to image translation, scaling, and rotation, and robust to illumination changes and affine or 3-D projection. This description method performs description by keypoints: keypoints, otherwise called points of interest, refer to image points of specific importance according to a set of rules. Keypoints are located on an image and description takes place just around these points. Experimental results on SIFT attest to its successful performance.

In order to explain the way SIFT works, a brief explanation over the method called Difference of Gaussians (DoG) must be given. Convolving one image with two Gaussian functions of different variance and subtracting one convolved image from the other creates a new image that preserve spatial information that lies between the range of frequencies that are preserved in the two blurred images. This method is called the method of Difference of Gaussians and is a simpler way to implement the method of Laplacian of Gaussian (LoG), which is calculating the Laplacian derivatives of pixels on images. In LoG theory, local extrema indicate local change in original image and therefore indicate keypoints for SIFT. So, the local extrema on the Difference of Gaussian created images also indicate keypoints. Figure 2.1,adapted from [23], shows a DoG example and figure 2.2, adapted from [24], shows a Log example in one dimension. It is obvious that

the result (the blue line plot) is the same for DoG and LoG. Consequently, DoG is used to locate keypoints for SIFT.

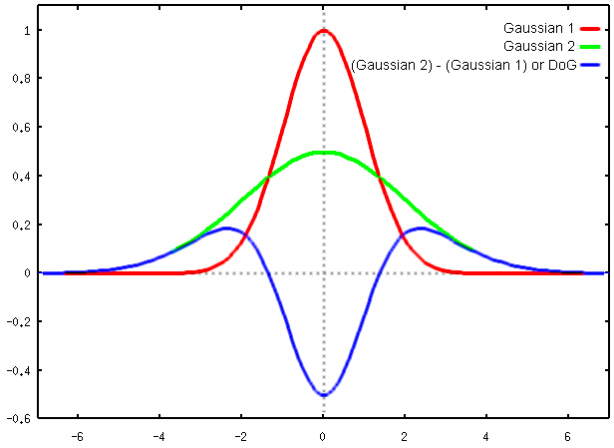

Figure 2.1: Difference of Gaussians. The Gaussian plotted in red color is subtracted by the one plotted in green color. The blue plot is the DoG. Adapted by [23]

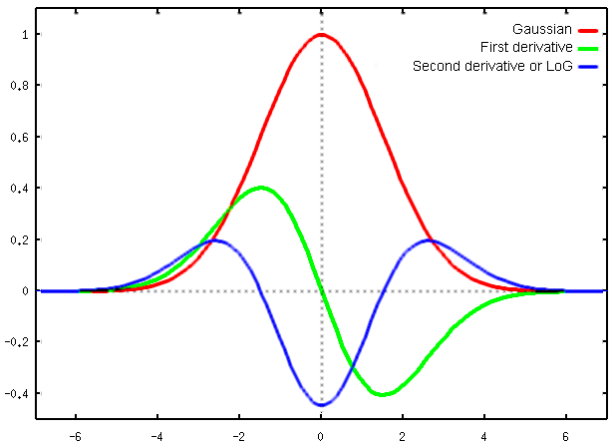

Figure 2.2: Laplacian of Gaussian. The red plot is a Gaussian. The green plot is the first derivative. The blue plot is The second derivative or LoG. Adapted by [24]

Keypoints are detected in different scales of the image. The original image is convolved with a set of Gaussian functions $G(x, y, \sigma_k)$, with variable standard deviation σ_k equal to $2^{\frac{k}{2}}$. The blurring accumulates the scaling of the image to smaller dimensions. Then the differences of successive Gaussian-blurred images are taken. Considering the original image as $I(x, y)$, a convolved image as $L(x, y, k)$ and a DoG image as $D(x, y, \sigma_k)$, the DoG image occurs from:

$$L(x, y, k) = G(x, y, \sigma_k) * I(x, y) \tag{2.1}$$

and

$$D(x, y, \sigma_k) = L(x, y, k_i) - L(x, y, k_j) \tag{2.2}$$

where k is a scaling factor and $\sigma{=}\sqrt{2}$ is the scale. The convolved images are grouped by octave (an octave corresponds to doubling the value of σ), and the value of k_i is selected to obtain a fixed number of convolved images per octave. Local extrema that occur at the multiple scales of the Difference of Gaussians are considered as good candidates for keypoints. More precisely, each pixel on a DoG image is compared to its eight neighbours at the same scale and nine corresponding neighbouring pixels in each of the two neighbouring scales. If the pixel value is a maximum/minimum among all compared pixels, it is considered as a keypoint. The keypoints found are then filtered so the less informative are eliminated. The combination of Difference of Gaussians and different scales results to collecting keypoints of different sizes, which leads

to a good image description. [31, 51, 50].

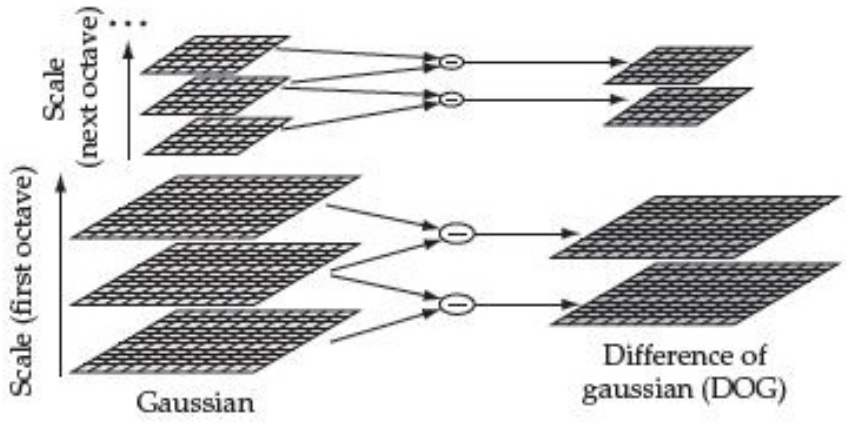

Figure 2.3: Differences of Gaussian images in SIFT. The original image is repeatedly smoothed by Gaussian functions. Gaussian images are subtracted to produce the DoG images. [28]

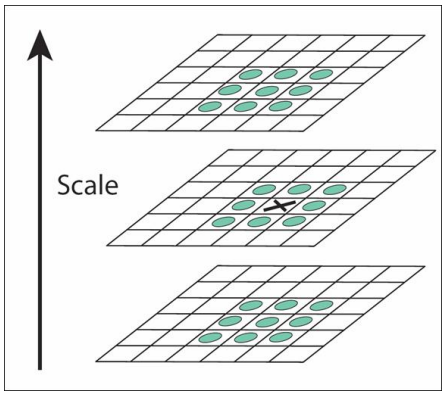

Figure 2.4: Selection of local extrema: each pixel on a DoG image is compared to its eight neighbours at the same scale and nine corresponding neighbouring pixels in each of the two neighbouring scales. If it is an extremum among all compared pixels, it is a keypoint. [40]

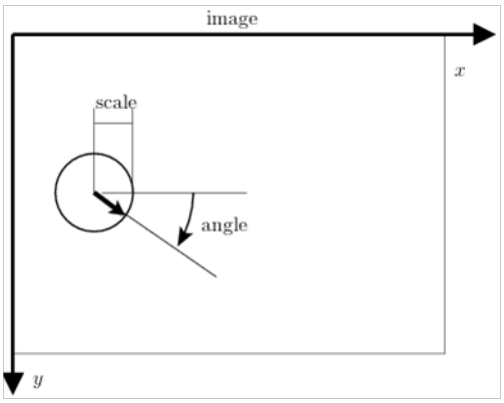

Figure 2.5: SIFT keypoint is a geometric frame of four parameters: the keypoint center coordinates x and y, its scale, and its orientation [43].

2.2.1 Grayscale SIFT

A SIFT keypoint, illustrated in figure 2.5 [43], is a circular image region with an orientation. It is described by a geometric frame of four parameters: the keypoint center coordinates x and y (point of interest), its scale (the radius of the region), and its orientation (an angle expressed in radians). [31, 43]

After locating the keypoints, SIFT creates local descriptors for local neighbourhoods, from a grid of histograms of oriented gradients. An image gradient is a directional change in the intensity or color in an image. The gradient of a pixel $p(x, y)$ is a 2-D vector with the components given by the derivatives in the horizontal and vertical directions of the image. The gradient vector points in the direction of largest intensity (darker colors), and the length of the gradient vector corresponds to the rate of change in that direction. In SIFT theory, the gradient at each pixel is regarded as a sample of a three-dimensional elementary feature vector, formed by the pixel coordinates and orientation. Consider a SIFT descriptor as a 3 dimensional spatial histogram of image gradients characterizing the appearance around a keypoint. In more detail, a set of 16 orientation histograms of 8 bins are created around a keypoint center. These histograms are computed from gradient magnitude and orientation values of samples in a 16×16 neighbourhood region around the keypoint, such that each histogram contains samples from a 4×4 subregion of the neighbourhood region. Samples are weighed by the gradient norm and accumulated in a 3 dimensional histogram h, which forms the local descriptor of the keypoint neighbourhood region. An additional Gaussian weighting function is applied to give less importance to gradients farther away from the keypoint. A SIFT descriptor is illustrated in figure 2.6, adapted from [43].

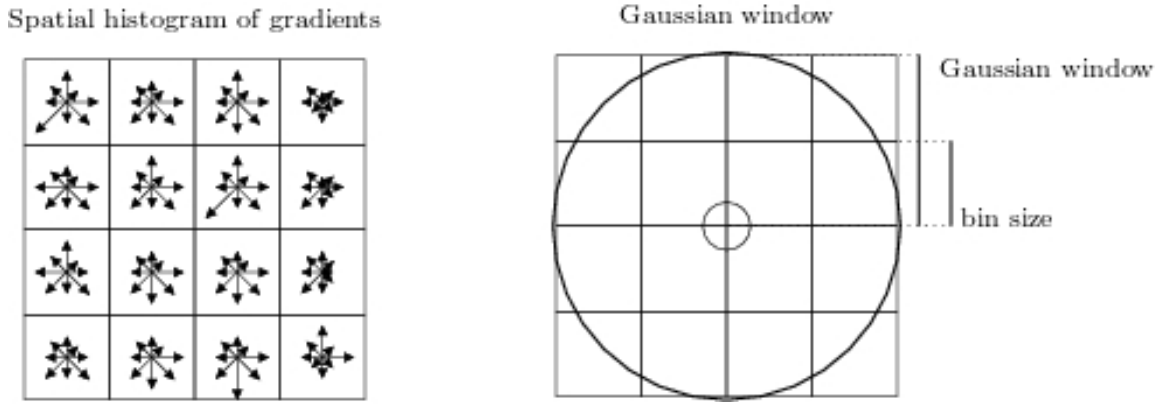

Figure 2.6: The SIFT descriptor is a spatial histogram of the image gradients: 16 histograms at 16 locations around a keypoint with 8 orientations each. The descriptor is further adjusted with an additional Gaussian weighting function. Adapted by [43]

A subset of SIFT descriptors found on an image are presented in figure 2.7. For creating these figures, an implementation of SIFT was used, created by Vedaldi and Fulkerson [43]. The number of the descriptors

shown on these images is extremely small regarding to the real number of the descriptors found to reduce clutter. The yellow circles have a radius analogous to the scale where the descriptor was created. The yellow line (radius) shows the main orientation of the descriptor and the descriptor itself is in green color, with the orientations of each bin being obvious at each 4×4 subregion. The 3-D histogram (consisting of $8 \times 4 \times 4 = 128$ bins) is stacked as a single 128-dimensional vector. [31, 43]

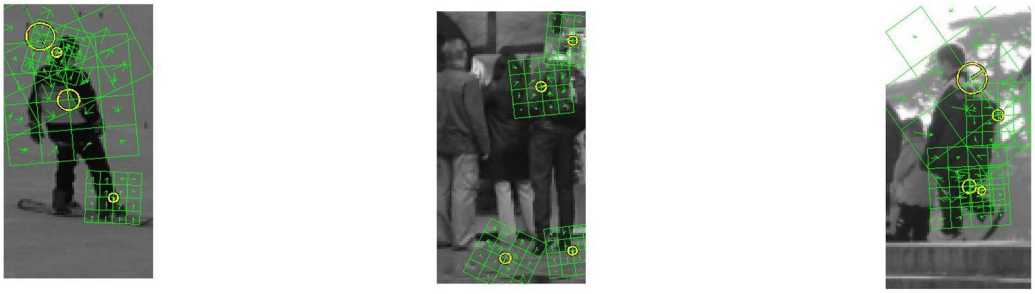

Figure 2.7: SIFT descriptors on images

2.2.2 Color SIFT

Alternative SIFT versions have been developed since the original was released, some of which use color components of the image instead of just one greylevel component. Burghouts and Geusebroek [8] propose and evaluate three colour SIFT methods and prove better performance of the algorithm against the original one. Three color components are derived from a linear transformation of the three RGB channels of the image, one representing intensity of the image and the rest two being chromatic components. The local greyvalue invariant gradients as proposed in the original SIFT version and a set of local photometric invariant gradients are calculated and compared based on the discriminative power they offer, the invariance to transformations and the information content they can carry. Information content refers to the ability of an invariant to distinguish between color transitions and photometric events such as shadow, shading and highlights. According discriminative power, it is shown that SIFT using color invariant gradients outperform SIFT using greyvalue gradients. In the part of invariance and information content, color SIFT methods have comparable results to the original SIFT version.

2.3 Histogram of Oriented Gradient (HOG)

Another very important image description method is the Histogram of Oriented Gradient (HOG) which was introduced by Dalal and Triggs [19] in 2005 and it was originally developed for human detection. The idea is to evaluate well-normalized local histograms of image gradient orientations in a dense grid, taking under

consideration that local object appearance and shape within an image can be described by the distribution of intensity gradients or edge directions, even without knowledge of their exact corresponding positions in the image.

2.3.1 Grayscale HOG

The use of orientation histograms was a subject of research for several years before 2005 [37, 38, 41] and set the theoretical base for both HOG and SIFT. The method works by dividing an image into smaller spatial regions, called cells, and for each cell computing a local 1-D histogram of gradient directions or edge orientations for the pixels within it. Therefore, HOG performs dense description. Dense description means that every image pixel is an important feature or a part of a feature and the whole image is subjected to description (features can be collected the one next to the other or overlapping each other). The term feature refers to a part of an image or a specific structure in the image, which can be a curve, a segment, a region with a homogeneous color/texture, or just a point (keypoint) [12]. Turning attention back to cells, cells can be rectangular (R-HOG) or circular (C-HOG). Before computing the gradients, it is essential that the original image undergoes Gaussian smoothing followed by a discrete derivative mask such as uncentred [-1, 1], centred [-1, 0, 1] or cubic-corrected [1, -8, 0, 8, -1], as well as 3×3 Sobel masks and 2×2 diagonal ones

$$\begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix} \text{ and } \begin{bmatrix} -1 & 0 \\ 0 & 1 \end{bmatrix} \tag{2.3}$$

Experiments showed that the simple 1-D [-1, 0, 1] masks at $\sigma = 0$ work best. Less essential is to normalized color and gamma values to the original image, though sometimes it can have better results. Figure 2.8, taken from the presentation of Dalal et al. [20] on HOG used for object detection, illustrates the explained method.

The histograms of the cells are evenly spread over 0 to 180 degrees, if the gradient is "signed", or 0 to 360 degrees,if the gradient is "unsigned". The histograms of the cells are finally contrast-normalized by a norm of gradients across a larger region of the image, called a block. The blocks can overlap, in which case some cells contribute more than once to the image description. The normalization of the cells in a block can be done by four different ways as proposed in [19]. Let v be the unnormalized descriptor vector containing all cell histograms in a block, $||v||$ k be its k-norm for k = 1,2 and e be some small constant. Then one of the following schemes can be used for normalization:

$$L2-norm: f = \frac{v}{\sqrt{||v||_2^2 + e^2}} \tag{2.4}$$

$$L1-norm: f = \frac{v}{(||v||_1 + e)} \tag{2.5}$$

$$L1-sqrt: f = \sqrt{\frac{v}{(||v||_1 + e)}} \tag{2.6}$$

$$L2-Hys: f = \frac{v}{\sqrt{||v||_2^2 + e^2}}, \text{ after limiting the maximum values of } v \text{ to } 0.2 \tag{2.7}$$

All schemes provide significant improvement to the performance of HOG, especially in invariance to changes in illumination or shadowing. The L2-Hys, L2-norm, and L1-sqrt schemes have similar good performance and the L1-norm has slightly less good performance than the rest. The normalized descriptor blocks are called the Histogram of Oriented Gradient (HOG) descriptors. The combination of the normalised oriented histograms of all the cells represent the descriptor of the image.

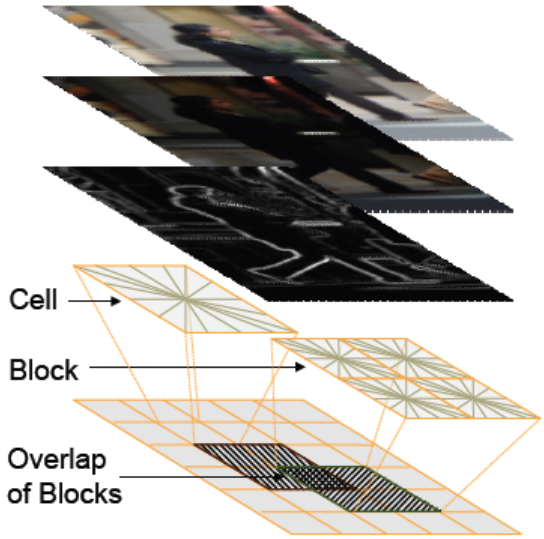

Figure 2.8: Histogram of Oriented Gradient. The image is segmented in cells. A block consists of a set of cells. Blocks can overlap. [20]

2.3.2 Color HOG

Research on HOG have also been focused in using color to increase power. In case of a color image, gradients are calculated separately for each color channel and then the one with the largest norm is taken as the pixel's gradient. Rishabh and Satish explain this in [36], where their implementation of HOG for human detection in RGB images is described. Furthermore, in Villamizar et al. [44] paper according detection on images, it

is proposed to use a color-based detector with HoG descriptors in order to create a system that improves detection performance in outdoor scenes under cast shadows.

2.4 GIST of image: Spatial Envelope

Other approaches than dense and keypoint description have also been explored. A very interesting approach in image description has been proposed by Oliva and Torralba [34]. In this work, the meaning of object in an image is differentiated by the meaning of the scene of an image. What is referred to as scene of an image, is the "gist"of the image rather than all the separate details. It is proposed that specific information about object shape or identity is not essential for scene recognition and that modelling a holistic representation of the scene gives enough information for its probable semantic category.

2.4.1 Grayscale GIST

Oliva and Torralba propose a computational model of the recognition of real world scenes based on a very low dimensional representation of the scene, for which they use the term Spatial Envelope. The model of Spatial Envelope generates a multidimensional space in which scenes sharing membership in semantic categories (e.g. forests, mountains, buildings, streets, highways, coasts) are projected in a similar way. A set of perceptual dimensions is proposed for categorizing images: naturalness, openness, roughness, expansion and ruggedness of the image scene. Figure 2.9 [34] presents an example for the degree of roughness of images with different content when projected in 3-D, with the third dimension corresponding to the intensity of the image pixels. These dimensions describe the dominant spatial structure of an image scene and represent the holistic spatial scene properties, termed Spatial Envelope properties.

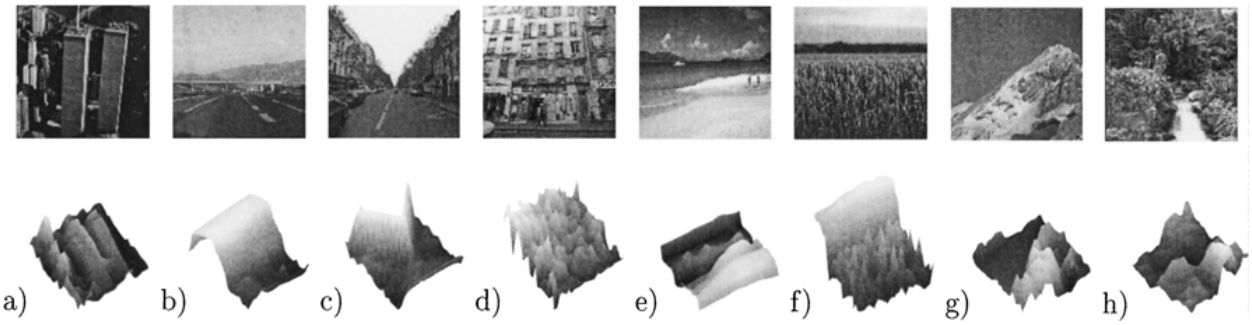

Figure 2.9: Scenes with different spatial envelopes and their surface representation, where the height level corresponds to the intensity at each pixel. The images show: a) skyscrapers, b) an highway, c) a perspective street, d) view on a flat building, e) a beach, f) a field, g) a mountain and e) a forest. [34]

Spatial Envelope scene properties in Oliva and Torralba's paper [34] are evaluated by the use of energy spectra. The spectra of many images of a semantic category (examples of basic level scene categories: tall buildings, highways, city close-up views, e.t.c) are averaged in one energy spectrum per category and the spectral signatures for each category are created. They use spectral signatures to categorize images.

Most of GIST method implementations as proposed in [22], resize the original image to a fixed size square image, where a number of orientation histograms are calculated on a grid and create the description of the image.

2.4.2 Color GIST

In [33] Oliva and Schyns examine the contribution of color in gist of scenes and prove that color is an important property of image description and recognition. The GIST method can be implemented to use image color by extracting the orientation histograms in all image channels and combining them into one descriptor [30].

2.5 Discussion on Existing Descriptors

Keypoint descriptors, like SIFT, perform remarkably well in several applications such as matching invariant keypoints or object detection. This breakthrough in computer vision led to draw attention from dense descriptors. Dense descriptors tend to be generally more powerful and computationally simpler than keypoint descriptors. As it is suggested by experiments in the paper of Dalal and Triggs [19], even the best current keypoint based approaches are likely to have false positive rates at least 1 or 2 orders of magnitude higher than the dense grid approach of HOG for human detection, mainly because keypoint detectors are not able to detect human body structures reliably. Keypoint detectors miss the perceptual information of the image or cannot describe properly objects as a unity. On the exact opposite side, descriptors that concentrate on general structure, the gist, of the image scene cannot capture details of objects. They are efficient for categorizing images simply and quickly, but not for object recognition. In contrast to both, dense descriptors can catch the appearance of objects and describe them efficiently. This is the reason why HOG method outperforms other methods in human detection.

In the case of logo detection and recognition, the same issue as in human detection appears. The appearance of publicity logos are of major importance as they consist of a set of objects (e.g. letters, shapes) and attributes (e.g. shape, texture, colors) that make sense only when combined in an exact way. Consequently, it is reasonable to ask for a dense image description that will lead to detecting and recognizing

logo structures in images properly.

A major disadvantage though of dense description methods is that they are not invariant to all possible image transformations. They usually are invariant to light or background variations but not to rotation or affine transformations. This is an issue in favour of keypoint description methods which are usually well invariant.

The basic introduction of the method proposed in this project is that it will conduct dense description, as it is the most suitable, but with respect to invariance. This is the reason why the construction of the descriptor vectors must contain information on both the exact appearance of an object in an image and general information about the image signal at that image area that can help identify the same pattern in different positions or angles. The challenge is to introduce a dense descriptor that will capture the most possible discriminative and invariant information without the one effecting negatively the other.

Chapter 3

Need for a new image description method

3.1 Logo detection and recognition

The original motivation for the description method proposed through this work was to perform publicity logo recognition on images taken by mobile phone cameras. Publicity logo detection and recognition is a complicated matter due to the wide variety of existing logos: different shapes, sizes, color combinations and structures. Additionally, real-world images containing logos are more complicated due to the infinite number of differences in background, light, view angle, distance from the camera, etc. Therefore, the description method to be used in such a case must be robust to every possible main content of the image and be able to locate and identify a logo correctly.

Figure 3.1: Examples of real-world images containing publicity logos [35, 7].

One main goal towards the evolution of a new method is that it must be strongly discriminative in order to be successful, without ignoring the need for vigorous invariance, exactly for the reason that publicity logo appearance can vary a lot in real-world images. But dense description methods, that are suitable for our purpose, lack in invariance. Our proposed method intends to be innovative especially for this case, which implies it should be able to maximize both discrimination and invariance without the one existing at the expense of the other.

A secondary but equally important issue to be examined for this particular goal, is the case of memory

constraint. According to the conclusion of Chapter 2, a dense description is the most suitable for logo detection and recognition. As assumed, the larger in size the descriptors of an image are, the more information they can represent and therefore allow robust recognition. But allowing a large descriptor size leads to the need of a lot of memory resources and huge calculations that affect the speed of performance. Setting the constrains that the process is real-time and on a computer with very limited memory capacities, the description method to be proposed must provide as small-sized descriptors as possible in order to perform in acceptable time limits. This problem is not examined in this project but is a subject for future work.

Logo detection and recognition is only one case of object detection in real world images that motivated this project. There are more cases as complicated as logos, where our proposed method could also be suitable. For example, such cases are detection and recognition of humans or faces in real world images.

3.2 Proposed descriptors

The introduced description method proposed in Chapter 4 aims to create image descriptors that can describe an image densely in an efficient way that obey the constrains set on the previous section. In this primary approach, attention is focused on two basic problems:

1. imposing invariance against the three basic image transformations, which are translation, rotation and scaling of the image, and

2. obtaining a good discrimination power.

Experiments presented in Chapters 5 and 6 prove that the proposed descriptors contain invariant information regarding all three basic image transformations, and have also a good discriminative power.

Chapter 4

Creation of a new Invariant Image Descriptor

4.1 Color space

Color has a high discriminative power and provides an important amount of information on the content of images [25, 8, 26, 41]. Consequently, it is reasonable that computer vision and image processing research is often focused on color representation of images. A color space, or color model, is a system for describing color numerically. Well known color spaces include the RGB color space for scanners, cameras and displays, CMYK for color printing and YUV for TV/video. The color space to be used depends on the research subject or on the application because every color space has different advantages and disadvantages.

In this project, the color space LC_1C_2 is used for the images. This color space was introduced by Shih and C. Liu [39] for face recognition because it provides an effective image representation. The abbreviation stands for luminance (L) channel and two chrominance channels (C_1, C_2). This color space is a linear transformation of the input RGB color space. The L channel is designed to capture intensity or luminance characteristics of an image. The C_1 and C_2 channels are designed to extract color properties such as hue and saturation. The transformation from the RGB color space to the LC_1C_2 color space is easily given by:

$$L = R + G + B \tag{4.1}$$

$$C_1 = R - B \tag{4.2}$$

$$C_2 = R + B - 2 \times G \tag{4.3}$$

In figure 4.1 there is an image of the Google Logo [42] and the three LC_1C_2 components are in figures 4.2, 4.3 and 4.4.

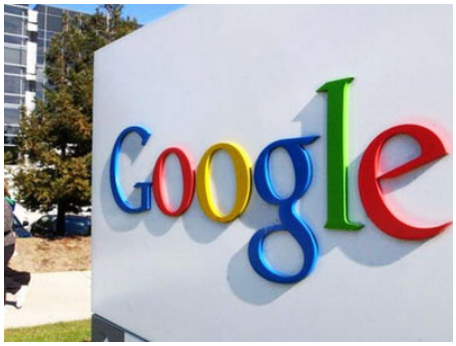

Figure 4.1: Original RGB image [42].

Figure 4.2: L component

Figure 4.3: C1 component

Figure 4.4: C2 component

4.2 Half-Octave pyramid

4.2.1 Gaussian derivatives

The next step is to decide how to extract meaningful information from the image representation. It is known since the 1980s that Gaussian derivatives provide a scale and rotation invariant image description [14, 16]. The use of Gaussian derivatives on image pixels are the basis of both HOG and SIFT (see Chapter 2). Hall and Crowley [27] used the Gaussian derivatives for face detection with log polar histograms. Crowley et al. [12] explain that the first and second Gaussian derivatives capture information about changes of the surface normal and measure the intensity of edges, the second order Gaussian derivatives (Laplacian of the Gaussian) are good descriptors for compact image features such as bars, blobs and corners and the higher order Gaussian derivatives are more sensitive to the image noise. They state that sets of Gaussian derivatives, second order derivatives especially, can be used to describe image neighbourhoods over a range of orientations and scales and allow an excellent description of local appearance for object detection and recognition. These sets are referred to as Gaussian Jet. They conclude that combining normalization of scale and orientation provides an invariant feature vector that can be used for robust detection and recognition.

A scale invariant local Jet for an $N \times N$ image requires computing second order derivatives of the image

at Log(N) scales. In the 1980's, a linear time pyramid algorithm was introduced for computation of the second order Gaussian derivatives Jet [9, 14, 16]. The algorithm creates an image structure named the Half-Octave Gaussian Pyramid. An integer coefficient version of the Half-Octave Gaussian Pyramid algorithm introduced by Crowley and Riff [15] uses repeated convolutions of the binomial kernel (1,2,1).

4.2.2 Implementing Half-Octave Gaussian pyramid

The Half-Octave Gaussian pyramid for an $N = W \times H$ image is composed of up to K images [13, 12]:

$$K = 2 \times Log_2(min(W, H)) \tag{4.4}$$

Each of the $k \in [1, K]$ images of the pyramid is convolved with the Gaussian filter $G(x, y, 2^{(k+1)/2})$ and resampled with a sample distance of $2^{(k-1)/2}$ (this is the reason why the pyramid it is called Half-Octave [48]), so that the scale/distance ratio is constant to ensure scale invariant impulse response. To produce the base-level pyramid image $P(x, y, 0)$, for $k = 0$, the original image is initially convolved with a filter with $\sigma_0 = 1$:

$$k = 0 : P(i, j, 0) = P(x, y) * G(i, j, \sigma_0) \tag{4.5}$$

where * is the convolution symbol. The $k = 0$ image can be retained instead of this image for high-resolution analysis if desired. The next level pyramid image, for $k = 1$, is produced by convolving the previous image, for $k = 0$, with a low pass Gaussian filter with $\sigma_1 = \sqrt{2}\sigma_0$:

$$k = 1 : P(i, j, 1) = P(i, j, 0) * G(i, j, \sqrt{2}\sigma_0) \tag{4.6}$$

which is the same as convolving:

$$k = 1 : P(i, j, 1) = P(i, j, 0) * G(i, j, \sigma_0) * G(i, j, \sigma_0) \tag{4.7}$$

The same process continues for $k > 1$, by convolving each previously sampled image with a low pass Gaussian filter:

$$k = 1 : P(i, j, k) = P(i, j, k - 1) * G(i, j, 2^{k/2}\sigma_0) \tag{4.8}$$

Every k image of the pyramid has half the number of samples of the $k - 1$ image and double the number of samples of the $k + 1$ image which results in a total of $2 \times N \times N$ samples in all k levels (images) of the pyramid. Aliasing is minimized (less than 1% of signal energy) by the fact that the image has been low-pass

filtered by previous convolutions. This algorithm has linear algorithmic complexity (i.e. $O(N)$) and gives a discrete representation of scale space with $2 \times M$ total samples. The aim of its use in this project is to extract imformation at different image scales.

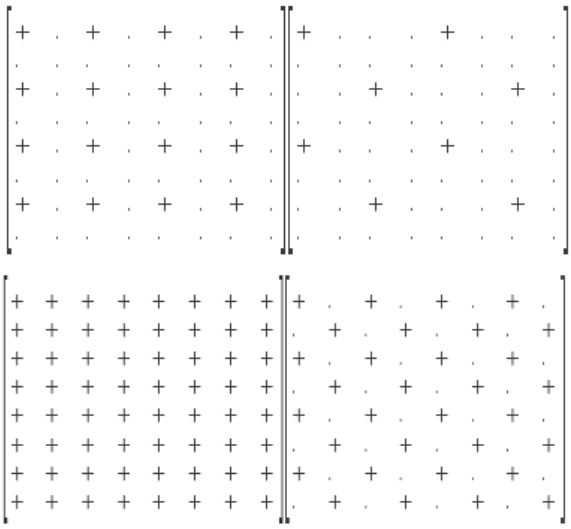

Figure 4.5: Illustration of sampling in the Half-Octave Gaussian pyramid. The symbol "+"represents image samples. Each image has half the pixels of the previous image, viewing the images from the bottom left towards top right. [13]

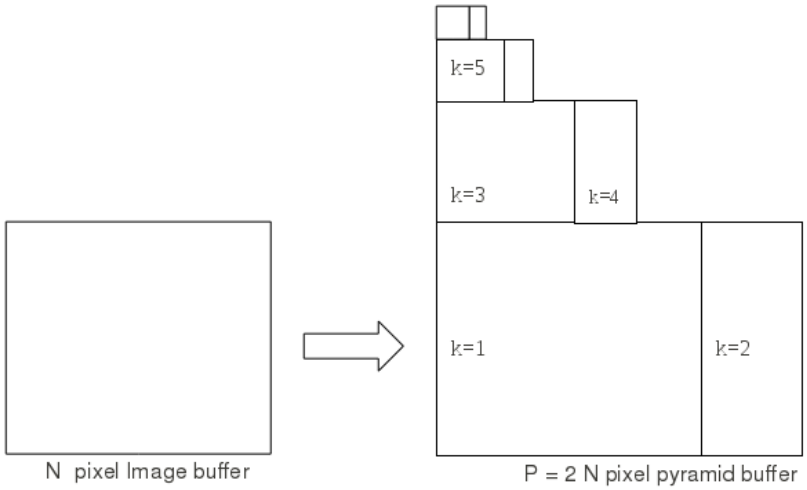

Figure 4.6: The pyramid is composed of $P = N(1 + 1/2 + 1/4 + 1/8 + \ldots) = 2N samples$ [13].

The implementation of the Half-Octave Gaussian pyramid used, made by Combe [10], constructs a collage image consisting of the convolved images created at each k level of the pyramid. The images are accessed by their coordinates, which are also part of the output. The enumeration of the pyramid levels starts from the lowest level to the highest. The first level is the lowest and largest and the last level is the highest and smallest. The pyramid was used on each of the three LC_1C_2 image components separately. The exact way is described in the next section.

Figure 4.7: The result of Half-Octave Gaussian pyramid implementation ont the L channel of the image [42]: the output is an image consisting of all the convolved images created at each k level of the pyramid.

4.3 Laplacian profiles

As mentioned previously, second order Gaussian derivatives (Laplacian of the Gaussian) are good descriptors for compact image features. We need to use this characteristic of the Laplacian values of image pixels in order to build the new descriptor. The Laplacian values of image pixels can be easily calculated with the method of the Half-Octave Gaussian pyramid. In [13], this process is well described: for a 2-D Gaussian function at scale σ, $G(x, y, \sigma)$, the equation is:

$$\frac{\partial G(x,y,\sigma)}{\partial \sigma} = \frac{\partial^2 G(x,y,\sigma)}{\partial x^2} + \frac{\partial^2 G(x,y,\sigma)}{\partial y^2} = \nabla^2 (G(x,y,\sigma) * P(i,j)) = \nabla^2 P(i,j,k) \tag{4.9}$$

For a 2-D Gaussian impulse response, the Laplacian can be computed either as a sum of second derivatives at a given scale k or as a difference of pyramid samples at adjacent scales k and $k-1$. The formula for both looks like:

$$\nabla^2 P(i,j,k) = P_{xx}(i,j,k) + P_{yy}(i,j,k) \approx P(i,j,k) - P(i,j,k-1) \tag{4.10}$$

The second computational manner stems from the Difference of Gaussians (DoG) that was already mentioned in Chapter 2 in the SIFT section. The implementation by Combe [10] computes the Laplacian values by difference of Pyramid samples at adjacent scales.

In order to create the descriptors, Laplacian values are collected at a subset of pyramid levels. The levels at the top of the pyramid lack of informative power compared to lower levels because the scaling is too intense. The Laplacian value at the bottom level cannot by computed because there is no level below. For the above reasons, only a subset of the middle levels of the pyramid are used for the descriptors.

Starting from the highest chosen level at a given pixel position and going down to the lowest chosen level, a Laplacian value is calculated at each level and stored into a vector. This vector is referred to as the Laplacian profile. The position at each level, where the Laplacian value is calculated, is indicated by the selected pixel position at the highest level following the rule that a pixel value at a level k is the interpolated value of the corresponding pixel neighbourhood at level $k-1$.

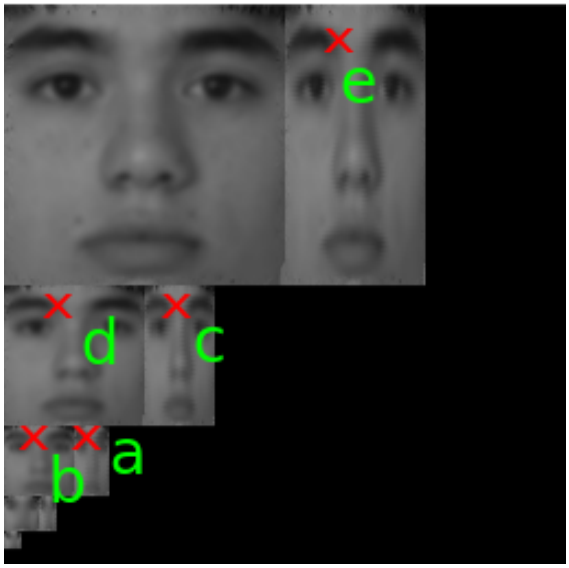

Figure 4.8: Laplacian profile: the vector created by computing the Laplacian values of corresponding coordinates at a selected subset of pyramid levels. The red x indicate the position and the green letters the levels computed Laplacian values. The structure of the Laplacian profile vector is $[a, b, c, d, e]$. The process is shown only at the L channel for simplicity.

In figure 4.8 is shown that from all the created pyramid levels, only the levels indicated with red x are taken into account for the creation of the descriptor. For simplicity, the process is shown only at the L channel. It is the same at the C_1 and C_2 channels. Using Laplacian values to a number of different pyramid levels imposes scaling invariance to the description method.

4.4 Radial Fourier transform using the Radial Discrete Fourier transform

The Fourier transformation is well known and broadly used as it defines a relationship between a signal in the time domain and the signal representation in the frequency domain without any addition or loss of information in the process [47]. It has the ability to capture the magnitude and the phase of a signal and provide significant details about its form.

In order to impose translation, rotation and scaling invariance as well enhance the discriminative power of the descriptors, the Laplacian values of a small neighbourhood of a circular ring at four pixels is taken around the Laplacian profile corresponding coordinates at each selected level except the first and the last one (for example in figure 4.8 these are levels b, c and d, see figure 4.9 for result). But these values are too few to add a significant amount of information to the descriptor. On the contrary, the use of a Fourier transform can provide more serious information extracted from these four pixel neighbourhoods, capturing the essence of the image around a Laplacian profile.

Consider the four neighbours of a Laplacian profile pixel existing on the periphery of a circle. Theoretically, the use of Radial Fourier transform in terms of angular frequency is the most efficient form of Fourier transform to capture the magnitude and the phase of the four pixel neighbourhood, as they exist on a circle. The formula for a continuous signal z using angular frequency is:

$$F(\omega) = \int_{\infty}^{-\infty} f(z) e^{-i\omega z} dz \tag{4.11}$$

where ω is the angular frequency.

Practically, the image is a discrete signal and the four neighbour Laplacian values are also discrete signals. In order to calculate the Radial Fourier transform, the generalized formula of Discrete Fourier transform (DFT) [45] for discrete signals is an efficient way as it was derived to particularly handle discrete signals. The Radial DFT (RDFT) formula, for a sequence of N complex numbers x_0, ..., x_{N-1} that are transformed into the frequency space sequence of N complex numbers X_0, ..., X_{N-1}, interpreting the angular frequency ω in its discrete equivalent $\omega = \frac{2\pi nk}{N}$ and replacing the integral by summation, is given by:

$$X_k = \sum_{n=0}^{N-1} x_n e^{-\frac{2\pi i}{N} kn} \tag{4.12}$$

for k= 0, 1, ..., N-1 and i being the imaginary unit

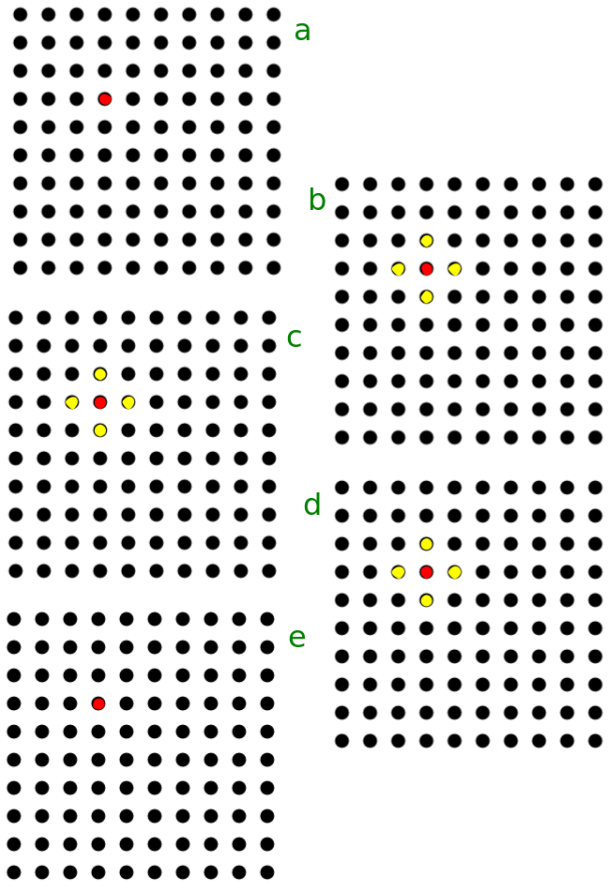

Figure 4.9: Laplacian values of four neighbours are taken around the Laplacian profile corresponding coordinates of all selected levels except the first and the last one. The rectangular areas of black dots as rectangular areas cut off from the previous figure around the red x (Laplacian values). Consider the four neighbours of a Laplacian profile corresponding pixel existing on the periphery of a circle.

For the four Laplacian values x_0, x_1, x_2 and x_3 of every four sample neighbourhood of each level, the Radial Discrete Fourier transform values are:

$$X_0 = x_0 e^{-\frac{2\pi i}{4}0\times 0} + x_1 e^{-\frac{2\pi i}{4}0\times 1} + x_2 e^{-\frac{2\pi i}{4}0\times 2} + x_3 e^{-\frac{2\pi i}{4}0\times 3} \tag{4.13}$$

$$X_1 = x_0 e^{-\frac{2\pi i}{4}1\times 0} + x_1 e^{-\frac{2\pi i}{4}1\times 1} + x_2 e^{-\frac{2\pi i}{4}1\times 2} + x_3 e^{-\frac{2\pi i}{4}1\times 3} \tag{4.14}$$

$$X_2 = x_0 e^{-\frac{2\pi i}{4}2\times 0} + x_1 e^{-\frac{2\pi i}{4}2\times 1} + x_2 e^{-\frac{2\pi i}{4}2\times 2} + x_3 e^{-\frac{2\pi i}{4}2\times 3} \tag{4.15}$$

$$X_3 = x_0 e^{-\frac{2\pi i}{4}3\times 0} + x_1 e^{-\frac{2\pi i}{4}3\times 1} + x_2 e^{-\frac{2\pi i}{4}3\times 2} + x_3 e^{-\frac{2\pi i}{4}3\times 3} \tag{4.16}$$

Solving the equation, by using Euler's formula [46] when necessary, they become:

$$X_0 = x_0 + x_1 + x_2 + x_3 \tag{4.17}$$

$$X_1 = x_0 + x_1 e^{-\frac{\pi i}{2}} + x_2 e^{-\pi i} + x_3 e^{-\frac{3\pi i}{2}} = x_0 + x_1(-i) + x_2(-1) + x_3 i = x_0 - x_1 i - x_2 + x_3 i \tag{4.18}$$

$$X_2 = x_0 + x_1 e^{-\pi i} + x_2 e^{-2\pi i} + x_3 e^{-\frac{3\pi i}{4}} = x_0 + x_1(-1) + x_2 + x_3(-1) = x_0 - x_1 + x_2 - x_3 \tag{4.19}$$

$$X_3 = x_0 + x_1 e^{-\frac{\pi i}{2}} + x_2 e^{-\pi i} + x_3 e^{-\frac{3\pi i}{2}} = x_0 + x_1(-i) + x_2(-1) + x_3 i = x_0 - x_1 i - x_2 + x_3 i \tag{4.20}$$

The results show that X_0 has only a real part and is the sum of x_0, x_1, x_2 and x_3, X_1 and X_3 are the same with real and imaginary part and X_2 has only a real part but signed. The Radial Discrete Fourier transform can be used to replace, in the descriptor vector, the four neighbour Laplacian values of each level taken around the Laplacian profile corresponding coordinates with meaningful information about the magnitude and phase of the image signal at this limited local area. One pair of meaningful information is the absolute value and the sign of X_2. The second pair is the magnitude and phase of the complex number X_1 (or X_3). The X_0 number is dependent on the actual location, the coordinates, where the calculations on the image take place [52]. For this reason, it was decided to be ignored in the construction of the descriptors. Considering a complex number $z = x + iy$, magnitude A and phase Φ are computed by the formulas [45]:

$$A = \sqrt{x^2 + y^2} \tag{4.21}$$

$$\Phi = atan2(x, y) \tag{4.22}$$

As X_1 (or X_3) are complex numbers, their magnitude A and phase Φ can be calculated by the above formulas. Magnitude depicts the amount of a certain existing frequency and phase can be interpreted as the position of this frequency in the image. Image translation and rotation impose a scaling by multiplication with a linear phase factor to the X_k values. Their module is not affected by such multiplication so they are independent of translation and rotation to the original neighbourhood. Therefore, magnitude and phase should not be affected. Magnitude A along with the absolute value and sign of X_2 are expected to provide information to assist translation invariance, and phase Φ is expected to provide rotation invariance. Eventually, four values, which are the absolute value and the sign of X_2, the magnitude A and the phase Φ from X_1 (or X_3), provide important local information which can describe an area relatively in a unique way that can force a descriptor to identify this area regardless of translation or rotation of the image. Repeating the same in more than one pyramid level following the Laplacian profile of the area, provides extra invariance of the descriptor to scaling of the image. The order of the four values in the descriptor is absolute value of X_2 -

A - sign(X_2) - Φ. To summarize, the final descriptor vector is visualized in figure 4.10. For simplicity, the process is shown at only one channel. It is the same for all channels. In order to create a descriptor of all the three L, C_1 and C_2 channels, the three final descriptors of each channel are concatenated by order L - C_1 - C_2 :

$$\text{Color descriptor} = \textit{concat}(\text{ descriptor for } L,\ \text{descriptor for } C_1,\ \text{descriptor for } C_2\ ) \tag{4.23}$$

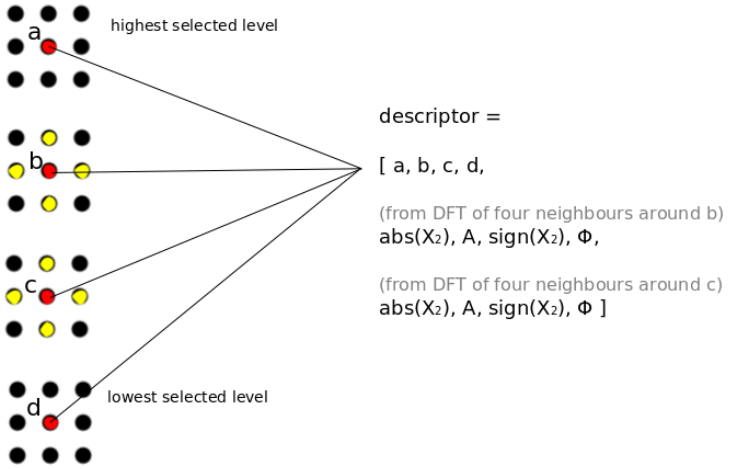

Figure 4.10: The structure of the final descriptor vector in case of four selected levels in the pyramid and only for one channel of the image. The red dots represent the pixels corresponting to the Laplacian values of the Laplacian profile. The yellow dots represent the four neighbour areas around the Laplacian profile pixels of the middle selected pyramid levels.

4.4.1 Implementing Radial Discrete Fourier transform

The implementation of RDFT for the four neighbour Laplacian values was done by using Euler's form for the $e^{-\frac{2\pi i}{N}kn}$ factor of the RDFT formula, that is:

$$e^{-\frac{2\pi i}{N}kn} = \cos(\frac{2\pi}{N}kn) - i\sin(\frac{2\pi}{N}kn) \tag{4.24}$$

Presenting the transformation of x_0, x_1, x_2 and x_3 to the frequency space values X_0, X_1, X_2 and X_3 in matrix form, we get:

$$\begin{bmatrix} X_0 \\ X_1 \\ X_2 \\ X_3 \end{bmatrix} = \begin{bmatrix} e^{-\frac{2\pi i}{4}0\times0} & e^{-\frac{2\pi i}{4}0\times1} & e^{-\frac{2\pi i}{4}0\times2} & e^{-\frac{2\pi i}{4}0\times3} \\ e^{-\frac{2\pi i}{4}1\times0} & e^{-\frac{2\pi i}{4}1\times1} & e^{-\frac{2\pi i}{4}1\times2} & e^{-\frac{2\pi i}{4}1\times3} \\ e^{-\frac{2\pi i}{4}2\times0} & e^{-\frac{2\pi i}{4}2\times1} & e^{-\frac{2\pi i}{4}2\times2} & e^{-\frac{2\pi i}{4}2\times3} \\ e^{-\frac{2\pi i}{4}3\times0} & e^{-\frac{2\pi i}{4}3\times1} & e^{-\frac{2\pi i}{4}3\times2} & e^{-\frac{2\pi i}{4}3\times3} \end{bmatrix} \times \begin{bmatrix} x_0 \\ x_1 \\ x_2 \\ x_3 \end{bmatrix} \tag{4.25}$$

and using Euler's form while making the multiplications, the equation becomes:

$$\begin{bmatrix} X_0 \\ X_1 \\ X_2 \\ X_3 \end{bmatrix} = \begin{bmatrix} \cos(0) - i\sin(0) & \cos(0) - i\sin(0) & \cos(0) - i\sin(0) & \cos(0) - i\sin(0) \\ \cos(0) - i\sin(0) & \cos(\frac{\pi}{2}) - i\sin(\frac{\pi}{2}) & \cos(\pi) - i\sin(\pi) & \cos(\frac{3\pi}{2}) - i\sin(\frac{3\pi}{2}) \\ \cos(0) - i\sin(0) & \cos(\pi) - i\sin(\pi) & \cos(2\pi) - i\sin(2\pi) & \cos(3\pi) - i\sin(3\pi) \\ \cos(0) - i\sin(0) & \cos(\frac{3\pi}{2}) - i\sin(\frac{3\pi}{2}) & \cos(3\pi) - i\sin(3\pi) & \cos(\frac{9\pi}{2}) - i\sin(\frac{9\pi}{2}) \end{bmatrix} \times \begin{bmatrix} x_0 \\ x_1 \\ x_2 \\ x_3 \end{bmatrix} \tag{4.26}$$

By replacing sines and cosines and dismantling the relevant matrix, the equation becomes:

$$\begin{bmatrix} X_0 \\ X_1 \\ X_2 \\ X_3 \end{bmatrix} = \left(\begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 0 & -1 & 0 \\ 1 & -1 & 1 & -1 \\ 1 & 0 & -1 & 0 \end{bmatrix} - i \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & -1 \\ 0 & 0 & 0 & 0 \\ 0 & -1 & 0 & 1 \end{bmatrix} \right) \times \begin{bmatrix} x_0 \\ x_1 \\ x_2 \\ x_3 \end{bmatrix} \tag{4.27}$$

which is very simple to be interpreted into source code in order to get the real and imaginary part of X_0, X_1, X_2 and X_3.

4.5 Normalization

The last but very necessary operation is the normalization of the descriptor vector. Normalizing descriptor vectors makes them comparable. A normalized vector v, also called unit vectors or norm, has a magnitude of 1 and it is symbolized as $\hat{v}$. The normalized vector $\hat{v}$ is given by:

$$\hat{v} = \frac{v}{||v||}, \text{ where } ||v|| \text{ is the norm of the vector } v \tag{4.28}$$

It was decided that the norm of the vector, $||v||$, must be the L_2 norm, also named Euclidean norm, because the elements of the descriptor vectors have signed numbers. The L_2 norm is:

$$L_2 = \sqrt{\sum_{i=0}^{n} x_i^2}, \text{ where } n \text{ is the vector size and } x_i \text{ the vector elements} \tag{4.29}$$

This normalization forces the descriptor vectors to be comparable. After the normalization, the test results improve significantly, especially for complicated datasets.

4.6 Visualization of the descriptor

At this point, the descriptors are completed. To make the procedure described step by step in the previous sections clearer, it is wise to summarize it visually. Figure 4.11 explaines briefly the procedure. To clarify the way this method works, think that moving higher to the pyramid levels, due to the levels becoming smaller, the small four pixel neighbourhoods around the Laplacian profiles correspond to larger areas on the original image. This happens because a pixel at higher levels is interpolated from a neighbourhood of pixels at a lower level, see figure 4.12 for illustration. This way a descriptor, corresponding to a couple of coordinates on the higher selected pyramid level, can represent a relatively big part on the original image. The higher the pyramid level the descriptor starts from, the larger the described area is on the original image. Consequently, the four pixel neighbourhoods at each level correspond to a different area size in the original image but around the same center which corresponds to the coordinates of the starting point of the descriptor at the highest level. Subsequently, the descriptors catch the appearance of the area in several proportions defined by concentric disks of different radius.

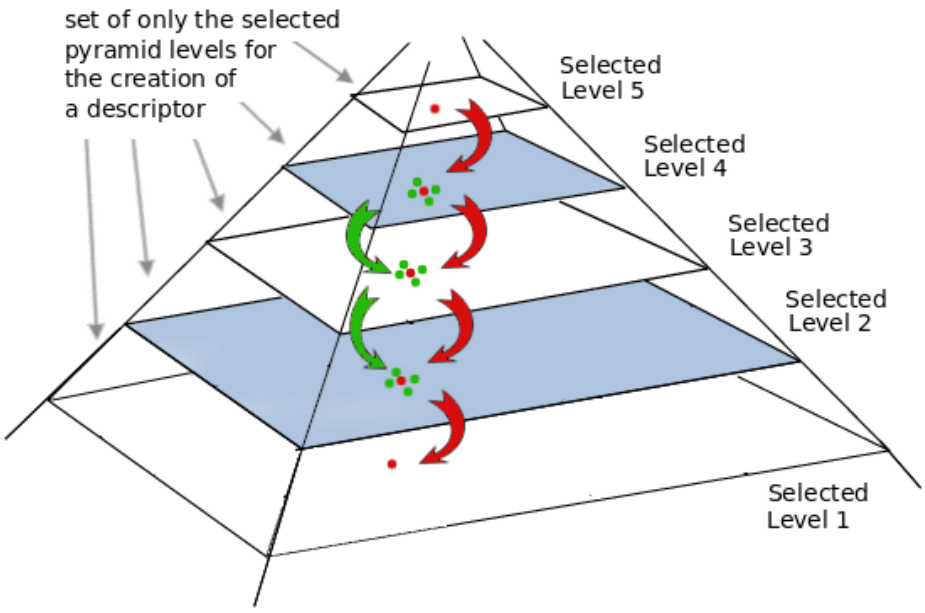

Figure 4.11: Constructing the descriptor for 5 levels of the pyramid. Consider these five levels as the selected levels of a pyramid for the construction of the descriptor corresponding to the pixel coordinates at the top level of this image (red dot at Level 5). First the procedure follows the red arrows to collect the Laplacian profile values and then follows the green arrows to add the RDFT details of the Laplacian values of the four neighbour areas around the Laplacian profile at middle levels. Using pixels to a number of different pyramid levels imposes scaling invariance, using Laplacian profile values and Radial Fourier derived values and magnitudes of local four pixel neighbourhoods imposes translation invariance and using phase information for the local four pixel neighbourhoods imposes rotation invariance. Pyramid skeleton adapted from [29].

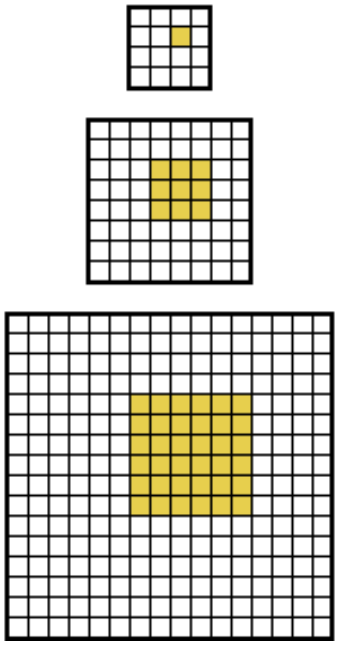

Figure 4.12: A small neighbourhood in a higher level corresponds to a larger neighbourhood to lower levels. Adapted from [32].

Figure 4.13: The appearance of the area is caught in several proportions defined by concentric disks of different radius.

4.7 Support Vector Machines

A Support Vector Machine (SVM) is a data classification algorithm for supervised learning. The SVM theory in its present form was introduced by Cortes and Vapnik [11] in 1995. The concept is to map data, that are difficult to classify, to a predetermined higher dimensional space via a Kernel function K where they can be separated by a hyperplane or set of hyperplanes into classes. The support vectors are the samples (consider the samples as vectors with elements the coordinates at each dimension) that define the hyperplane. SVM classifiers are very commonly used in computer vision because the size of data is usually enormous. Dalal and Triggs [19] used SVM for human detection. Dorkó and Schmid [21] used SVM over keypoints as intermediate part detectors for general object recognition, and tested two types of final classifiers. For the experiments in Chapter 6, the online available SVM library LIBLINEAR [5] was used, a linear SVM classifier for data with millions of instances and features that works relatively fast. LIBLINEAR was used for creating classification models from image descriptors of a training set and then classification of image descriptors of a test set.

A SVM finds the hyperplane that: a) maximizes the margin between training data of different classes or b) maximizes the margin and minimizes misclassifications in case the training data are not separable. Margin is the distance between the nearest training data samples of any class. The first kind is called Hard Margin SVM and is used when the data are seperable in a higher dimensional space. The second kind is

called Soft Margin SVM and is used when the data are not linearly separable even in a higher dimensional space by Hard Margin SVM. The difference between Hard and Soft Margin SVM is that Soft Margin allows a few outliers not to be correctly classified when the data are not linearly seperable so as to be able to find a solution. Misclassified samples and samples within the margin are penalizes but allowed to exist. In Soft Margin SVM, a constant C balances between the two parts of the criterion, classification and penalizing. The larger the value of C is, the more intense the penalization is, so the SVM tends to be Hard Margin. In the experiments that will follow, we use the INRIA Person Dataset by Dalal [17]. Human bodies in images are as hard as logos to be described in images due to their complexity in appearance. Due to the size and complexity of the dataset, a Soft Margin SVM is believed to be more suitable for experiments. This is already proposed in [18] and [36] where the same dataset was used.

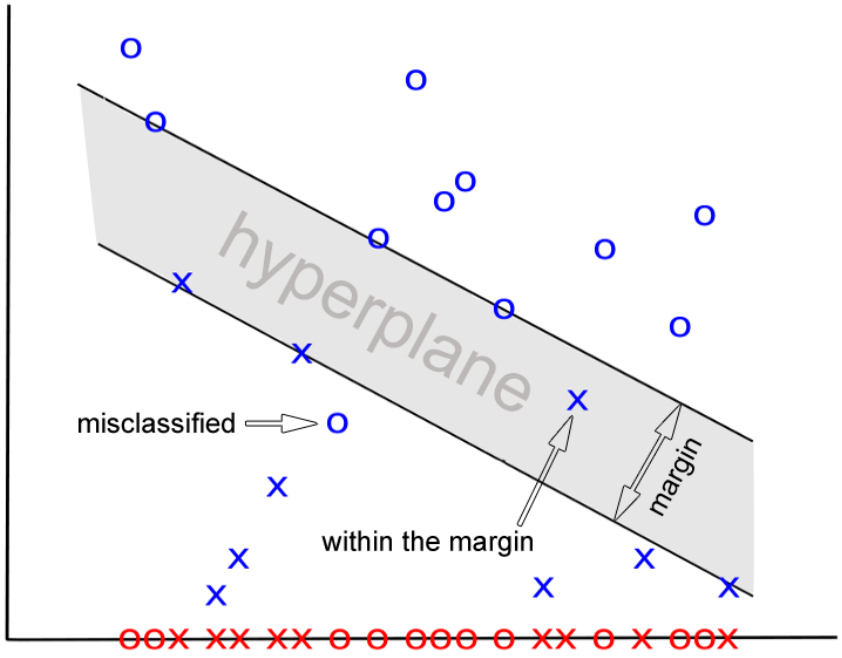

Figure 4.14: Soft Margin SVM: in this example, a SVM needs to be trained with samples from two classes in one dimension. The samples in one dimension (red x and o) are very difficult to separate. But after projecting them to a higher dimensional space of two dimensions (blue x and o) with a Kernel function K, they can be easily separated by a linear hyperplane. There is only one x which is one the hyperplane and one o that is misclassified. These outliers are penalized but allowed by the SVM in order to find a linear solution. Samples on the margin are the support vectors.

Chapter 5

Evaluation of invariance

5.1 Experiments

5.1.1 Experimental hypothesis on invariance

The first round of experiments is designed to test the invariance of the new method concerning translation, rotation and scaling. The description method is expected to be translation, rotation and scaling invariant. The reasons are those explained in Chapter 4: a) using pixels at a number of different pyramid levels is expected to impose scaling invariance to the description method, b) using Laplacian profile values and Radial Fourier derived values and magnitudes of local four pixel neighbourhoods around the Laplacian profiles is expected to impose translation invariance and c) using phase information for the local four pixel neighbourhoods around the Laplacian profiles is expected to impose rotation invariance. However, this ideal invariance is expected to be degraded by quantization and sampling noise required for compacting the descriptor. The extent of this degradation is unknown.

5.1.2 Initial experimental evaluation of invariance

The rule for these experiments is to select a pixel on an image and define a particular image transformation of a specified size, e.g. translate the image 10 pixels to the right or rotate the image 20 degrees with this pixel as rotation center. A descriptor vector is created for this pixel for the original image and one for each of the possible transformed images indicated by the parameters. For example, for x pixels of translation to the left, a total of $x + 1$ descriptors are made, one for the original image position, one for the original image position + 1 pixel, one for the original image position + 2 pixels, etc. The selected pixel follows the transformations, meaning it is translated with translation, it is the rotation center of rotation and it is interpolated in scale changes. Even though only a single pixel is selected on the image, the descriptor vector corresponds to a larger area on the image, as explained in Chapter 4.

In order to perform the transformations, the open-source online library OpenCV 2.0 [3] was used. The

description is formed from the highest selected pyramid level towards the lowest in order to create a descriptor from a pixel selected at the original image. Its position on the highest selected pyramid level is located according to the constrains set by the pyramid interpolation method. Invariance is measured by the Root Mean Square Error (RMSE) [49], otherwise named Root Mean Square Deviation (RMSD), between the descriptor vector for the original image pixel position and the descriptor vector for each transformed image pixel corresponding position. The formula is:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(x_{1,i} - x_{2,i})^2}{n}} \tag{5.1}$$

where $x_{1,i}$ and $x_{2,i}$ elements of two vectors. The experiments show the existence of invariance to all three transformations.

5.2 Results

5.2.1 Translation

The first experiment shows translation of an image to the right for 114 pixels.

Figure 5.1: Translation 114 pixels to the right.

The selected pixel is shown as a red spot on both the original image and the translated copy (figure 5.1). The plot in figure 5.2 presents the RMSE for descriptors created for 3 (green), 4 (blue), 5 (red), 6 (magenta) and 7 (black) pyramid levels. The number of pyramid levels selected, for this case and for the rest of the following cases, depends on the image size. Bigger images can provide pyramids with more levels.

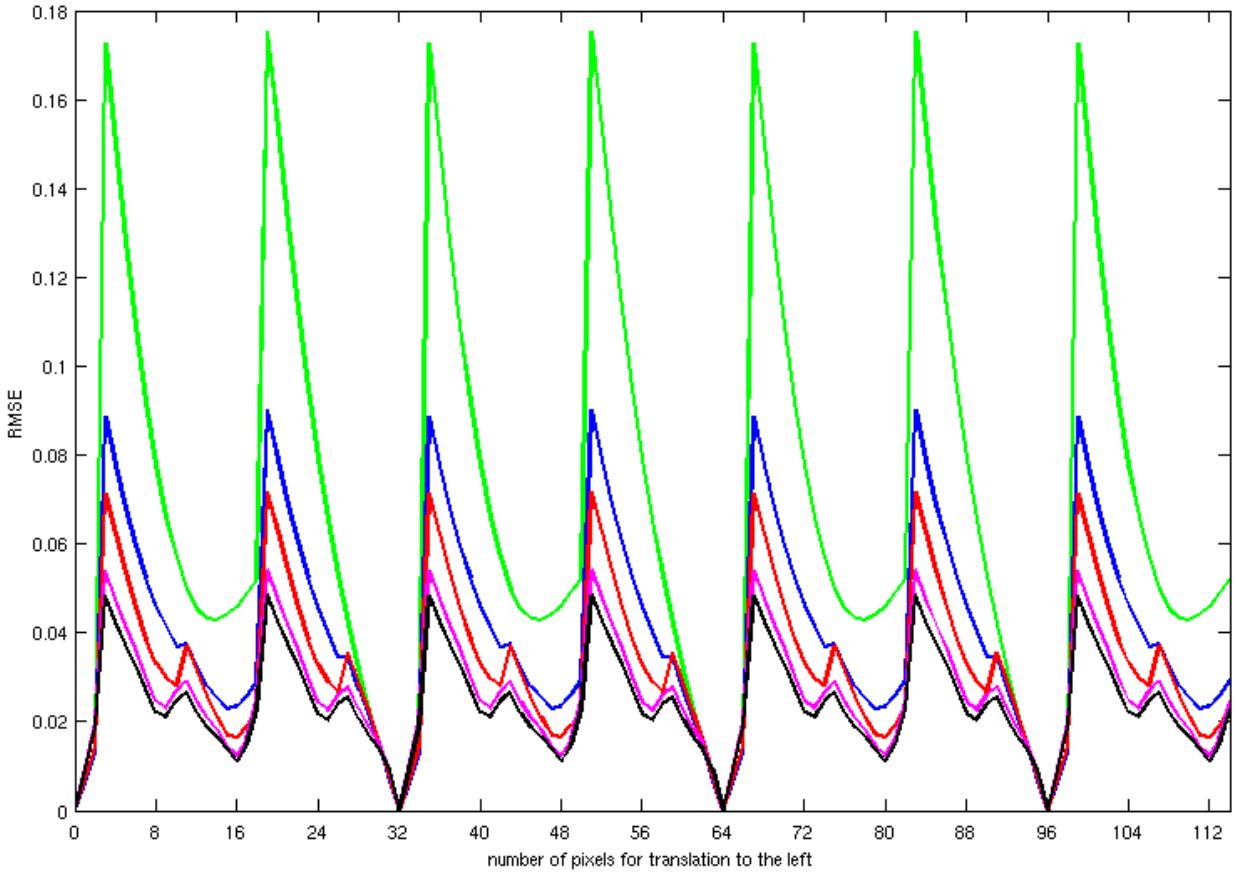

Figure 5.2: Translation to the right. RMSE for descriptors created for 3 (green), 4 (blue), 5 (red), 6 (magenta) and 7 (black) pyramid levels. The highest selected pyramid level for the construction of the descriptors is the 4^{th} from the top (3 levels discarded from the top of the pyramid, according to Chapter 4 in the Laplacian profiles section). Two aspects are the most important. First is using more pyramid levels for the construction of the descriptors, resulting in smaller RMSE, which means more invariance. Second is that the plot repeats every 32 pixels of tranlstion.

According to the RMSE plot, there are two basic effects that need to be commented. The first is that using more pyramid levels for the construction of the descriptors results in a smaller RMSE, which means that the descriptors created using more pyramid levels for translated images are more alike, so more invariant. The second is that RMSE is periodic with a period of 32 pixels of translation. It increases with some fluctuations before starting to decrease until reaching zero and then it repeats the exact same way. A probable explanation is the fact that the descriptor includes elements derived from Radial Fourier transform. Generally, translating a function leaves the magnitude unchanged and adds a constant to the phase. Specifically, For an image of size $M \times N$ pixels, the RDFT repeats itself every M points in the x-direction and every N points in the y-direction [1]. The periodic nature of RDFT forces the RMSE plot to repeat with respect to position. In figure 5.4 the original image from figure 5.3 is translated to a distance that is equal to its width and/or its height times a integer factor.

Then, the representation for each of the images in the frequency space is illustrated. It is obvious from the frequency space representations that the RDFT repeats itself every time the image is translated with an integer product of its dimensions. The same effect occurs when the image is translated and the pyramid is created. Consider the neighbourhoods around Laplacian profile corresponding pixels as little images. Then each one of this little images is translated, slower to higher levels, and the RDFT repeats every time the translation of each one of them completes an integer product of its dimension. The higher the level is, the smallest the effect is to the RDFT as the change is very slow due to the small size of the higher levels. Every time all neighbourhoods in every pyramid level synchronize in completing a translation equal for each to their dimensions, which occurs every 32 pixels for this image case, the error is minimized to zero because the descriptors in the original image and the translated image become equal.

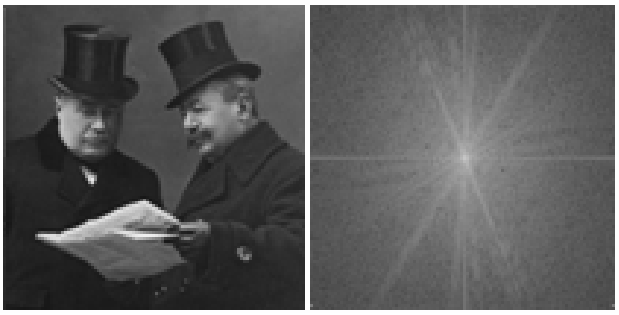

Figure 5.3: The Discrete Fourier transformation (DFT) of an $M \times N$ image. Adapted from [6].

Figure 5.4: At the left side, translation of the $M \times N$ image for a product of M points in the x-direction and for a product of N points in the y-direction. At the right side there are the Discrete Fourier transforms (DFT) of each one of the translated images separately. For each translation of M points in the x-direction and every N points in the y-direction the DFT repeats [1]. Adapted from [6].

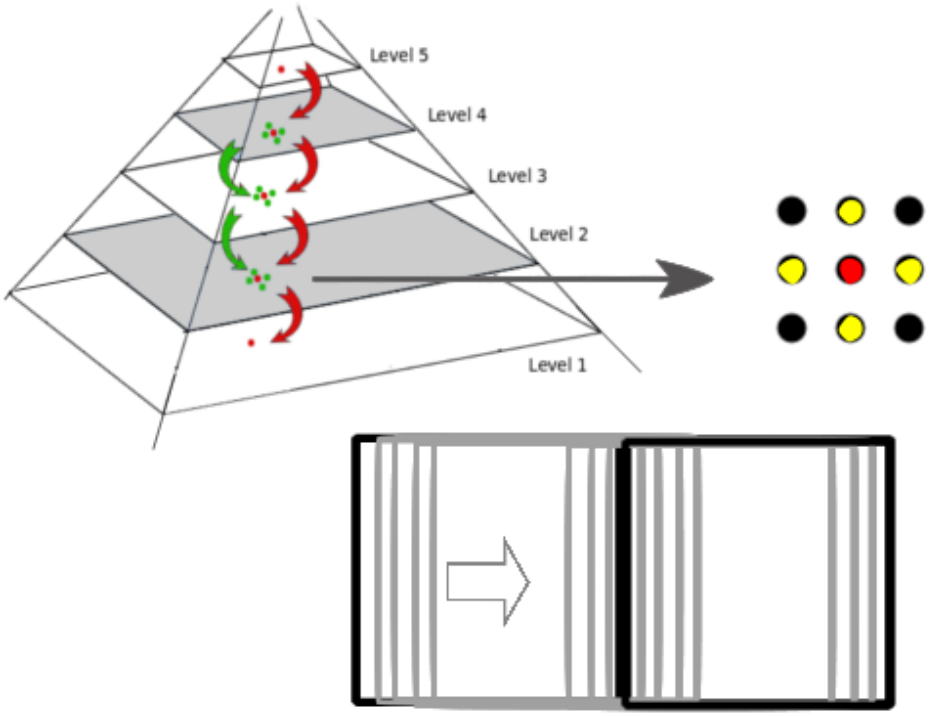

Figure 5.5: Consider the neighbourhoods around Laplacian profile pixels as little images. While the original image translates, the neighbourhoods of the Laplacian profile pixels also translate, slower at higher levels due to their small size. The RDFT repeats every time the translation of each of neighbourhoods completes an integer product of its dimension. Pyramid skeleton adapted from [29].

The repetition occurs every 32 pixels of translation due to our decision to discard the first three levels of the pyramid, for the reasons indicated in Chapter 4 in the Laplacian profiles section. In case there were more levels discarded from the top of the pyramid, then the repetition will occur faster, as the translation of neighbourhoods around Laplacian profile pixels is more significant due to the bigger size of the lower pyramid levels. For example, if 4 levels from the top of the pyramid are discarded, the repetition occurs every 16 pixels of translation, see figure 5.7. The translation this time is downwards. The direction of translation makes no difference to the results. The interval which the repetition occurs is a product of 2 due to the interpolation used with the pyramid which causes each level to have half the samples of each lower level and double the samples of its higher level. The conclusion is that repetition depends on the total number of pyramid levels created from an image, therefore the size of the image is relevant as bigger images create higher pyramids, and the subset of levels selected to create the descriptor can be larger.

To summarize, the descriptors examined show the existence of invariance to translation in a periodic form, as for a fixed interval, different under different circumstances, the RMSE is decreasing to zero. Moreover, the addition of pyramid levels to the construction of the descriptor leads to better results in invariance to translation. Concerning translation to the left and upwards, the same results are obtained, such as for translation to the right or downwards that were already examined.

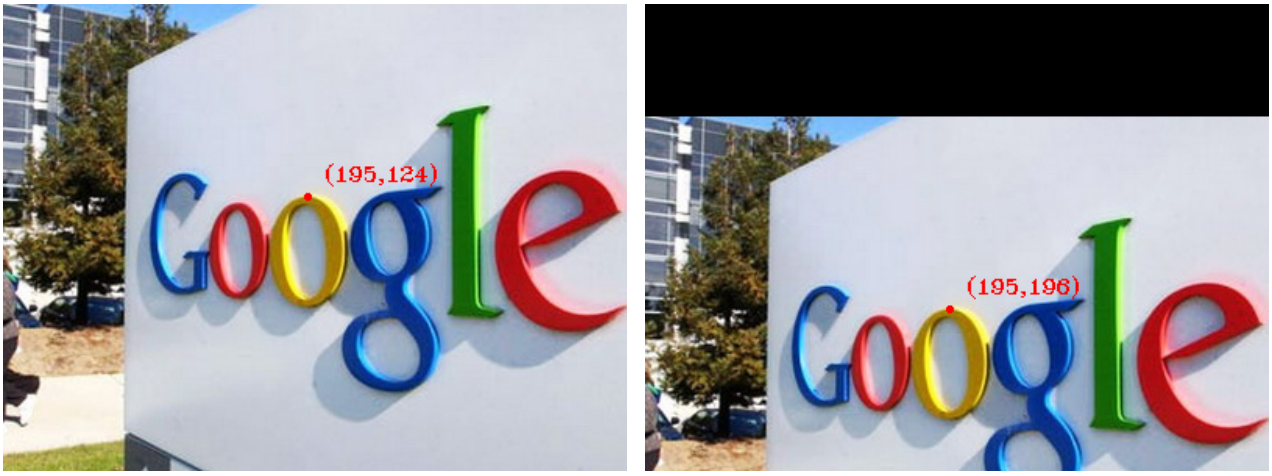

Figure 5.6: Translation 72 pixels downwards [42].

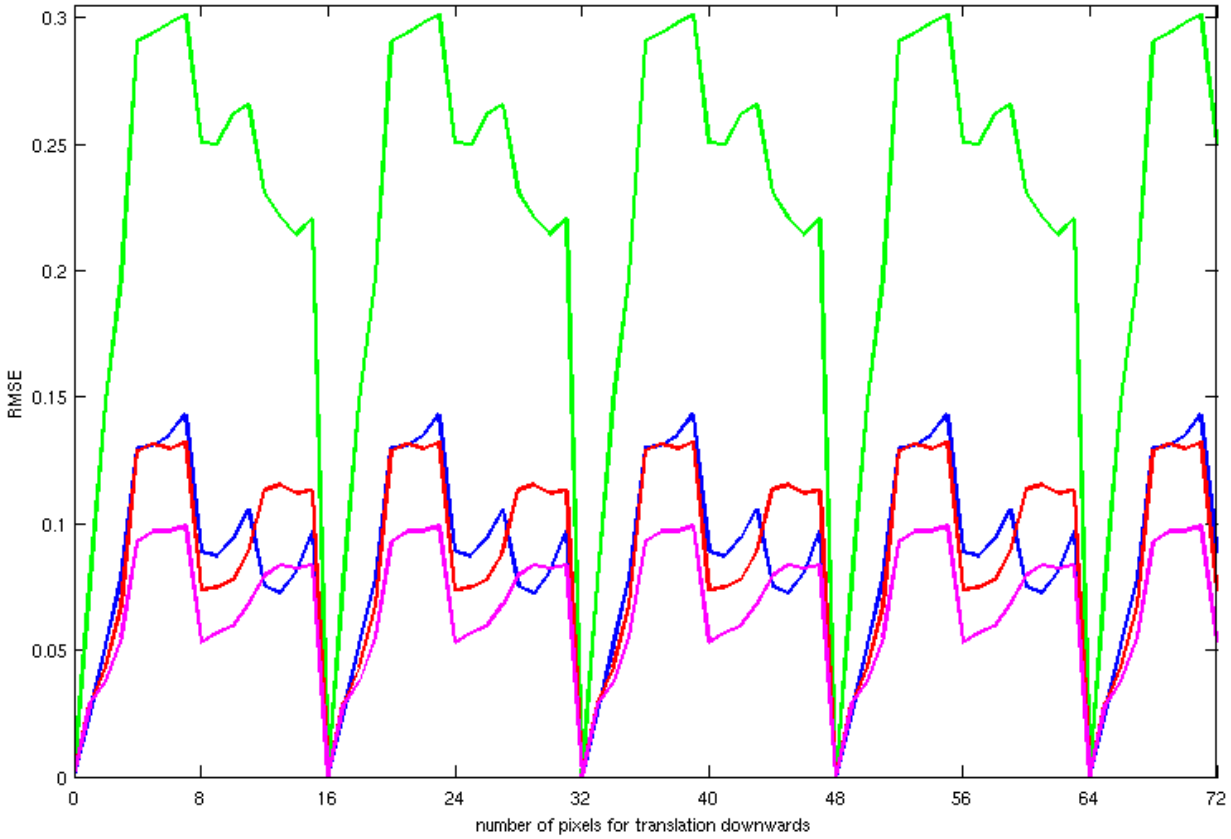

Figure 5.7: Translation downwards. RMSE for descriptors created for 3 (green), 4 (blue), 5 (red) and 6 (magenta) pyramid levels but the higher selected pyramid level for the construction of the descriptors is the 5^{th} from the top (4 levels discarded from the top of the pyramid, according to Chapter 4 in the Laplacian profiles section). Again, using more pyramid levels for the construction of the descriptors results in smaller RMSE, which means more invariance. Now, the plot repeats every 32 pixels of tranlstion.

5.2.2 Rotation

Regarding rotation, figure 5.12 contains results on rotation to the left. The RMSE plot reveals that using more pyramid levels for the construction of the descriptors results in smaller RMSE, which means that the descriptors for rotated images are more similar, so there is more invariance to rotation. Another important aspect is that the RMSE increases as the degrees of rotation increase, but only until a point, and then decreases radically until it reaches zero. This fact attest to rotation invariance of the new method descriptors, as the RMSE reaches a maximum for a large number of rotation degrees and for small rotation the descriptors between rotated images tend not to differ a lot.

Figure 5.8: Original image [35].

Figure 5.9: Rotation 30°.

Figure 5.10: Rotation 180°.

Figure 5.11: Rotation 360°.

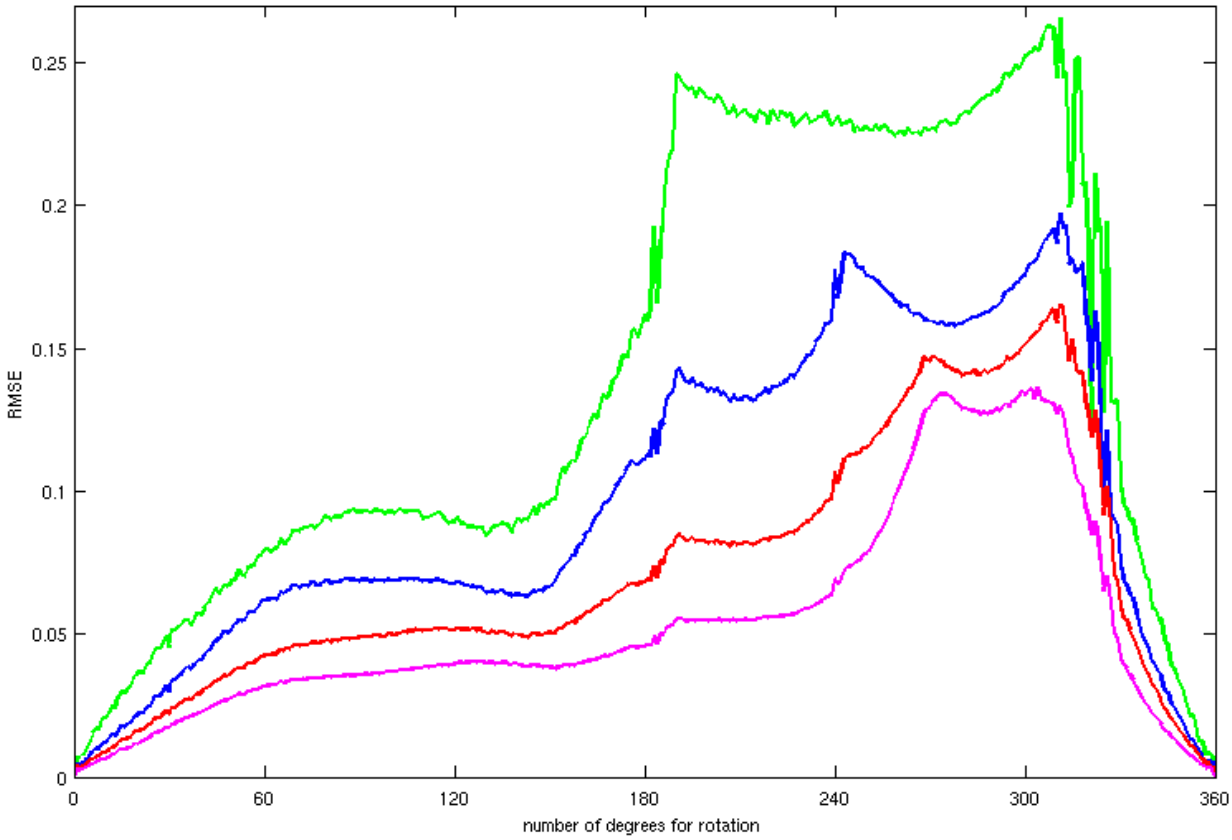

Figure 5.12: Rotation results for 360° to the left. RMSE for descriptors created for 3 (green), 4 (blue), 5 (red) and 6 (magenta) pyramid levels.

Figure 5.12 shows the variation in the descriptor resulting from rotation of the image around the selected point. The lack of symmetry in this case was a surprise. One could assume that the RMSE should have increased evenly until 180° and then should decrease evenly to reach zero at 360°. To assure that the result is not arbitrary, figure 5.13 shows the results of the same test with the same image pixel but with rotation to the right. The result for the rotation in the different direction is the exact symmetric to the central vertical axis (180°) of the result in figure 5.12.

The main reason is that RDFT represents an image as a summation of cosine-like images. The RDFT derived X_k values are complex numbers characterized by their magnitude and phase. As mentioned in Chapter 4, rotation should not effect RDFT. Though, as explained in [2], when rotating an image, new frequencies appear due to the fact that RDFT always treats an image horizontally and vertically. Hense, the rotated image is treated as a different image by the RDFT. The area content that is described around the selected pixel in the image is not symmetrical, so the corresponding area in every rotated image is very different, causing the RDFT X_k values to vary and the RMSE plot to fluctuate.

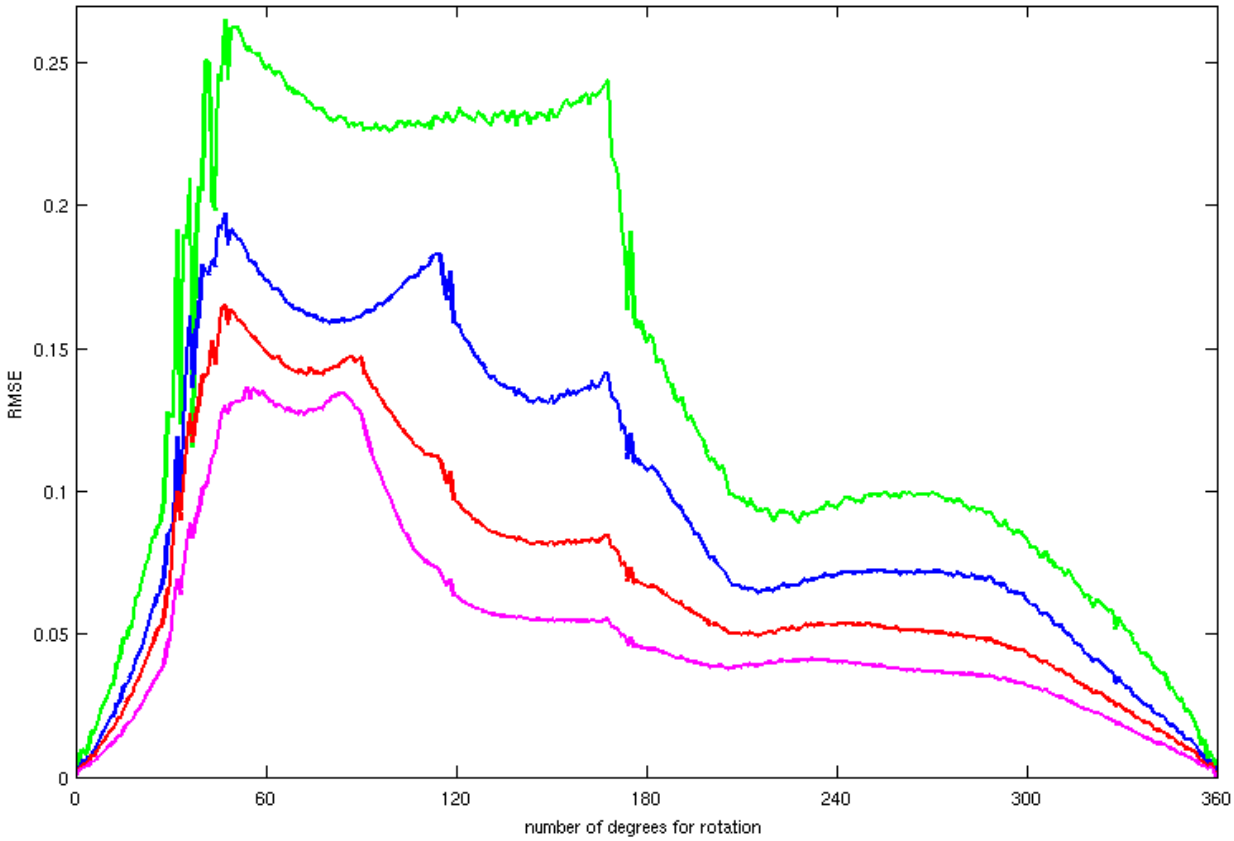

Figure 5.13: Rotation results for 360° to the right. RMSE for descriptors created for 3 (green), 4 (blue), 5 (red) and 6 (magenta) pyramid levels. The result is the exact symmetric to the central vertical axis of the result in figure 5.12.

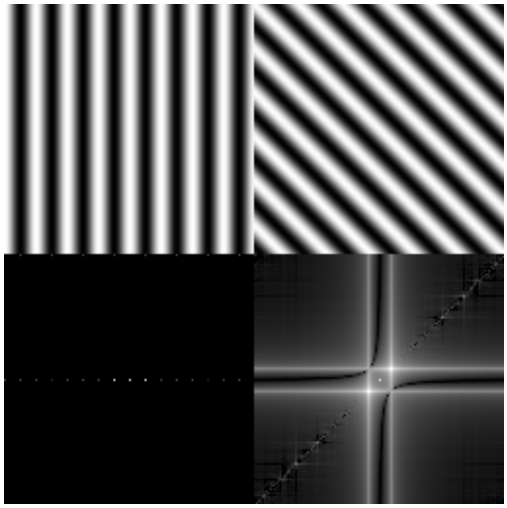

Figure 5.14: The top images displayed are horizontal cosines, the right being the rotation of the left. The bottom images are their Fourier transforms. The rotated cosine has a Fourier transform that is much more complicated, with strong diagonal and plus sign shaped horizontal and vertical components. This occurs because rotating an image causes new frequencies to appear as RDFT always treats an image horizontally and vertically. The rotated image is considered as a different image. Considering that the area described by the descriptor corresponding to a selected pixel is not symmetrical, the RDFT varies in a non fixed way and causes the RMSE plot to not be symmetrical. [2]

In order to validate this explanation, a image with a pattern that is symmetrical every $\frac{1}{3}$ of a circle is rotated to the right for 360°. The image is in figure 5.15 and the RMSE is shown in figure 5.18. A pixel is selected close to the center of the pattern so the neighbourhood around it will repeat with rotation. The RMSE has three peaks which are imposed by the three similar shapes in the image. While the image rotates, the same set of frequencies appear every time the rotation covers 120° and forces RMSE to decrease close to zero.

In conclusion, in order to have an RMSE plot that is increasing evenly until 180° and then decreasing evenly to reach zero at 360°, the described area in the original image must be totally asymmetrical and not repeating. In contrast, if the described area is symmetrical, the symmetry will cause a similar set of frequencies to appear at related moments during rotation and the RMSE to decrease at those moments.

Figure 5.15: Original image. The pattern in the center of the image will repeat every $\frac{1}{3}$ of a circle (120°). The rotation center should be close to the pattern center in order to capture the symmetry. [4]

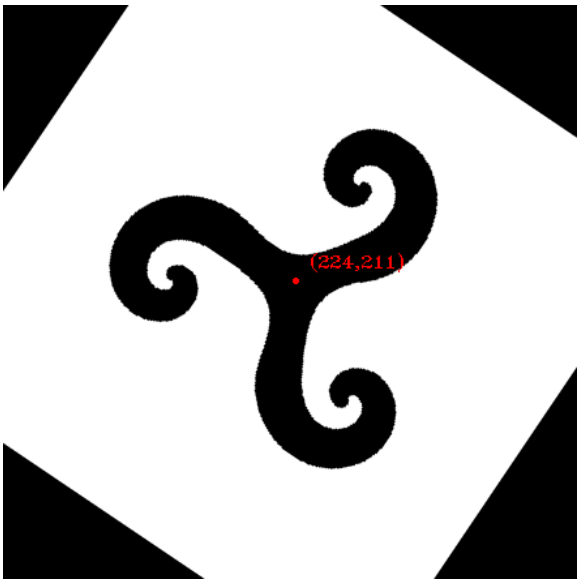

Figure 5.16: Rotation 33°.

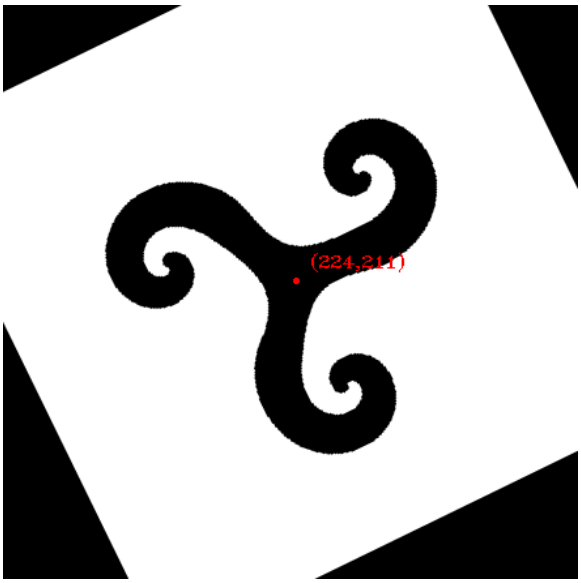

Figure 5.17: Rotation 153°.

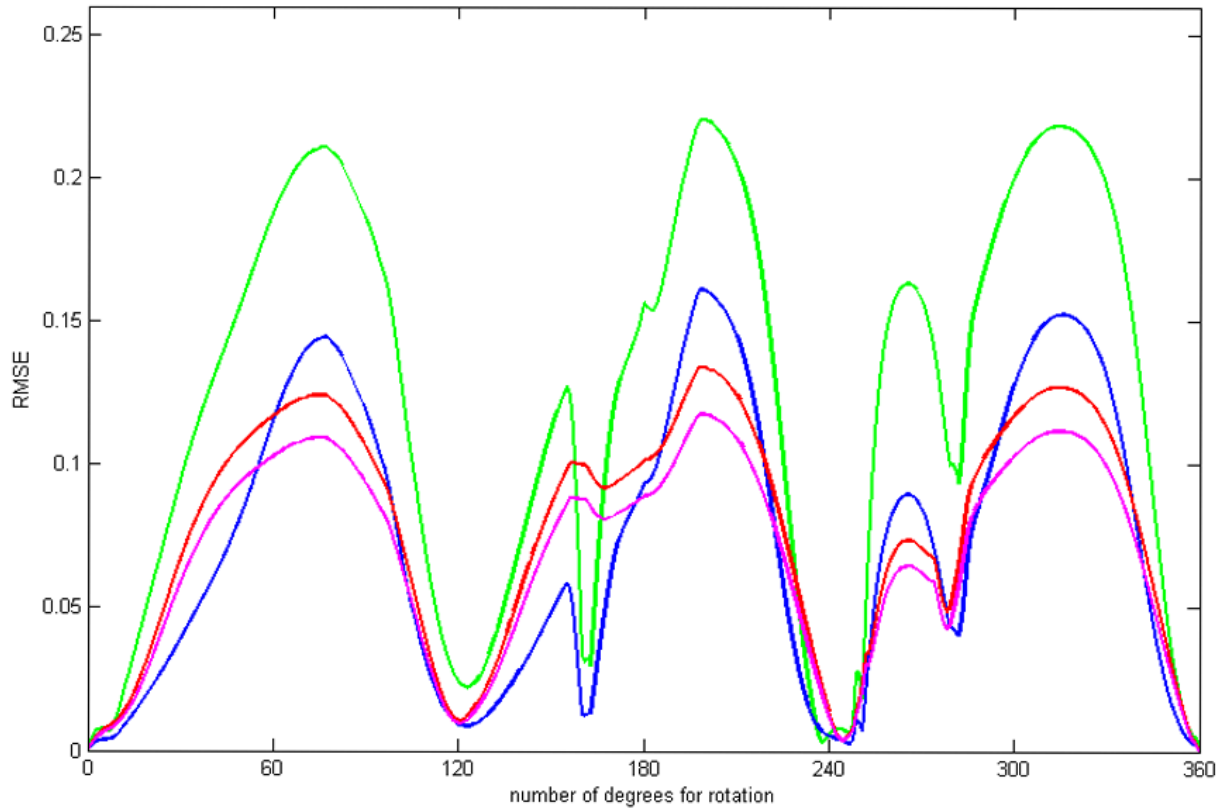

Figure 5.18: Rotation results for 360° to the right. RMSE for descriptors created for 3 (green), 4 (blue), 5 (red) and 6 (magenta) pyramid levels. RMSE repeats approximately every $\frac{1}{3} of a circle$. The small fluctuation in the RMSE occur because the pattern is not exactly identical at each of the three directions and also because the selected pixel is not exactly on the center of the pattern.

5.2.3 Scaling

Invariance to scaling is tested with images scaled with a factor of $\sqrt{2}$, meaning that each scaled image is $\sqrt{2}$ smaller than the previous one. The results of RMSE have again two aspects to be discussed. First, using more pyramid levels for the construction of the descriptors results in smaller RMSE, which means more invariance, exactly as shown for translation and rotation. Second, after a number of scales, the RMSE is stable, earlier for descriptors constructed with less pyramid levels. This occurs because when the image becomes too small after several times of scaling, the constructed pyramid can have too few levels, even just one. Then, the constructed descriptors are actually the same vector of three elements (one Laplacian value of each of the LC_1C_2 channels) and contain no useful information. Consequently, the description of an image scaled to smaller size is not feasible at any scale.

Figure 5.19: Original image [7].

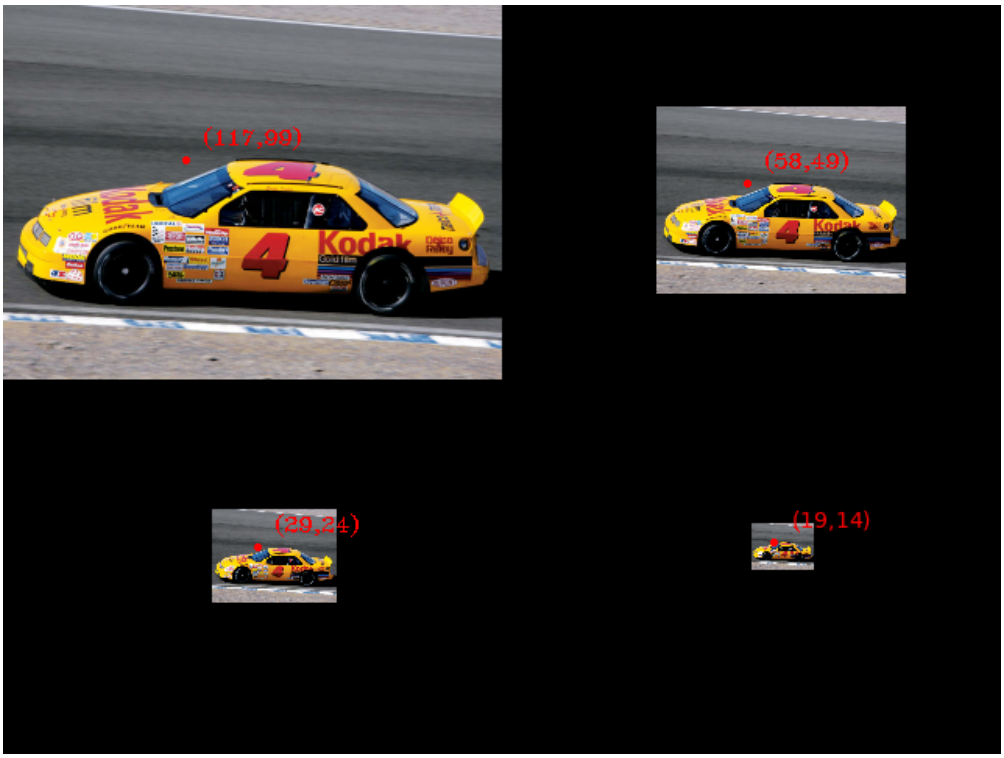

Figure 5.20: Scaled smaller of 2, 4, 6 and 8 times with scaling factor $\sqrt{2}$.

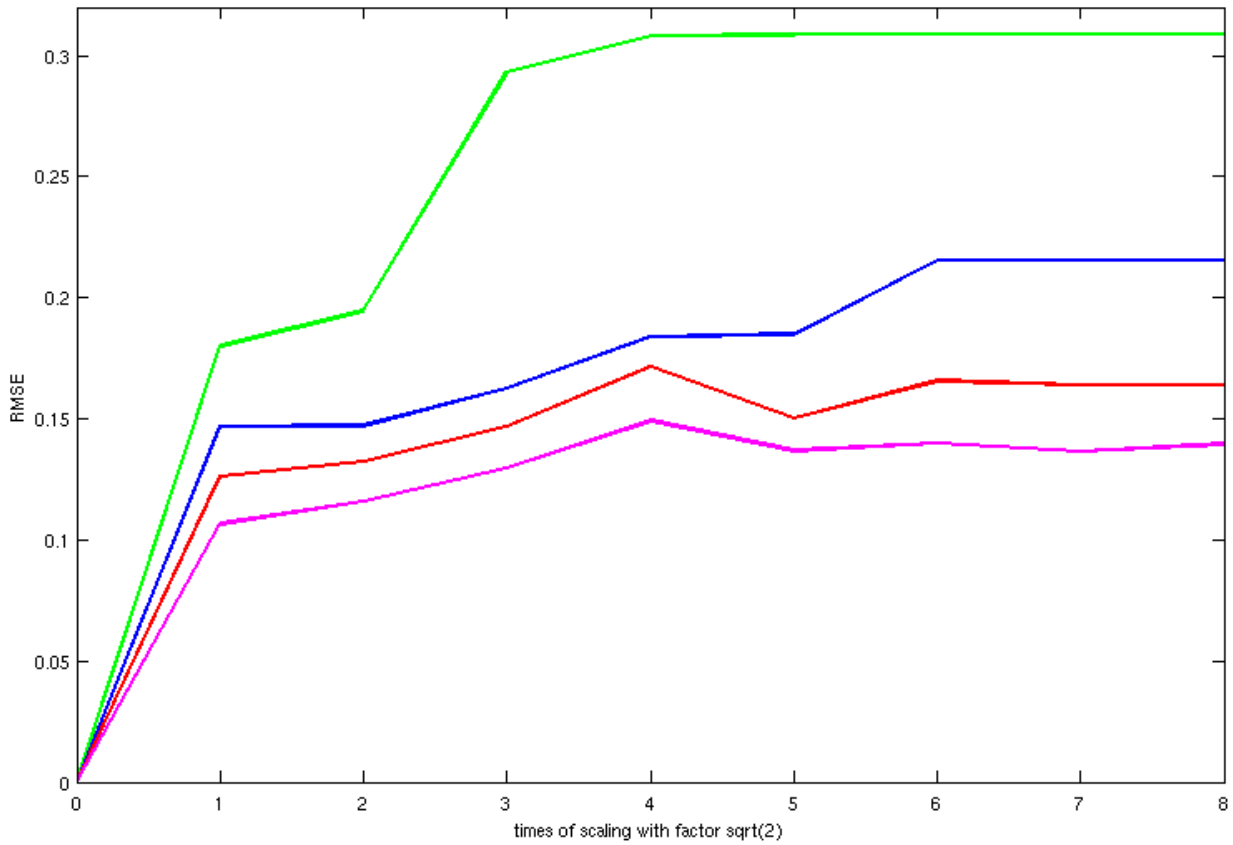

Figure 5.21: Scaling to smaller images with scaling factor $\sqrt{2}$. RMSE for descriptors created for 3 (green), 4 (blue), 5 (red) and 6 (magenta) pyramid levels.

Concerning scaling to larger images, there is no limit of scaling apart from hardware constrains. The RMSE increases at each scaled image until it reaches a point that is almost stable. This is because for a very large scaled image the interpolated corresponding pixel local neighbourhood does not change significantly after some time.

Figure 5.22: Enlarged image with scaling factor $\sqrt{2}$. The original is in figure 3.1.

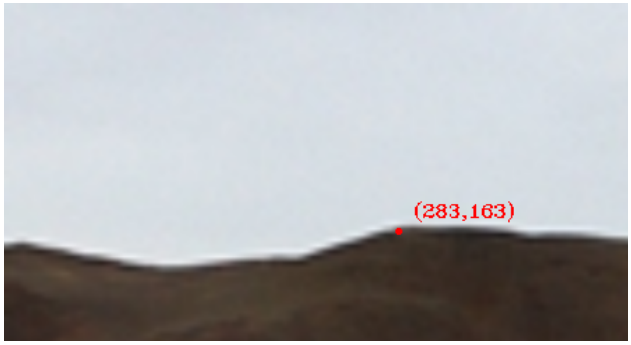

Figure 5.23: 3 times larger with scaling factor $\sqrt{2}$.

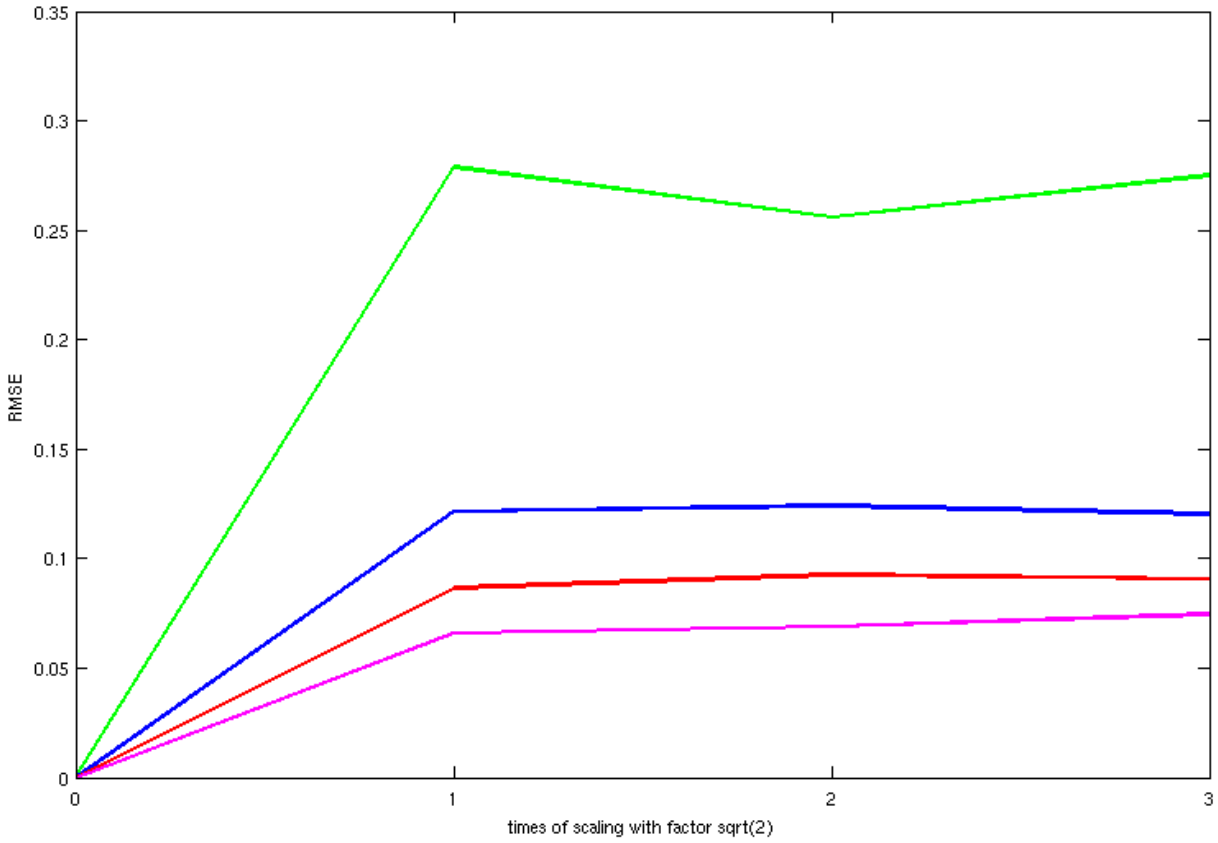

Figure 5.24: Scaling to larger images with scaling factor $\sqrt{2}$. RMSE for descriptors created for 3 (green), 4 (blue), 5 (red) and 6 (magenta) pyramid levels.

Chapter 6

Evaluation of discrimination

6.1 Experiments

6.1.1 Experimental hypothesis on discrimination

A second round of experiments was designed to test the discriminative power of the proposed method. The expectation for a discriminative power lays in the use of Laplacian values, which can capture local image extrema, and the use of the Radial Fourier transform of local areas, which can provide an appearance identification for these areas. It is expected that results can improve with the addition of pyramid levels in the construction of the descriptors. Depending on this facts, it is expected that the proposed method can be discriminative.

6.1.2 Initial experimental evaluation of discrimination

The discrimination power of the proposed method has been tested on the INRIA Person Dataset by Dalal [17]. Human detection is a well examined problem in computer vision and the INRIA Person Dataset is a widely used database for this purpose. Therefore, there is assurance that this database is efficient for dense description testing. Consequently, it was wise to use this database for the first experimentations of the proposed method.

LIBLINEAR [5] was used on descriptors created densely on 2416 positive images and 4832 negative images, all of size 96×160. Positive refers to images containing humans and negative to images not containing any humans. The size of the images, 96×160 pixels, allows the creation of 9 level pyramids using the implementation of the Half-Octave Gaussian pyramid by Combe [10]. Possible meaningful descriptor vectors can be made starting from the 7^{th} level of the Gaussian pyramid and be constructed for either 3, 4, 5 or 6 levels of the Gaussian pyramid or starting from the 6^{th} level of the Gaussian pyramid and be constructed for either 3, 4 or 5 levels. The reason is that, from the 9 levels of the pyramid, the 2 or 3 highest levels and the bottom level were abandoned and only levels between the 7^{th} or 6^{th} and the 2^{nd} were taken

into account, as indicated in Chapter 4 in the Laplacian profiles section. Therefore, local descriptor vectors per highest selected level pixel can occur having one of the different sizes shown to the table below:

Chosen levels	Descriptor vector size
from 7 to 5 (3 levels)	$(3 + (1 \times 4)) \times 3 = 21$
from 7 to 4 (4 levels)	$(4 + (2 \times 4)) \times 3 = 36$
from 7 to 3 (5 levels)	$(5 + (3 \times 4)) \times 3 = 51$
from 7 to 2 (6 levels)	$(6 + (4 \times 4)) \times 3 = 66$
from 6 to 4 (3 levels)	$(3 + (1 \times 4)) \times 3 = 21$
from 6 to 3 (4 levels)	$(4 + (2 \times 4)) \times 3 = 36$
from 6 to 2 (5 levels)	$(5 + (3 \times 4)) \times 3 = 51$

Table 6.1: Descriptor creation starting from the 7^{th} or the 6^{th} level. This size corresponds to only one pixel position from the highest selected pyramid level, here the 7^{th} or the 6^{th}. The right column shows how the descriptor size occurs. For each set of selected levels there is the same number of Laplacian profile values plus 4 values per each middle selected level. The descriptor vectors are created for a all 3 channels of the image L, C_1 and C_2. For more details, consult Chapter 4.

For a dense image description, a large descriptor vector is created from the combination of all local descriptors, each of them corresponding to one pixel position of the highest selected level of the pyramid. The combination is simple: while scanning the highest selected pyramid level, computed descriptor vectors at each pixel position are stacked linearly in a common vector. At the two higher middle selected levels of the pyramids (see Chapter 4 in the Laplacian profiles section), the four pixel neighbourhoods around the Laplacian profiles (see Chapter 4 in the Radial Fourier transform section) overlap as they are too close together. The final vector is a global image descriptor for a given image and its size is (x-dimension of the highest selected level) $\times$ (y-dimension of the highest selected level) $\times$ (the size of a local descriptor vector).

The global image descriptors for all 2416 positive images and 4832 negative images created different SVM models, one for each different case of table 6.1. Six images of various dimension containing human figures were used as testset. These images come with groundtruth accompanying files. Groundtruth is a term used to express the real content of images according particular classes of objects or structures. For example, the groundtruth for images containing humans is the representation of the exact number, positions and sizes of humans in each image. In this project the groundtruth for images is the top-left and bottom-right coordinates of rectangles surrounding humans in the testset images. This groundtruth was created by Dalal and Triggs and published with their paper on human detection with HOG [19]. The procedure for testing discrimination includes selecting patches from a testset image of dimension 96×160. The window moves 8 pixels per time either for the x or the y coordinate. The patches are also collected to different scales of the test image. After the image is scanned completely, it is resized with a factor of 95% and a new set of patches is collected from the resized image. Then the image is resized again and the patch collection

repeats, until the resized image size becomes smaller than the patch size. What this particular procedure actually does is to create a database out of images (meaning the patches) of positive and negative content. Every patch is examined by extracting its global descriptor and classifying it by one of the SVM models, depending on the size of the local descriptors (according to the number of selected pyramid levels) and the higher pyramid level used for its construction. When patches are identified as positive, they are compared to the groundtruth of the image. If a positive classified patch matches at least 60% percent with the image groundtruth, then the result counts as True Positive (TP). In the opposite case, it is considered a False Positive (FP). The performance is checked by three measures which are Detection Rate (DR), Error Rate (ER) and False Positive Per Window (FPPW). Their definitions are presented in the next three formulas:

$$DR = \frac{\text{True Positive Patches}}{Groundtruth} \tag{6.1}$$

$$ER = \frac{\text{Groundtruth - True Positive Patches}}{\text{Groundtruth}} \tag{6.2}$$

$$FPPW = \frac{\text{False Positive Patches}}{\text{Number of Created Windows}} \tag{6.3}$$

where True Positive Patches refer to all created patches identified positive correctly, False Positive Patches refer to all patches mistaken for positive and Number of Created Windows refers to all patches created for an image in all scales while checked with a particular SVM model. The SVM models used were Soft-Margin with $C = 0.01$. The selection of C was made after a series of testing and cross-validation. Though, it was expected that this value will be the choice as it was proposed before in [18] and [36] where the same dataset was used. The seven models showed the maximum accuracy in comparison to models for a different C value. Table 6.2 lists the 5-fold cross-validation accuracy values per model by the number of levels used in the descriptors that were used to train the models. According to this table, performance improves with the addition of pyramid levels to the construction of the descriptors.

Used levels	Accuracy
from 7 to 5 (3 levels)	96.1507%
from 7 to 4 (4 levels)	97.42%
from 7 to 3 (5 levels)	97.42%
from 7 to 2 (6 levels)	97.9857%
from 6 to 4 (3 levels)	96.3852%
from 6 to 3 (4 levels)	97.4614%
from 6 to 2 (5 levels)	97.9443%

Table 6.2: 5-fold cross-validation accuracy of SVM models. The column *Used levels* indicates the highest level and the number of pyramid levels used for the creation of the descriptors that the SVM models were trained with.

6.2 Results

The first experimental results are very modest but attest to the existence of a discrimination character in the new method. According to DR and ER, the results for descriptors from the 7^{th} pyramid level are better in comparison to the results for descriptors from the 6^{th} pyramid level. Remember that the higher the pyramid level is, the less pixels it has, so the less local descriptors are created covering bigger parts on the original image. DR and ER for the descriptors from the 7^{th} pyramid level are stable regardless of the amount of used levels in the construction of the vectors. The FPPW plot in figure 6.4 decreases with the addition of levels, which shows that the performance improves as it becomes more accurate.

The unexpected turnout is that for descriptors from the 6^{th} pyramid level, DR decreases, so ER increases, with the addition of pyramid levels in the construction of the vectors. The addition of pyramid levels used for the construction of the descriptors was expected to have a positive effect on performance. This fact indicates that larger descriptor vectors do not necessarily perform better as the success of description can also lay on the manner information is organised in the vectors. The organization of information in the descriptor vectors refers to the vector element values, for example Laplacian values and RDFT information, and their size, which in our case depends on the pyramid levels used. A more accurate explanation is that, because the starting pyramid level for the descriptor construction is too low, so it has bigger dimensions but not enough lower levels underneath, the created descriptors are too many and correspond to smaller areas on the original image, making the global image descriptor too noisy and less descriptive. Apparently, less descriptor vectors that cover larger areas on the original image can be more informative.

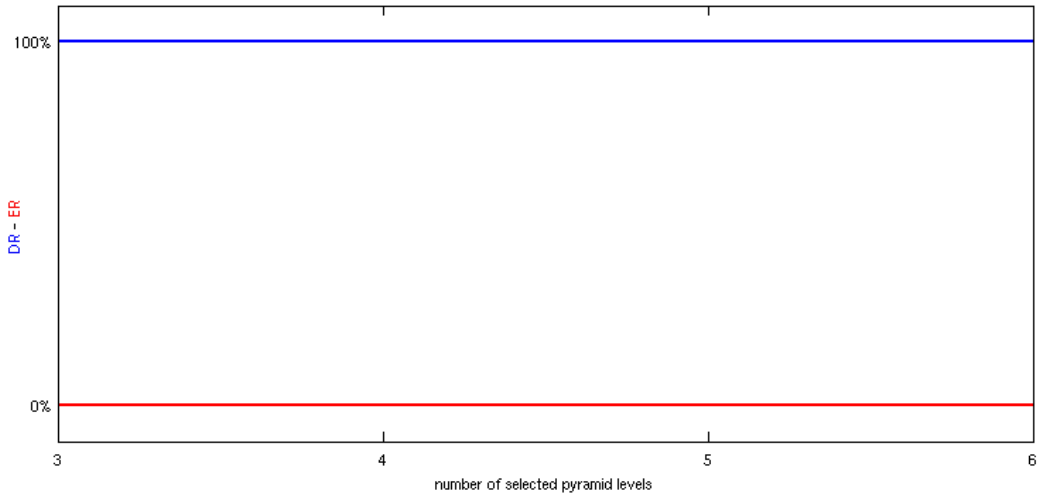

Figure 6.1: Plot of Detection Rate and Error Rate on different descriptor size, depending on used pyramid levels, starting from 7^{th} pyramid level. All different size descriptor SVM models detected successfully every human in the images. Detection Rate is plotted in blue and Error Rate in red.

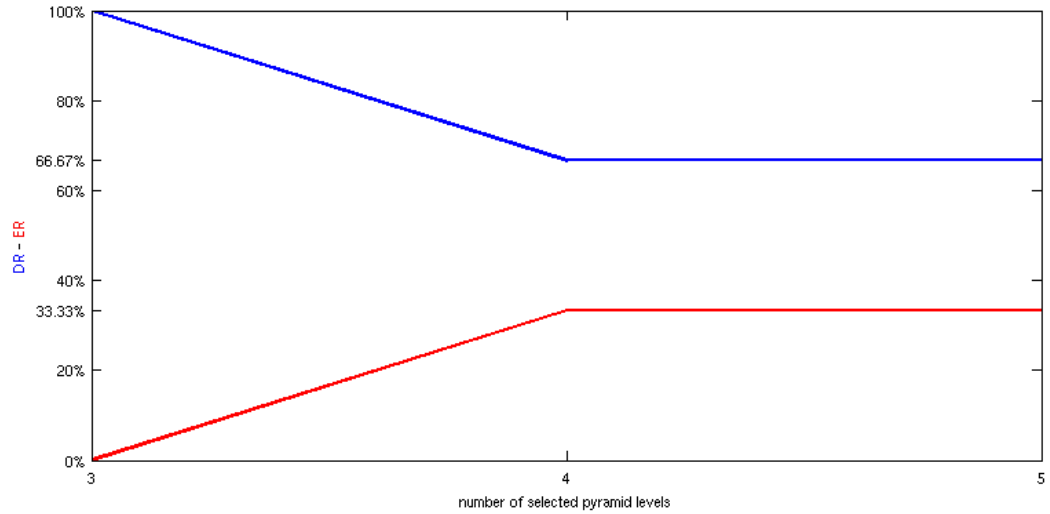

Figure 6.2: Plot of Detection Rate and Error Rate on different descriptor size starting from 6^{th} pyramid level. Only the SVM model of descriptors made from 3 pyramid levels detected successfully every human in the images. The other two models detected four out of a total of six people in the images. Detection Rate is plotted in blue and Error Rate in red.

On the other hand, the FPPW indicates that descriptors created staring from the 6^{th} pyramid level stand less chances to make a mistake. This fact in relation to DR and ER indicate that descriptors from the 6^{th} pyramid level are weaker in identifying humans than descriptors from the 7^{th} pyramid level but offer more certainty of the result. The assumption taken is that a larger number of local descriptors combined into the global descriptor can provide more accurate detail on the image appearance. Moreover, FPPW shows the behaviour that was expected, decreasing in both cases with the addition of pyramid levels to the creation of the vectors, which means that the performance improves by becoming more accurate.

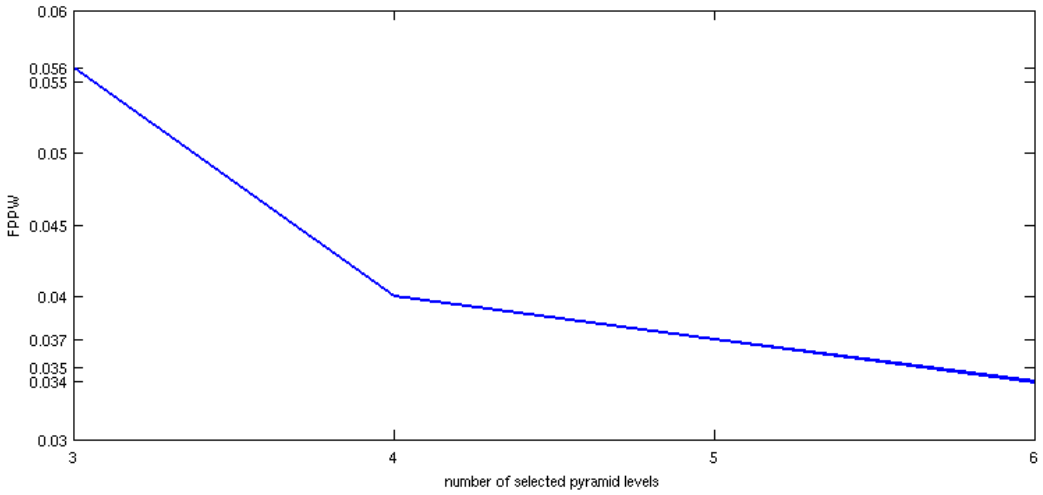

Figure 6.3: Plot of False Positives on different descriptor sizes starting from the 7^{th} pyramid level. It decreases with the addition of pyramid levels in the construction of the descriptors.

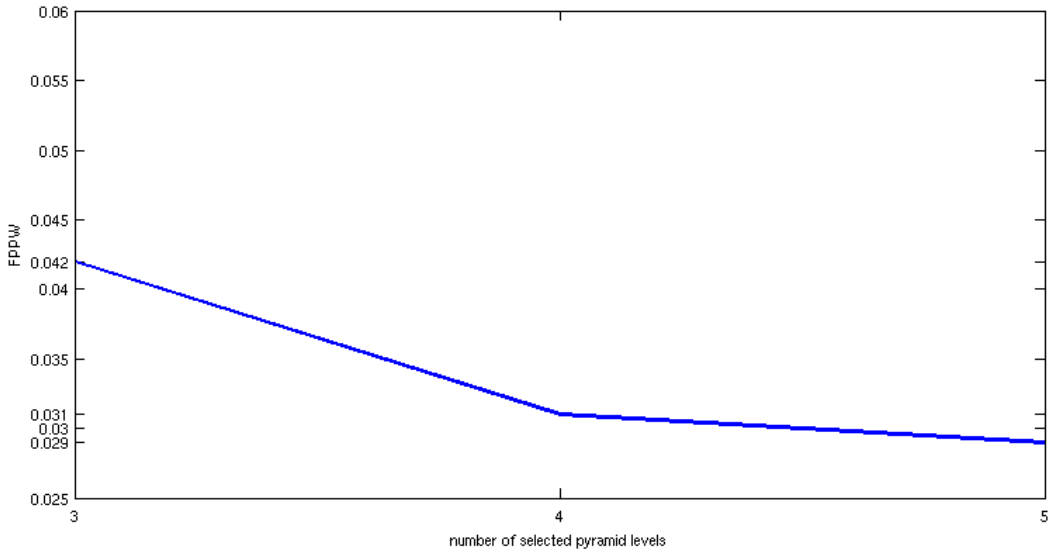

Figure 6.4: Plot of False Positives on different descriptor sizes starting from the 6^{th} pyramid level. It decreases with the addition of pyramid levels in the construction of the descriptors.

The DR and ER contradict the calculated 5-fold cross validation accuracy on the SVM models. According to table 6.2, DR must increase with the addition of pyramid levels in the construction of the vectors, while ER must decrease. The difference between the two facts could most probably be explain by the data distribution. Therefore, tests on more testset images were done using the same SVM models or other SVM models made from the same trainset but with different C parameter values. All had similar results as the above, contributing to the belief that descriptors from the 7^{th} pyramid level have better DR but descriptors from the 6^{th} pyramid level have better FPPW. To conclude, results show that the starting level of the descriptors is as important as the number of levels used for the construction of the vectors.

Chapter 7

Conclusion

7.1 Lessons learned

In this report we have provided initial results with experimental evaluation of a new invariant image descriptor. While these results are encouraging, much remain to be done. According to the first experimental results, invariance to translation, rotation and scaling and discrimination power exist for the new description method. The proposed descriptors can combine dense description with translation and scale invariance and especially invariance to rotation which is hard for dense description. A basic fact demonstrated is that the strength of descriptors generally improve with the addition of pyramid levels in computations. Comparing to the experiments conducted before including Radial Fourier transforms in the process (they are not included in this report), Radial Fourier transforms provided significantly better performance without using a larger number of elements in the descriptor vector. The conclusion of this project is that this method has a promising future.

7.2 Discussion and Future work

An important issue to be examined further in the future is the number of neighbours used around the Laplacian profiles. It has been shown in the previous Chapters that the Radial Fourier transform of four neighbours around the Laplacian profile corresponding pixels in middle levels is a good approach as the magnitude and phase provided meaningful details on the neighbourhood appearance regarding translation and rotation. Though, a combination of all eight neighbours could have a more powerful result. Moreover, the neighbourhood does not necessarily need to be fixed for every level. The number of neighbours can grow larger while descending the pyramid levels in a shape, for example in a circle or rectangle, and the neighbours to be used can belong to the periphery or the whole area of the shape.

Another important avenue for further research is the creation of descriptors in a tree structure following the four neighbours of Laplacian profile values to lower levels. This would capture the appearance of the

image in larger neighbourhoods while the size of the levels increases by descending the pyramid, which can result as expected to more efficient description of the local area. There are different ways this idea can be implemented. One way is to consider pixel neighbourhoods corresponding to the four neighbours of the Laplacian profile of the level above and every pixel of this neighbourhoods would initiate another set of neighbourhoods to the level below. A second way is to consider pixel neighbourhoods corresponding to the four neighbours of the Laplacian profile of the level above that don't further initiate new neighbourhoods to lower levels. Also, the manipulation of the new neighbourhoods can vary. One probable way is Radial Fourier transforms and another is to just take the Laplacian values.

The size of the descriptor varies with the length of the Laplacian profile. Larger descriptor sizes were found to have better results but there is the question of memory constraint. The descriptor size against the availability of memory is an important issue. The Principal component analysis method (PCA) can help explore more accurately the limits of this trade-off. Another matter is the structure of the descriptor. The complex number X_0, derived from the Radial Fourier transforms in Chapter 4, was not used for the construction of the descriptors. Perhaps, including the value of X_0 in the descriptor can have an effect on performance. Invariance to other image transformations except translation, rotation and scaling, such as illumination changes and affine transformations, must also be examined. The method should be enhanced to be efficient for any kind of image transformation to work in the best possible way.

The new proposed method, after reaching a satisfactory level of performance, must be compared with the state of the art existing algorithms, for instance HOG. Theoretically, it is expected to have results as good as the state of the art methods, although many aspects concerning performance remain to be examined.

The final purpose of the future plans is to eventually test the method on publicity logo indexing and recognition. Of course, the method can be used for other purposes, though the ultimate goal is to be efficient for difficult datasets such as logos.

Bibliography

[1] Image Enhancement in the Frequency Domain. Available at `baggins.nottingham.edu.my/~hsooihock/G52IIP/Introduction.ppt`. Introduction to Image Processing Course, University of Nottingham.

[2] INTRODUCTION TO FOURIER TRANSFORMS FOR IMAGE PROCESSING. Available at `http://www.cs.unm.edu/~brayer/vision/fourier.html`.

[3] Opencv 2.0. Available at `http://opencv.willowgarage.com/documentation/index.html`, 2009.

[4] Triskele. Available at `http://twistedone151.wordpress.com/tag/symmetry/`, June 2009. Physics Friday 78, Twisted One 151's Weblog.

[5] LIBLINEAR. Available at `http://www.csie.ntu.edu.tw/~cjlin/liblinear/`, April 2011. A Library for Large Linear Classification, Version 1.8.

[6] AllFreeware. 2d FFT/iFFT plugin. Available at `http://www.all-freeware.com/details/44603/2d-fft-ifft-plugin.html`.

[7] J. Boone. Racing car. Available at `http://www.stockcarracing.com/featurestories/scrp_0709_historic_stock_car_racing/photo_03.html`, February 2009.

[8] G. J. Burghouts and J. Geusebroek. Performance evaluation of local colour invariants. *cvia*, 113:48–62, 2009.

[9] P. Burt and E. Adelson. The laplacian pyramid as a compact image code. IEEE *Transactions on Pattern Analysis and Machine Intelligence*, 31:532–540, 1983.

[10] C. Combe. Half-octave gaussian pyramid implementation, 2011. for the purposes of the MinImage project.

[11] C. Cortes and V. Vapnik. Support-Vector Networks. *Machine Learning*, 20(3):273–297, 1995.

[12] J. Crowley, F. Devernay, F. Huguet, J. Ruiz-Hernandez, and A. Lux. State of the art in real-time view-invariant detection and recognition. Technical Report 1.4, INRIA Grenoble Rhône Alpes Centre de Recherche, mar 2009. Contribution to MinImage, Deliverable Report 4.1.1.3, Analyse Faisabilité / Etat de l'art.

[13] J. Crowley, A. Meler, J. Ruiz-Hernandez, C. Combeand, and A. Lux. Results of algorithmic analysis for real-time invariant image description. Technical Report 2.2, INRIA Grenoble Rhône Alpes Centre de Recherche, April 2009. Contribution to MinImage, Deliverable Report 4.1.2.2, Analyse Algorithmique.

[14] J. L. Crowley and A. Parker. A representation for shape based on peaks and ridges in the difference of low pass transform. IEEE *Transactions on Pattern Analysis and Machine Intelligence*, 6(2):156–170, March 1984.

[15] J. L. Crowley and O. Riff. Fast computation of characteristic scale using a half octave pyramid. volume 31, Isle of Skye, Scotland, UK, 2003.

[16] J. L. Crowley and R. M. Stern. Fast computation of the difference of Low-Pass Transform. IEEE *Transactions on Pattern Analysis and Machine Intelligence*, 6(2), March 1984.

[17] N. Dalal. Inria Person Dataset. Available at `http://pascal.inrialpes.fr/data/human/`, 2005.

[18] N. Dalal. *Finding people in images and videos*. PhD thesis, Institut National Polytechnique de Grenoble, July 2006.

[19] N. Dalal and B. Triggs. Histograms of Oriented Gradients for Human Detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, INRIA Rhˆ ne-Alps, 655 avenue de l'Europe, Montbonnot 38334, France, 2005.

[20] N. Dalal, B. Triggs, and C. Schmid. Histogram of Oriented Gradients (HOG) for object detection. Available at `vision.stanford.edu/teaching/cs223b_winter0910/lecture/HoG.pdf`, February 2010.

[21] G. Dorkó and C. Schmid. Selection of scale-invariant parts for Object Class Recognition. In *9th International Conference on Computer Vision*, volume 1, pages 634–640, Nice, France, October 2003. IEEE.

[22] M. Douze, H. Jégou, H. Sandhawalia, L. Amsaleg, and C. Schmid. Evaluation of GIST descriptors for web-scale image search. In *International Conference on Image and Video Retrieval*, July 2009.

[23] N. Drakos. Difference of Gaussian (DoG). Available at `http://fourier.eng.hmc.edu/e161/lectures/gradient/node11.html`, 1993.

[24] N. Drakos. Laplacian of Gaussian (LoG). Available at `http://fourier.eng.hmc.edu/e161/lectures/gradient/node10.html`, 1993.

[25] J. Geusebroek, R. van den Boomgaard, A. Smeulders, and H. Geerts. Color invariance. *ieeetpami*, 23(12), dec 2001.

[26] T. Gevers and A. Smeulders. Color-based object recognition. *patrec*, 32:453–464, 1999.

[27] D. Hall and J. Crowley. Face detection by robust generic features computed from luminance. *Reconnaissance des Formes et Intelligence Artificiel (RFIA)*, 2004.

[28] M. R. Heinen and P. M. Enge. A computational attention model for robot vision. *Journal of the Brazilian Computer Society*, 15(3), September 2009.

[29] IDL. Object Programming: Working with Image Objects. Available at `http://idlastro.gsfc.nasa.gov/idl_html_help/Image_Tiling.html`, March 2007. IDL Online Help.

[30] H. Jhuang and S. Chikkerur. Video Shot Boundary detection Using Gist. In *TRECVID Workshop*, 2002.

[31] D. Lowe. Object recognition from local scale-invariant features. In *Proceedings of the 7th IEEE International Conference on Computer Vision*, pages 1150–1157, Kerkyra, Greece, August 1999.

[32] C. Murray. Georaster Overview and Concepts. Available at `http://unixservers.eu/documentation/oracle/warehousebuilder/10.2/appdev.102/b14254/geor_intro.htm`, November 2005. Oracle Spatial GeoRaster, 10g Release 2 (10.2), Part Number B14254-02.

[33] A. Oliva and P. G. Schyns. Diagnostic colors mediate scene recognition. *Cognitive Psychology*, 41:176–210, 2000.

[34] A. Oliva and A. Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *International Journal of Computer Vision*, 42(3):145–175, 2001.

[35] Pepe's-Chile. The coca-cola logo. Available at `http://www.joeskitchen.com/chile/2009/02/03/world%E2%80%99s-largest-coca-cola-logo-in-northern-chile/`, February 2009. World Largest Coca-Cola Logo in Northern Chile, In honor of Coca-Cola's 100th anniversary which was celebrated in 1986.

[36] I. Rishabh and A. Satish. Human Detection in RGB Images. Irvine, CA - 92612.

[37] B. Schiele and J. Crowley. Recognition without correspondence using multidimensional receptive field histograms. *International Journal of Computer Vision*, 36(1):31–50, 2000.

[38] B. Schiele and J. L. Crowley. Object recognition using multidimensional receptive field histograms. In *Fourth European Conference on Computer Vision*, April 1999.

[39] P. Shih and C. Liu. Evolving effective color features for improving frgc baseline performance. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 156–163, 2005.

[40] A. Sukcharoenpong, J. Jiang, and M. Al-Shahri. Modeling the World from Internet Photo Collections. Available at `http://dpl.ceegs.ohio-state.edu/courses/GeodSci830/2009spring/Final/Group1/`. GS830: Advanced Methods of Processing Digital Imagery in Photogrammetry.

[41] M. Swain and D. Ballard. Color indexing. *International Journal of Computer Vision*, 7(1):11–32, 1991.

[42] Topnews.us. Google logo. Available at `http://topnews.us/images/google-logo_6.jpg`.

[43] A. Vedaldi and B. Fulkerson. The vlfeat open source library. Available at `http://www.vlfeat.org/index.html`, 2005-2010.

[44] M. Villamizar, J. Scandaliaris, A. Sanfeliu, and J. Andrade-Cetto. Combining Color-Based Invariant Gradient Detector with HOG Descriptors for Robust Image Detection in Scenes under Cast Shadows. Available at `http://dx.doi.org/10.1109/ROBOT.2009.5152429`, 2009.

[45] Wikipedia. Discrete Fourier transform. Available at `http://en.wikipedia.org/wiki/Discrete_Fourier_transform`.

[46] Wikipedia. Euler's formula. Available at `http://en.wikipedia.org/wiki/Euler's_formula`.

[47] Wikipedia. Fourier transform. Available at `http://en.wikipedia.org/wiki/Fourier_transform`.

[48] Wikipedia. Octave. Available at `http://en.wikipedia.org/wiki/Octave`.

[49] Wikipedia. RMSE. Available at `http://en.wikipedia.org/wiki/Root_mean_square_deviation`.

[50] Wikipedia. Scale-invariant feature transform. Available at `http://en.wikipedia.org/wiki/Scale-invariant_feature_transform`.

[51] Y. Wong. *Invariant Local Feature for Image Matching*. PhD thesis, The Chinese University of Hong Kong, December 2006.

[52] Y. Yoo. Tutorial on Fourier Theory. Available at `http://www.cs.otago.ac.nz/cosc453/student_tutorials/fourier_analysis.pdf`, March 2001.