A NEW APPROACH TO FACE RECOGNITION BASED ON GENERALIZED HOUGH TRANSFORM AND LOCAL IMAGE DESCRIPTORS

A Thesis

Submitted to the Faculty of Graduate Studies and Research

In Partial Fulfilment of the Requirements

For the Degree of

Masters of Science

In

Computer Science

University of Regina

By

Marian Moise

Regina, Saskatchewan

September, 2012

ABSTRACT

In this thesis a new approach to face recognition is presented. Face recognition is one of the most prolific research fields and also one of the most demanding. It is influenced not only by face-related attributes such as pose, position, scale, facial expression, accessories and physiognomy changes, but also by environmental factors like illumination, background, occluding objects, and, lastly, camera characteristics.

The generalized Hough transform is improved so that it can find the image region that best matches the template face image. Its reference point and hit rate are later used to discriminate between faces. The transform takes into account not only the position of the points, but also the value of their corresponding descriptors, which are compared against one another using the matrix cosine similarity measure and contribute to the final score.

The proposed method does not require any training data and it can be extended to object recognition. Moreover, it embeds not only the local structure of the face, represented by the local descriptors, but also the global shape of the face, captured by the generalized Hough transform and used later to discriminate between faces.

The main advantage of the new method stems from the fact that image descriptors better embed features than do simple pixels. As descriptors have higher memory requirements, they are computed only for the most interesting points in an image based on the Canny-edge detector.

Based on the locally adaptive regression kernels descriptor, a new descriptor,

namely, the gradient distance descriptor, is proposed in this work and test results for face identification using the Yale face database prove that it performs better than other descriptors. Moreover, the new method for face identification improves the recognition rate with at least 20% in comparison with Fisherfaces, no matter which descriptor is used.

As there are a plenitude of descriptors described in the literature, the proposed method for face recognition can be further improved by combining multiple descriptors in order to provide better invariance to the affine transformations and to increase the discriminative power.

ACKNOWLEDGEMENT

I would like to address my sincere gratitude to my supervisor Dr. Xue-Dong Yang who has encouraged me to follow this path, has provided me with the needed means and has shared with me his expert knowledge.

Along my way through the labyrinth, I was guided by the more experienced colleagues Dr. Richard Dosselmann and Brien Beattie. Also I have benefited from the experience and good advice of my brother and my sister-in-law, Daniel L. Moise and Dr. Gabriela Moise, respectively.

The financial support of my supervisor Dr. Xue-Dong Yang, the Faculty of Graduate Studies and Research and the Department of Computer Science helped to satisfy my daily basic needs and kept me focused on my research.

POST DEFENSE ACKNOWLEDGEMENT

At this point I wish to thank to my internal committee members, Dr. Cory Butz and Dr. Malek Mouhoub for their prompt and comprehensive feedback. Further appreciation is extended to my external examiner, Dr. Joseph Piwowar, for his insightful suggestions.

TABLE OF CONTENTS

LIST OF FIGURES

ABBREVIATIONS

DCT Discrete Cosine Transform

EKF Extended Kalman Filter

FLD Fisher's Linear Discriminant

GDD Geodesic Distance Descriptor

GHT Generalized Hough Transform

LARK Locally Adaptive Regression Kernels

LBP Local Binary Patterns

PCA Principal Component Analysis

SIFT Scale-Invariant Feature Transform

SSLD Self-Similarities Local Descriptor

WLD Weber Local Descriptor

Chapter 1

INTRODUCTION

Humans are endowed with the capability to categorize an object based on a single glance regardless of its pose, illumination condition, surface texture, deformation and whether the object is partially occluded. Given a training set of objects, humans are capable of generalizing the concept of a specific object, like a "house", and to recognize objects that have not been seen before. Developing a vision system that is able to identify different objects is a stupendous task. Most of the object recognition systems nowadays represent objects by using two-dimensional or three-dimensional models, based on commonalities extracted from multiple images, and match this new representation against the models in a database [LT09].

According to [LT09], the most common object recognition applications are in surveillance, industrial inspection, content-based image retrieval, robotics, medical imaging, human computer interaction, intelligent vehicle systems and biometric recognition systems. The most reliable methods for biometric recognition appear to be those based on the iris or a fingerprint, though a lot of effort has been recently directed at face recognition.

1.1 Motivation and Applications

Although iris-based or fingerprint person recognition systems are more reliable, it seems that face recognition is better for identification as it is less intrusive. The most suitable application for face recognition is video surveillance in which there is a watch list of possible suspects and its implications are not as drastic as in the identification case [CWS95].

Face recognition has a broad suite of additional applications ranging from the matching of face images taken in controlled environments to real-time matching of surveillance video images which impose higher computational requirements. Face matching is also used for passports, credit cards, driver's licenses, personal identification, mug shot matching and crowd surveillance [CWS95].

Generation of new face images from the input set has useful application such as witness face reconstruction, electronic mug shot books, reconstruction of faces from remains and computerized ageing [CWS95]. Many times the technology is used without people being aware of it, such as in identifying those with criminal records in stadiums and for preventing voter flaw, as some people might register to vote under different names.

The basic problem of face recognition can be formulated as follows: given a set of face images or video sequences of a scene, the system is asked to determine which face images from the stored database appear in these images. In the case of content-based image retrieval, the search is eased by additional input parameters such as race, gender or age.

The major steps in solving such a problem include a preprocessing stage, followed by the extraction of features from the face image and then matching the face against the faces from the database in order to identify the individual. Preprocessing of the initial image plays a significant role and might include face detection, illumination compensation, face alignment and segmentation for background removal.

Both local and global features should be included in a face representation. It seems that the most important parts of the face are the eyes, mouth, hairline and nose [CWS95]. Many times the feature extraction is tightly coupled with the identification part, with most of these approaches being based on eigenvectors, neural networks, feature points, profile image and different descriptors [CWS95].

Therefore, the evaluation of a face recognition solution is a complex task that involves a series of scenarios. According to the evaluation framework presented in [PGM+03], the results should be reported for a verification experiment, an identification experiment and a watch list experiment on the face database. For the basic cases, the tests are done on pictures taken either indoor or outdoor, on the same day, or a few days apart, and for different poses. Then, a more detailed analysis is done based on the resolution of the face, image compression, media, distance from the camera, watch list gallery size, or rank, and demographic factors such as sex and age. Also, a couple of experiments are done to test the underlying technology in cases such as three-dimensional morphable models, normalization and video-based recognition algorithms.

Based on the kind of experiments of [PGM+03], an important set of conclusions has been drawn. It seems that a good face recognition system is one that is not sensitive to normal indoor lighting. Outdoor performance, in general, needs to be improved. Because the normalized face images improve the verification and watch list performance, three-dimensional morphable models are very effective when dealing with non-frontal faces. Demographic factors, such as sex and age, influence the recognition rate as males seem to be easier to recognize than females, and younger people are harder to recognize than older people. Moreover, between the face recognition vendor test in 2002 and that in 2006 [PGM+03], scientists have successfully reduced the error rate of single image and three-dimensional face recognition systems by at least an order of magnitude through advances in their algorithms, computing power

and camera sensor design.

1.2 Contributions of this Research

Image descriptors are features that can better characterize an image through attributes such as shape, orientation, edges, luminosity, color, texture, or that can remove unwanted parts of an image like background, blurred regions and outlier pixels. Descriptors can be computed for an entire image or an image patch. Global descriptors are computed for an entire image, like the Fourier transform [GW06], *discrete cosine transform* (DCT) [GW06], *principal component analysis* (PCA) [MK01], whereas local descriptors are computed only for small regions, like in the *locally adaptive regression kernels* (LARK) [SM11], local DCT [GW06], *self-similarities local descriptor* (SSLD) [SI07] and *scale-invariant feature transform* (SIFT) [Low04].

The *generalized Hough transform* (GHT) [Bal81] is a technique used to detect predefined shapes in images. There are numerous implementations of it which take into account not only scaling, but also small deformations.

In this research, a novel approach to face recognition using the Hough transform is introduced. It searches not only for the face sketch, but also for regions which resemble each other in the new space defined by the local image descriptors. The face sketches for both images are computed using the Canny-edge detector, as these face sketches contain the major face traits. For each point in the face sketch, a local descriptor is computed based on the values of the neighbourhood pixels.

This novel approach to face identification, given in Algorithm 17, combines the idea of finding different shapes, based on GHT, with the power of descriptors, which can better describe features from images even when the face images have been taken under different conditions related to the subject, camera characteristics or environment factors.

The power of our method arises from the fact that any descriptor can be used for finding the best matching face. As descriptors become more powerful, in the sense that they better describe the image content and are more discriminative, the proposed method will improve accordingly.

Based on the computation of the LARK descriptor, a new descriptor, called geodesic distance descriptor (GDD) , is proposed in Section 3.2. The evaluation of its performance on the Yale face database in Section 4.1 proves that it performs better than the other descriptors and it increases the recognition rate by at least 20% in comparison with Fisherfaces for the task of face identification.

The proposed approach can be used not only for face recognition, but also for generic object recognition. Moreover, it does not require any training data and the face images can be of different sizes and orientations when using specific descriptors or taking advantage of the more complex form of the GHT.

1.3 Thesis Structure

The remaining parts of this thesis are organized as follows: Chapter 2 reviews two well-known algorithms for face recognition, namely, Eigenfaces and Fisherfaces [BHK97]. The generalized Hough transform is introduced along with the selected descriptors and similarity metrics. As image preprocessing plays an important role, a description of the popular Canny-edge detector is provided.

The detailed structure of the new system for face recognition is presented in Chapter 3. Also given is a description of how the descriptors are computed, along the face sketch and an illustration of how the GHT works. Because the similarity metrics play a very important role in comparing descriptors, a comparison between two of the most popular similarity metrics, namely, the Pearson correlation coefficient and the matrix cosine similarity, is illustrated too. The proposed approach to face recog-

nition is compared with two of the most popular techniques, namely Eigenfaces and Fisherfaces.

In Chapter 4, the new approach to face recognition is thoroughly evaluated using the Yale database for different configurations of parameters for each of the following descriptors: GDD, LARK, SSLD and DCT. The performance of each descriptor is evaluated for different values of the associated parameters.

Finally, Chapter 5 presents possible research directions that might be followed to improve the method and to apply it to different problems. The proposed method can take advantage of the latest research in searching within clusters and data representation in order to speed up the search within the GHT bins. The face images can be aligned so that they have similar orientation. Different other descriptors can be employed as they might embed better the face traits.

Chapter 2

BACKGROUND AND RELATED WORK

The general concepts behind the modified GHT and two of the most referenced face recognition algorithms, namely, Eigenfaces and Fisherfaces, are presented so that in Chapter 3 a comparison of these approaches can be provided. As descriptors play the most important role in discriminating among face images, the details of how descriptors are computed and the most used metrics for comparing them are presented. Finally, as the generalized Hough transform needs an edge-based representation of the image, the Canny-edge detector computation method is explained.

2.1 Face Recognition

The problem of face recognition could be stated as follows: given a set of face images that are labelled with the name of each person for training, identify each person which appears in the test set of images, taking into account that the test set contains the same people as the training set of images [BHK97].

One of the simplest face recognition algorithms is using the correlation of the face images. As this method is too expensive both from a computational and memory point

of view, a dimensionality reduction is needed. A well-known technique for achieving is PCA and its basic idea is to project the face images into a hyperspace that will maximize the scatter between images [BHK97]. The principal components of this hyperspace resemble the human face and are also called ghost images. More details about the PCA technique are provided in Section 2.1.2 and an improved version of it is presented in Section 2.1.3.

2.1.1 Yale Face Database

The Yale face database [BHK97] comprises images taken in front of a simple background for fifteen subjects, both males and females, under eleven different circumstances: ambient lighting and different face expressions (happy, sad, winking, sleepy, surprised and neutral), with an additional light source positioned to the left, center or right side of the subject, and with the subject wearing glasses or not.

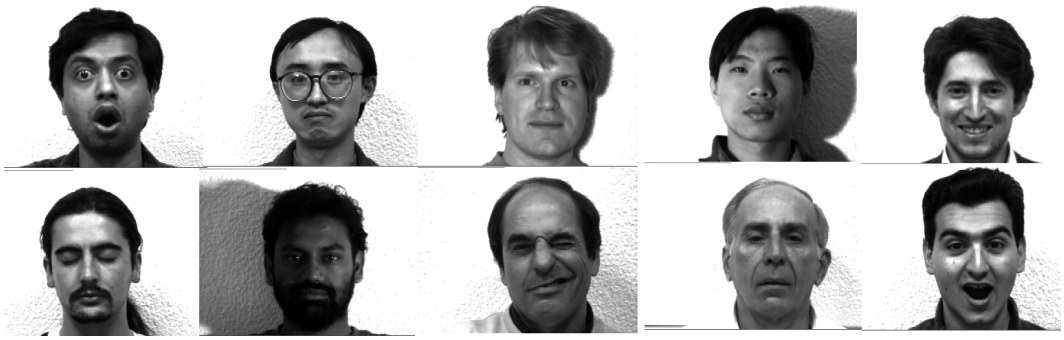


Figure 2.1: Yale face database samples [BHK97].

2.1.2 Principal Component Analysis

PCA projects face images into a subspace which embeds most of the data variance and whose main axes are the eigenfaces. These eigenfaces are the eigenvectors associated with the largest eigenvalues of the covariance matrix of the training data. The found eigenvectors correspond to the least-squares solution and are a powerful way to

represent the data. Thus, each face image can be reconstructed based on a weighted sum of the principal components computed from the training set of face images.

As described in [Tri09], in the initialization phase the training face images I_1,I_2,...,I_m are transformed from matrices I_i to vectors Γ_i and then normalized by their mean value Φ so that only the distinguishing features Φ_i from each face is stored and information that is common is removed:

$$I_i = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{p1} & a_{p2} & \cdots & a_{pn} \end{pmatrix} \rightarrow \Gamma_i = \begin{pmatrix} a_{11} \\ \vdots \\ a_{1n} \\ a_{21} \\ \vdots \\ a_{2n} \\ \vdots \\ a_{p1} \\ \vdots \\ a_{pn} \end{pmatrix}; \mu = \frac{1}{m}\sum_{i=1}^{m}\Gamma_i; \Phi_i = \Gamma_i - \mu. \tag{2.1}$$

This set of very large vectors Φ_i is then subjected to principal component analysis which seeks a set of m orthonormal vectors u_k and their associated eigenvalues λ_k that best describe the distribution of the data. The vectors u_k and scalars λ_k are the eigenvectors and eigenvalues of the covariance matrix:

$$C = \frac{1}{m}\sum_{i=1}^{m}\Phi_i\Phi_i^T = \frac{1}{m}AA^T, \tag{2.2}$$

where the matrix $A = [\Phi_1\Phi_2...\Phi_m]$. However, the matrix C is $p \times n$ by $p \times n$ and determining the $p \times n$ eigenvectors and eigenvalues is an intractable task for typical image sizes so a more computationally feasible method is needed. Fortunately, we can

determine the eigenvectors by first solving a much smaller m by m matrix problem by considering the transpose of the correlation matrix, and taking a linear combination of the resulting vectors [TP91]. The calculations are greatly reduced from the order of the number of pixels in the images $p \times n$ to the order of the number of images in the training set m which is relatively small $m << p \times n$. The associated eigenvalues allow to rank the eigenvectors according to their usefulness in characterizing the variation among the images. Normally the background is removed by cropping training images, so that the eigenfaces have zero values outside of the face area.

Each face image in the training set can be represented exactly as a linear combination of the eigenfaces. However, the faces can also be approximated using only the best k eigenfaces, those that have the largest eigenvalues, and which therefore account for the most variance within the set of face images. The primary reason for using fewer eigenfaces is computational efficiency.

When a new face image Γ is encountered, calculate a set of weights based on the input image and the k eigenfaces, denoted by U_i, by projecting the input image onto each of the eigenfaces:

$$\omega_i = U_i^T * (\Gamma - \mu), i = 1, ..., k. \tag{2.3}$$

The weights form a vector $\Omega = [\omega_1, \omega_2, ..., \omega_k]$ that describes the contribution of each eigenface in representing the input face image, treating the eigenfaces as a basis set for face images. The vector is used to find which of a number of predefined face classes best describes the face. The simplest method for determining which face class provides the best description of an input face image is to find the face class q that minimizes the Euclidean distance $\epsilon_q = |\Omega - \Omega_q|$, where Ω_q is a vector describing the face class q. A face is classified as belonging to class q when the minimum ϵ_q is below some chosen threshold θ_ϵ; otherwise it is classified as unknown. It seems that images of faces do not change radically when projected into the face space, while the projection of non-face images appear quite different. This basic idea is used to detect

the presence of faces in a scene by computing, at every location in the image, the distance ϵ between the local sub-image and face space.

One drawback of this method is that it maximizes the scatter of all the face images even if they belong to the same person. In case in which illumination changes, the principal components retain the light variation and cause the face images of different persons to be smeared together. A better approach is *Fisher's linear discriminant* (FLD) [BHK97] which not only maximizes the scatter across face images belonging to different persons, but also minimizes the scatter between face images belonging to the same person [BHK97]. In other words, FLD searches for those vectors in the underlying space that best discriminate among classes.

The performance of PCA is also related to the total number of eigenfaces used and above the limit of 45 eigenfaces no improvement has been noticed [BHK97]. The number of possible eigenfaces is equal to the number of face images in the training set. PCA can be used not only for the task of face recognition, but also for hand-print recognition, object recognition and robotics [MK01].

2.1.3 Fisher's Linear Discriminant

When the goal is classification rather than representation, the PCA solution may not yield the most desirable results. In such cases, one wishes to find a subspace that maps the sample vectors of the same class in a single spot and those of different classes as far apart from each other as possible. One technique derived to achieve this goal is known as Fisher's linear discriminant.

FLD is an example of a class specific method, in the sense that it tries to shape the scatter in order to make it more reliable for classification. This method selects U in such a way that the ratio of the between-class scatter and the within-class scatter

is maximized, where the between-class scatter matrix is defined as:

$$S_B = \sum_{i=1}^{c} N_i(\mu_i - \mu)(\mu_i - \mu)^T \tag{2.4}$$

and the within-class scatter matrix is defined as:

$$S_W = \sum_{i=1}^{c} \sum_{I_j \in C_i} (I_j - \mu_i)(I_j - \mu_i)^T, \tag{2.5}$$

where μ_i is the mean image of class C_i and N_i is the number of samples in class C_i. If S_W is non-singular, the optimal projection U_{opt} is chosen as the matrix with orthonormal columns which maximizes the ratio of the determinant of the within-class scatter matrix of the projected samples:

$$U_{opt} = \arg\max_U \frac{|U^T S_B U|}{|U^T S_W U|} = [u_1 u_2 ... u_k], \tag{2.6}$$

where u_i i=1,2,...,k is a set of generalized eigenvectors of S_B and S_W is corresponding to the k largest generalized eigenvalues λ_i, i.e.,

$$S_B u_i = \lambda_i S_W u_i. \tag{2.7}$$

Note that there are at most $c - 1$ nonzero generalized eigenvalues and this implies that the upper bound on k is $c - 1$, where c is the number of classes.

In the case of face recognition, the within-class scatter matrix S_W is always singular because the rank of S_W is at most $m - c$ and most of the time the number of images in the learning set m is much smaller than the number of pixels in each image $p \times n$. This means that it is possible to choose the matrix U such that the within-class scatter of the projected samples can be made exactly zero [BHK97].

To overcome the complication of a singular S_W, Fisher's method solves it by

projecting the set of images to a lower dimensional space so that the resulting within-class scatter matrix S_W is non-singular. This is achieved by using PCA to reduce the dimension of the feature space to $m - c$, resulting U_{pca}. Then the standard FLD is applied to reduce the dimension to $c - 1$ [BHK97], yielding U_{fld}. The optimal set of eigenvectors U_{opt} is obtained as follows:

$$U_{opt} = U_{pca} U_{fld}, \tag{2.8}$$

$$U_{pca} = \arg\max_U |U^T S_T U|, \; S_T = \sum_{j=1}^{m} (I_j - \mu)(I_j - \mu)^T, \tag{2.9}$$

$$U_{fld} = \arg\max_U \frac{|U^T U_{pca}^T S_B U_{pca} U|}{|U^T U_{pca}^T S_W U_{pca} U|}. \tag{2.10}$$

Optimization of U_{pca} is performed over $m \times (m - c)$ matrices with orthonormal columns, while the optimization for U_{fld} is performed over $(m - c) \times k$ matrices with orthonormal columns. The smallest $c - 1$ principal components are discarded when computing U_{pca} [BHK97]. Then the classification of a new image Γ is done by projecting it into the new face space described by U and finding the closest class mean μ_i projection based on Euclidean distance.

FLD method chooses the projection components or fisherfaces [BHK97] in such a way that they are not influenced by slight changes in lighting, facial expression and presence of glasses. It has been used in face recognition and mobile robotics. Moreover, it has been proposed for generic object recognition, but the results using a large database of objects have not been reported yet [MK01].

2.2 Generalized Hough Transform

The GHT is used in [Low04] to filter out the false matches of descriptors for the task of object recognition in case of cluttered images, small-sized images and occluded

objects. The idea behind it is to create clusters of features that can be used to estimate object location, orientation and scale based on matched features. As the pose estimation resulting from the GHT has large errors, each cluster is further used to approximate the affine transformations parameters based on a least square estimation that best relates the descriptors of the new object to the descriptors of training objects stored in the database. The final stage is based on a Bayesian analysis that performs a final verification and yields the location, orientation and scale of the new object relative to the objects from the database.

Another interesting application of the GHT is for sketch-based image retrieval, presented in [ACE07]. It can find a series of objects such as cars, horses, bottles, watches, saxophones, etc, based on a simple image sketch of these objects. The method is called deformation tolerant GHT and it can handle not only object deformations, but also non-uniform backgrounds and occlusions.

The task of real-time face detection and tracking has been solved in [Sch00] by combining the GHT, for head detection, and an *extended Kalman filter* (EKF) , for estimating the position and attitude of the head. The proposed representation is based on a mapping of facial features to a 3D template model of a face.

In order to detect multiple objects in an image, Hough forests [BLK12] can be trained with image patches to recognize specific objects. It has been successfully applied in [BLK12] to the task of detection of pedestrians in crowded places. Unfortunately, the increase in the accuracy is balanced by the increased computational complexity in comparison with the basic version of GHT transform.

The recognition of handwritten Chinese characters has benefited from the GHT too. In order to make the computation more efficient, the authors in [LD95] compute the R-table of GHT only for the stroke points and not for all the edge points of the characters.

The task of detecting the shape of 3D objects from a single image has been tackled

in [Kor07] by employing the GHT to estimate the lighting direction and the normal to the surface of the object. Good results have been obtained for the detection of spheres from a single image and for the relighting of these spheres.

The problem of template image matching has been solved in [LZ05] by taking advantage of the GHT. In [LZ05], the features are first extracted from image patches corresponding to the blocks of a grid superimposed on the image. These features are computed in such a way that they are robust to noise and illumination conditions and then are compared against the features from the other image in order to vote for the possible locations of the template image. The final set of positions results from clustering the values in the accumulator, so that the method is more tolerant to illumination noise and distortions. The authors claim that this method is robust to linear transformations applied to the intensity of the pixels and also to the presence of noise in the image.

The generalized Hough transform is a method for the detection of boundaries of an arbitrary non-analytic shape in images by using the information from a predefined boundary. These boundaries or edges are usually computed using a gradient-based method such as the Prewitt operator or Canny-edge detector. It seems that the human visual system has the capability to recognize objects even from a sketch [Bal81].

The basic idea of the GHT stems from the fact that a template of a shape can be computed based on the image edge map. This template, also called the R-table, stores the locations of the edge points relative to a reference point, which can be interpreted as the origin of the coordinate system. Then each edge point from the target image will cast a vote for the possible locations of the reference point based on the R-table and the maximum accumulated value will represent the new reference point of the shape. This method works even when the edge map is discontinuous as a result of a noisy image or partial occlusion of the object [Bal81].

In the GHT, any shape can be described by a set of three parameters y, s, θ, where

$y = (x_r, y_r)$ is the reference point of the shape, s is a scaling factor and θ is the initial rotation angle of the shape. The R-table is built based on the orientation of the shape edges relative to the reference point y and then simple transformations are applied to it in order to accommodate the scaling and rotation parameters [Bal81].

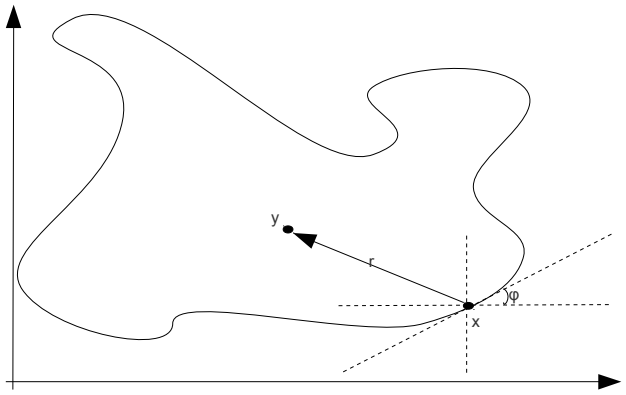


Figure 2.2: Arbitrary non-analytic shape for GHT.

After the reference point is chosen, the vector $\vec{r}$ and the gradient orientation ϕ for each point x on the shape are computed. The R-table stores the set of all possible vectors $\vec{r} = y - x$ for specific values of ϕ. In order to detect a single shape in an image, an accumulator array A is used and its maximum value location will represent the new origin of the shape. The accumulator array has a size equal to the image size plus a specific padding and the value stored at each of these locations represent the probability of being the reference point of shape. For each point x_{new} from the edge image, the gradient orientation ϕ_{new} is computed and, based on the R-table, all the possible values for $\vec{r}$ are located and used to increment the locations from the accumulator array corresponding to $x_{new} + \vec{r}$. The idea behind is based on the fact that the edge points located on the boundary of the searched shape will increment the same point in the accumulator array, while the edge points located elsewhere will increment different points in the array [Bal81].

Given a boundary B and the points x located on it, one can compute the R-table, illustrated in Table 2.1, based on the vector $\vec{r}$ and gradient direction $\phi(x)$ of the point x.

i	ϕ_i	$\vec{r}_{\phi_i}$
0	0	$\vec{r} = y\vec{x}, x \in B, \phi(x) = 0$
1	$\Delta\phi$	$\vec{r} = y\vec{x}, x \in B, \phi(x) = \Delta\phi$
2	$2\Delta\phi$	$\vec{r} = y\vec{x}, x \in B, \phi(x) = 2\Delta\phi$
$\vdots$	$\vdots$	$\vdots$

Table 2.1: R-table structure.

The R-table can be easily transformed so that it accounts for shape changes in scale, rotation and reference point translation. In case of a composite shape, the R-table can be computed by merging together the R-tables of the shape sub-parts. Moreover, each term in the R-table can be weighted based on a importance metric [Bal81].

According to [MM09], weights can be learnt automatically using a max-margin framework which optimizes the classification performance by using a set of predefined appearance codebooks and their relative position to the reference point. After the local features of the image are extracted, they are matched against a set of codebook entries and assigned weights based on their number of occurrences and location. For the matched codebook entries, the GHT is applied and the resulting set of possible locations is further pruned by using a learned distribution for reference points of the objects. Lastly, the region around these selected locations is cropped and an SVM classifier finds the exact reference point of the object.

2.3 Image Descriptors

Although image descriptors require additional memory and they increase the processing time, descriptors are preferable to raw pixel intensities as they better embed the

features of an image than do pixels. Moreover, some of the descriptors are invariant to scaling, rotation, shearing, translation, lighting conditions or small deformations.

2.3.1 Discrete Cosine Transform

With MP3 and JPEG using the DCT to compress data, DCT has become one of the most used transformations for image and signal compression. It is a technique resembling the discrete Fourier transform, except for the fact that it uses a series of cosines in the real domain as a decomposition basis instead of a combination of sines and cosines in the complex domain.

In the case of an image $f(x, y)$ of size $N \times N$, its DCT $T(u, v)$ can be expressed as follows [GW06]:

$$T(u,v) = \alpha(u)\alpha(v) \sum_{x=0}^{N-1} \sum_{x=0}^{N-1} f(x,y) \left(\cos \frac{(2x+1)u\pi}{2N} \right) \left(\cos \frac{(2y+1)v\pi}{2N} \right)$$

where

$$\alpha(u) = \begin{cases} \sqrt{\frac{1}{N}} \text{ for u=0} \\ \sqrt{\frac{2}{N}} \text{ for u=1,2,...,N-1} \end{cases}$$

, and similar for $\alpha(v)$.

The selection of the most significant coefficients is done by using a zigzag traversal pattern and by discarding the first coefficient, as it is highly influenced by lighting conditions, and the last ones which correspond to noise or higher levels of detail. By discarding part of the coefficients, the image can still be reconstructed and the image noise can be reduced.

2.3.2 Self Similarity Local Descriptor

The most interesting property of the SSLD is the fact that it properly represents the internal geometric layout of local similarities in images even in cases in which these

patterns are generated by different image properties such as colours, textures and edges [SI07]. As a result, it can be used for very challenging applications such as object detection in images using rough hand-sketches or action detection in cluttered video data without prior learning [SI07].

It is computed using a kind of correlation based on the sum of squared differences between a small patch P around the current pixel q and the other patches from a neighbourhood region R centred at q:

$$SSD_q(x,y) = \sum_{i,j \in P_q} (I_p(i,j) - I_r(x+i, y+j))^2, \text{ for each x,y} \in R_q. \tag{2.11}$$

The resulting distance surface $SSD_q(x,y)$ is normalized and transformed into a correlation surface $S_q(x,y)$:

$$S_q(x,y) = \exp\left(-\frac{SSD_q(x,y)}{\max(var_{noise}, var_{auto}(q))}\right). \tag{2.12}$$

where var_{noise} corresponds to acceptable photometric variations (in color, illumination or due to noise) and $var_{auto}(q)$ takes into account the patch contrast and its pattern structure, such that sharp edges are more tolerable to pattern variations than smooth patches.

The correlation surface is then transformed into a binned log-polar representation because it results in a compressed descriptor for each point and it better handles small deformations [SI07]. For each of the obtained bins from the log-polar representation only their maximum correlation value is stored. In this log-polar representation, the new coordinates are $\log(\rho)$ - logarithmic distance from the center of the image to a given point and θ - the angle of the point with the center:

$$\rho = \sqrt{(x-x_C)^2 + (y-y_C)^2} \text{ called radial distance,} \tag{2.13}$$

$$\theta = \arctan\left(\frac{y - y_C}{x - x_C}\right) \text{ called angular distance.} \tag{2.14}$$

This results in a space variant representation whose resolution is highest in the center and decreases with eccentricity [SI07] as it is illustrated in Figure 2.3. Another reason for using a log-polar representation of an image patch is that scaling and rotation in the Cartesian domain corresponds to pure translation in log-polar domain.

(a) (b)

Figure 2.3: Log-polar representation of an image.

The normalization of the SSLD is done by converting its values to the range [0,1] in order to be invariant to the differences in pattern and color distribution of different patches and their surrounding image regions [SI07]. The most interesting applications of this descriptor are the retrieval of the images from a database that resemble a rough hand-sketched query and the detection of complex actions from video sequences.

2.3.3 Locally Adaptive Regression Kernels

The LARK descriptor measures the similarity between a center pixel and surrounding ones from a local image region by using a "signal-induced distance" [SM11]. It is based on the geodesic distance and image gradients and it has the ability to represent

geometrical shapes even in noisy images.

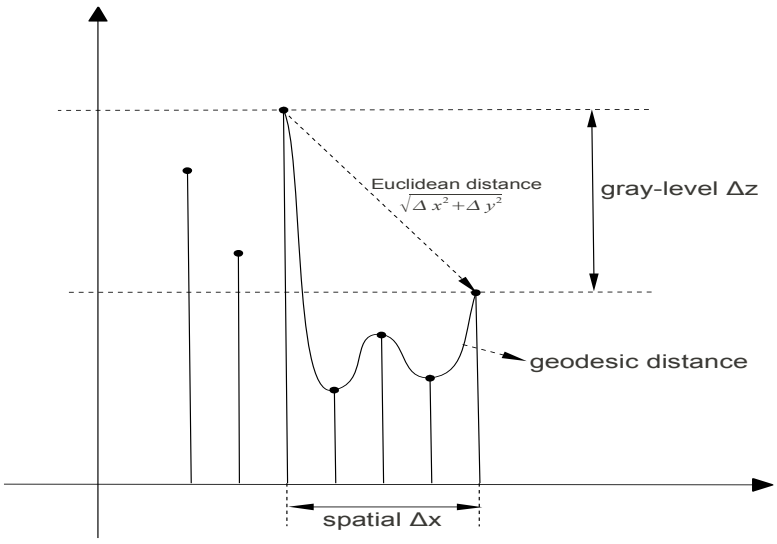


Figure 2.4: Geodesic distance versus Euclidean distance [SM11].

The grayscale values of an image can be represented as a function of their locations in the image, namely, I(x,y), as follows:

$$dI(x,y) = \frac{\partial I}{\partial x}dx + \frac{\partial I}{\partial y}dy = I_x dx + I_y dy. \tag{2.15}$$

In the three dimensional space defined by x, y and I, the differential arc length on the surface S(x,y) can be expressed as [SM11]

$$
\begin{aligned}
ds^2 &= dx^2 + dy^2 + dI^2 && (2.16)\\
&= dx^2 + dy^2 + (I_x dx + I_y dy)^2 && (2.17)\\
&= (1 + I_x^2)dx^2 + 2I_x I_y dx dy + (1 + I_y^2)dy^2 && (2.18)\\
&= \begin{bmatrix} dx & dy \end{bmatrix} \begin{bmatrix} I_x^2 + 1 & I_x I_y \\ I_x I_y & I_y^2 + 1 \end{bmatrix} \begin{bmatrix} dx \\ dy \end{bmatrix} && (2.19)\\
&= \Delta X^\top C \Delta X + \Delta X^\top \Delta X && (2.20)
\end{aligned}
$$

with

$$
\Delta X = \begin{bmatrix} dx & dy \end{bmatrix}^\top, \tag{2.21}
$$

and C being the local covariance matrix.

Based on the fact that the term $\Delta X^\top \Delta X$ is dependent only on the pixel position in the image and does not take into account the intensity of that pixel, the following approximation can be done:

$$
ds^2 \approx \Delta X^\top C \Delta X \tag{2.22}
$$

For a patch of size P, the LARK descriptor is defined as the similarity between the center pixel and its neighbours [SM11], namely,

$$
\begin{aligned}
K(C_l, \Delta X_l) &= \exp(-ds^2) && (2.23)\\
&= \exp(-\Delta X_l^\top C_l \Delta X_l), && (2.24)
\end{aligned}
$$

with l $\in$ {1,2,...,P} and ΔX defined in Equation (2.21).

As C_l is computed using the gradient (I_x, I_y) of each pixel from the patch, it is influenced by noise and perturbations from data. The covariance matrix C_l for the

patch P is computed using the average gradient $(I_{x_{avg}}, I_{y_{avg}})$ of the patch [SM11], namely,

$$C_l = \begin{bmatrix} I_{x_{avg}}^2 & I_{x_{avg}} I_{y_{avg}} \\ I_{x_{avg}} I_{y_{avg}} & I_{y_{avg}}^2 \end{bmatrix}, \tag{2.25}$$

with $I_{x_{avg}} = \frac{1}{P} \sum_{l=1}^{P} I_{x_l}$ and similarly for $I_{y_{avg}}$.

In order to smooth the largest variations of the image surface and make the local geodesic distance more stable, the covariance matrix is decomposed into eigenvectors using the singular value decomposition [SM11], namely,

$$\begin{aligned} G &= \begin{bmatrix} I_x \\ I_y \end{bmatrix} && (2.26) \\ &= USV^\top, && (2.27) \end{aligned}$$

with $UU^\top = VV^\top = I_n$. We then obtain:

$$\begin{aligned} C_l &= GG^\top && (2.28) \\ &= USV^\top VS^\top U^\top && (2.29) \\ &= US^2U^\top && (2.30) \\ &= \begin{bmatrix} u_1 & u_2 \end{bmatrix} \begin{bmatrix} s_1^2 & 0 \\ 0 & s_2^2 \end{bmatrix} \begin{bmatrix} u_1^\top \\ u_2^\top \end{bmatrix} && (2.31) \\ &= s_1 s_2 \left(\frac{s_1}{s_2} u_1 u_1^\top + \frac{s_2}{s_1} u_2 u_2^\top \right), && (2.32) \end{aligned}$$

where u_1, u_2 are the eigenvectors and s_1, s_2 are the singular values of the matrix G.

According to [SM11], the singular values are regularized in order to avoid numer-

ical instabilities, as

$$C_l^{reg} = (s_1 s_2 + \epsilon)^{\alpha} \left(\frac{s_1 + \tau}{s_2 + \tau} u_1 u_1^{\top} + \frac{s_{2+\tau}}{s_1 + \tau} u_2 u_2^{\top} \right), \tag{2.33}$$

where $\epsilon = 10^{-7}, \tau = 1, \alpha = 0.5$ [SM11].

Lastly, each local covariance matrix is normalized to unit norm so that the corresponding LARK descriptor better handles the illumination changes [SM11]. The LARK descriptor is most suitable for generic object detection [SM09] as it handles well the changes in imaging conditions, distortions and the background of the object, although it has been successfully applied in [SM11] to the task of face matching too.

2.3.4 Gradient Distance Descriptor

Based on the observation that a patch, surrounding a point of interest (x_c, y_c) for which the descriptor is computed, can be characterized by its major internal variations of the image gradient and based on how the LARK descriptor is computed, we derive a new descriptor using the weighted sum of the gradients $(I_{x_{avg}}, I_{y_{avg}})$ for the patch around the current pixel and the relative distance dx, dy between the current pixel and its neighbours from the current image patch. The idea behind this is that the farther pixels have less influence when computing the average gradient of the image patch and that the values of the descriptor are also influenced by the position of the pixels from the patch relative to the center pixel.

For a patch of size P around the center pixel $P(x_c, y_c)$, the GDD is defined as the similarity between the center pixel and each of its neighbours in the new space defined by the image gradients along x and y axes, namely

$$GDD_l(x_c, y_c) = \exp\left(-(I_{x_{avg}} dx_l + I_{y_{avg}} dy_l)^2\right), \tag{2.34}$$

with $l \in 1, 2, ..., P$

A good illustration of how the GDD is computed along a face sketch is provided in Figure 3.2. It is interesting that whereas the LARK descriptor representation resembles a sphere, the GDD resembles part of a toroid.

2.4 Metrics for Descriptors Comparison

Although one might understate the power of the similarity metrics for comparing the descriptors, these metrics play a key role in the whole system and, therefore, a comparison of three of the most common similarity metrics is proposed in Section 2.4.4. A detailed mathematical analysis of the Pearson correlation coefficient is done in [SB07].

2.4.1 Pearson Correlation Coefficient

Given two grayscale images of similar sizes, $M \times N$, represented by matrices A and B, the Pearson correlation coefficient can be defined as follows:

$$PCC(A,B) = \frac{\sum_m \sum_n (A_{mn} - \overline{A})(B_{mn} - \overline{B})}{\sqrt{\left(\sum_m \sum_n (A_{mn} - \overline{A})^2\right)\left(\sum_m \sum_n (B_{mn} - \overline{B})^2\right)}}, \tag{2.35}$$

where $\overline{A}, \overline{B}$ represent the mean of A and B, respectively.

The Pearson correlation coefficient is in the range [-1,1], having the value of zero for uncorrelated images, positive values in case the images are increasingly dependent and negative values in case images are decreasingly dependent. When comparing two image descriptors only the modulus of the Pearson correlation coefficient is considered, as the higher the modulus of the coefficient, the more related the images are.

It seems that even though the image of Figure 2.5b is a blurred version of the initial image 2.5a, their Pearson correlation coefficient still has a high value. Based on [SB07], one should choose a different metric when comparing descriptors, such as

(a) (b)

Figure 2.5: Pearson correlation coefficient PCC(A,B)=0.9894.

cosine similarity.

2.4.2 Normalized Root Mean Square Deviation

Given two grayscale images of size $M \times N$, represented by matrices A and B, the normalized root mean square deviation has a value in range [0,1], with a value of zero in case in which the images are identical, and is defined as

$$NRMSD(A,B) = \frac{\sqrt{\frac{\sum_m \sum_n (A_{mn} - B_{mn})^2}{M*N}}}{max(A_{mn}, B_{mn}) - min(A_{mn}, B_{mn})}. \tag{2.36}$$

2.4.3 Matrix Cosine Similarity

Given two images of similar size $M \times N$, represented by matrices A and B, the matrix cosine similarity(MCS) between them can be defined as:

$$MCS(A,B) = \frac{A \cdot B}{\|A\|\|B\|} \tag{2.37}$$

$$= \frac{\sum_m \sum_n A_{mn} B_{mn}}{\sqrt{\sum_m \sum_n A_{mn}^2} \sqrt{\sum_m \sum_n B_{mn}^2}}. \tag{2.38}$$

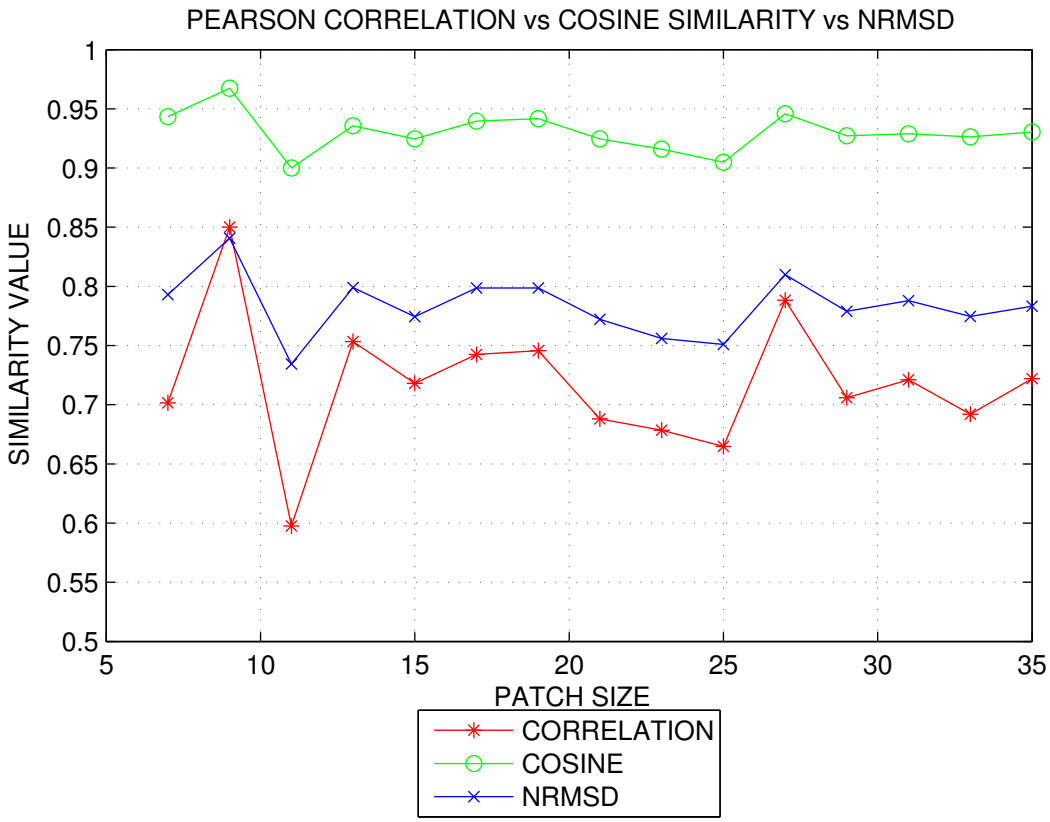
The value returned is in range [0,1], with a value of one in case in which the images are identical. The use of the modulus of their matrix cosine similarity for comparing images is based on the fact that for identical images the resulting value is one and for very different images the value gets close to zero.

2.4.4 Correlation versus Cosine Similarity

One of the most used statistical methods for template matching in images is the Pearson correlation coefficient, as it works well for image patches that have enough details. Taking into account that the descriptors are not uniform, therefore they have many details embedded, and the fact that most of their values are different than zero, the correlation metric is suitable for comparing them. Correlation metric has been successfully applied even for the task of face recognition in the frequency domain [KSX06].

According to [SB07], the correlation metric is not suitable for measuring the similarity between vectors as it is highly biased by the zero values. The cosine similarity and Chi square metrics are more recommended [SB07] according to the results of the Mantel test and Procrustes analysis. The main difference stems from the fact that the cosine metric uses the original values, whereas the correlation metric uses the centred values around the mean and this may result in a lower discriminative power as some information is discarded. It seems that correlation is measuring the proximity between data rather than their similarity [SB07].

Our empirical results of Figure 2.6 confirm that MCS produces better results than Pearson correlation and normalized root mean square deviation metrics in the case of randomly generated grayscale image patches of different sizes. The experiment simulates the effect of image processing on the comparison result between the initial image and the processed version of it using the aforementioned metrics. The images are processed to simulate different ambient lighting conditions, by adding a constant



(a) AMBIENT LIGHTING VARIATION SIMULATION.

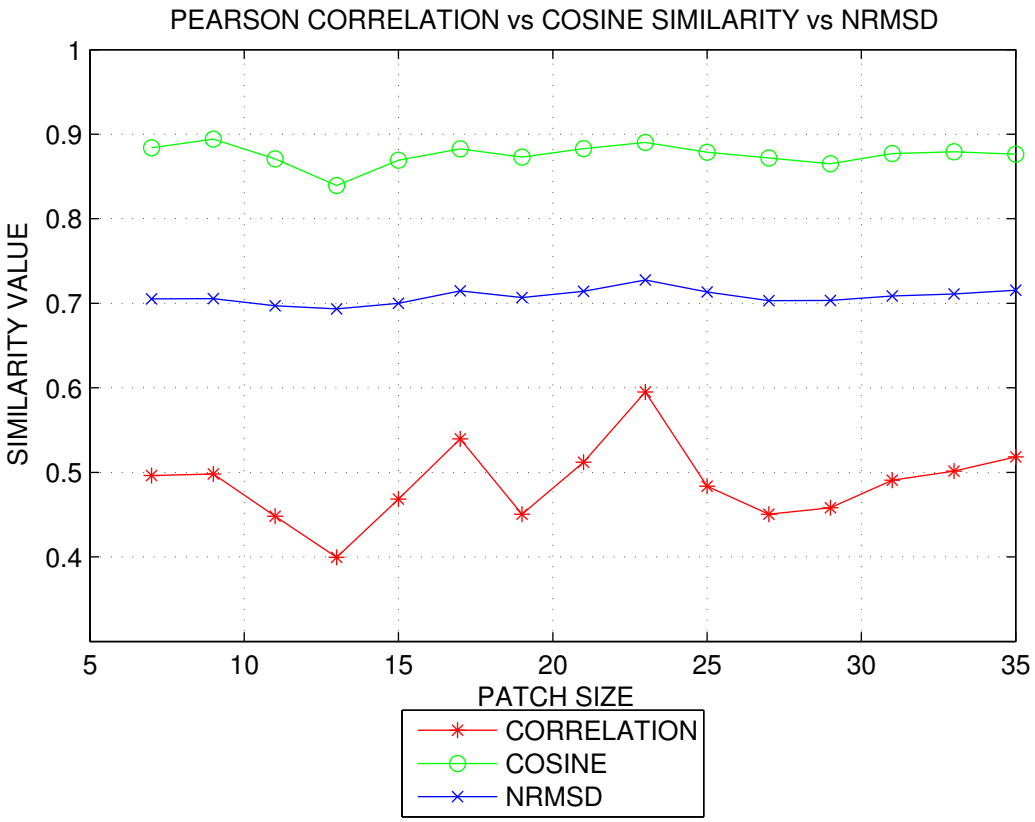
Figure 2.6

value to all the intensity values corresponding to the pixels from the generated images, adjusting the brightness of an image, by multiplying the grayscale intensities with a scaling factor, or image rotation.

2.5 Image Preprocessing

Preprocessing of the initial image plays a significant role and might include face detection, edge detection, illumination compensation, face alignment and segmentation for background removal. One of the most popular edge detection algorithms is Canny-edge detector and a description of it is presented this section.

The edge map of the object to be analyzed is computed using the Canny edge



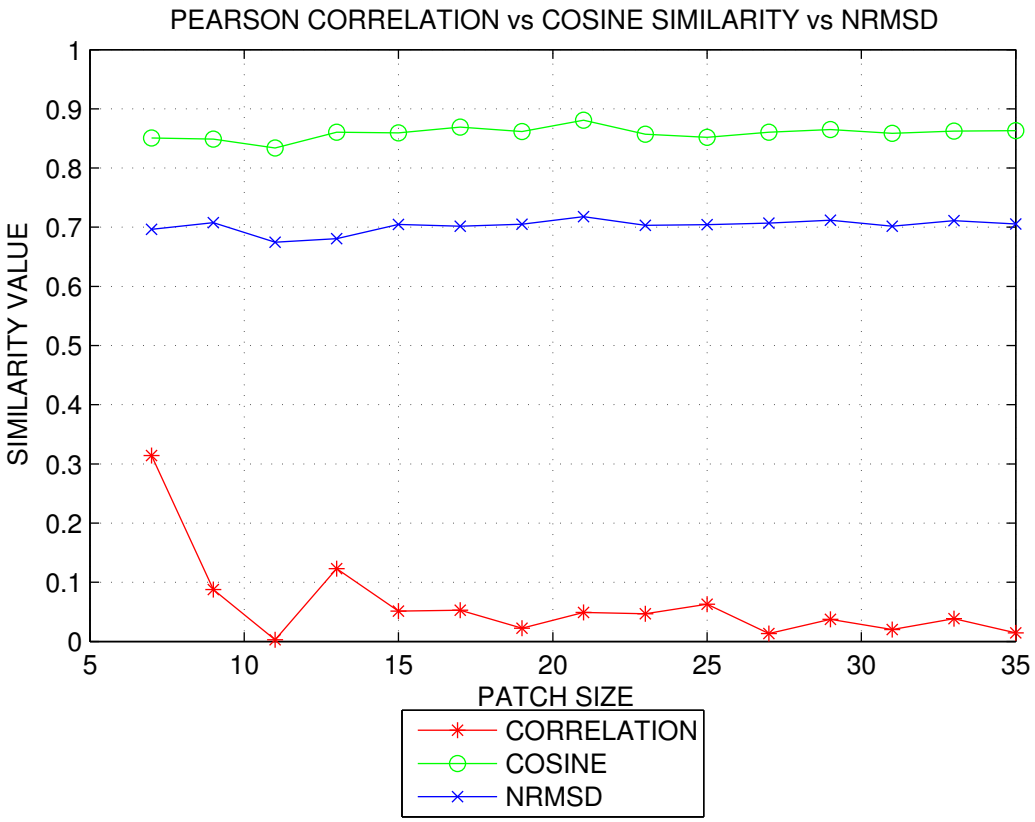
(b) BRIGHTNESS VARIATION SIMULATION.

Figure 2.6

detector. Not only does it speed up the GHT, it also selects the most interesting points for which to compute the descriptors by choosing the most significant edges. Although descriptors better embed the features of an image than do simple pixels, they have higher memory requirements and increase the processing time.

The first stage of the Canny edge detector is reducing the noise from the image by applying a Gaussian blur filter and its effect can be seen in Figure 2.7.

In the second stage, the Sobel operator is used to approximate the gradients of the image $I(x, y)$ along the x and y, directions and to derive the gradient magnitude $G = \sqrt{G_x^2 + G_y^2}$ and orientation $\theta = \arctan\left(\frac{G_x}{G_y}\right)$ as follows,



(c) IMAGE ROTATION SIMULATION.

Figure 2.6: Comparison between similarity matrices (original in colour).

$$G_x = \begin{bmatrix} -1 & 0 & +1 \\ -2 & 0 & +2 \\ -1 & 0 & +1 \end{bmatrix} I, \tag{2.39}$$

$$G_y = \begin{bmatrix} -1 & -2 & -1 \\ 0 & 0 & 0 \\ 1 & 2 & 1 \end{bmatrix} I. \tag{2.40}$$

After the magnitude and orientation of the image gradients are computed, the pixels for the major edges are selected based on the upper threshold set for the gradient magnitude. Given that a big gradient magnitude implies a significant change

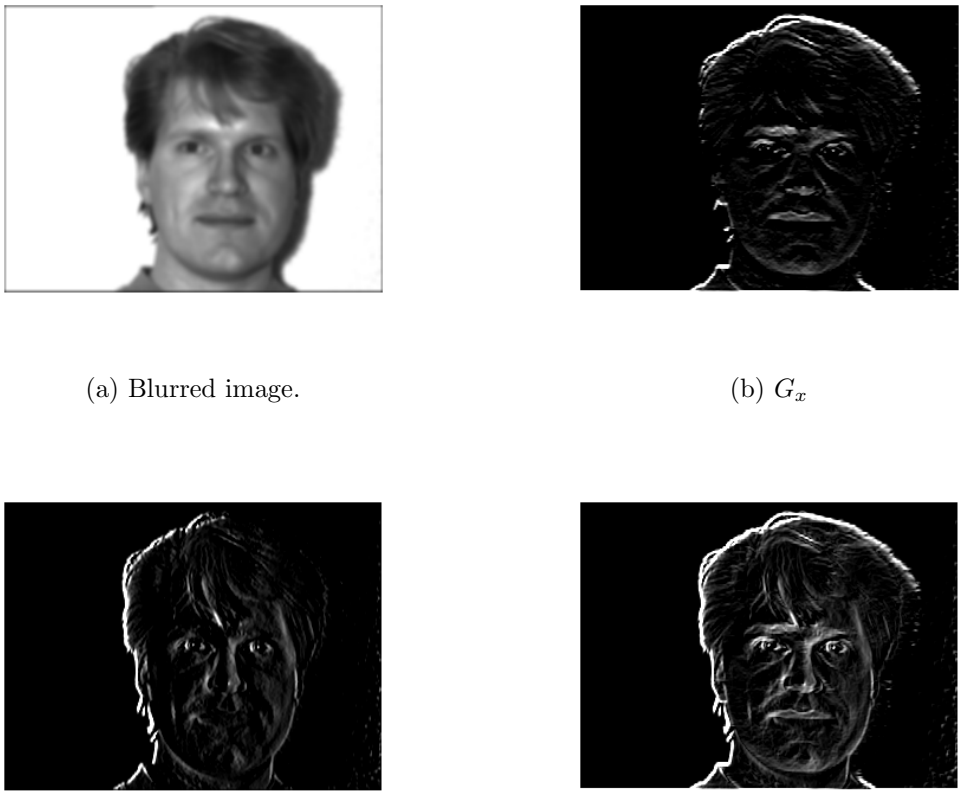
(a) (b)

Figure 2.7: Filtering the noise with a Gaussian filter of size 5×5 and standard deviation of 2.

in color and signals a major edge, whereas a small gradient magnitude is determined by a small change in color, the most significant edge pixels of the image are marked.

The third stage is depicted in Figure 2.9 and it reduces the width of the edges to one pixel using non-maximum suppression. The non-maximum suppression is a technique which considers the edge points as being the points having a local maximum gradient magnitude in the direction of the gradient orientation.

The gradient direction angle previously computed is rounded to 0,45,90 or 135 degrees which corresponds to the four major directions for edges: N-S, NW-SE, E-W or NE-SW. It is worth mentioning that the gradient is oriented perpendicular to the edge direction as color intensities change across the edges. As a result, an estimation of the edge orientation is done based on the gradient direction of current pixel. Then the gradient magnitude of the current pixel is compared with the magnitude of the adjacent pixels which have the same orientation. If the magnitude of the current pixel is higher, then it is marked for further processing, otherwise it is discarded. For example, if the current pixel gradient direction is 0 degrees, it is on a N-S edge and it has as adjacent pixels the E and W, then it is not discarded as long as its gradient magnitude is higher than that of the adjacent pixels. In Figure 2.9, the current pixel



(a) Blurred image.

(b) G_x

(c) G_y

(d) $G_x + G_y$

Figure 2.8: Applying the Sobel operator to a blurred image.

is kept for further processing, as the gradient magnitude of the current edge G is higher than the gradient magnitude of its E and W neighbours, that is, $G > G_E$ and $G > G_W$.

The last stage is to apply a thresholding technique based on hysteresis: using a upper threshold to start the edge curves and a lower threshold to connect them. As a positive side effect, it also removes noise points with high gradient magnitudes that are not along an edge curve.

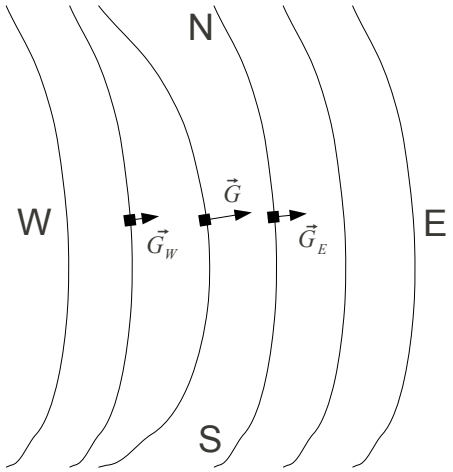
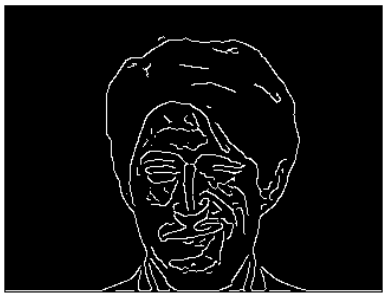


Figure 2.9: Non-maximum suppression for N-S edges.

2.6 Summary

In this chapter, a short overview of the two most common face recognition methods is presented, namely, PCA and FLD. Their performance is later evaluated on Yale face database, so a description of the database is presented too. As our new approach to face recognition is based on the GHT and a set of descriptors, an overview of the GHT is done in Section 2.2 and the employed descriptors are presented in Section 2.3. Moreover, a new descriptor is proposed in Section 2.3.4, namely, GDD. The metrics used to compare the descriptors are presented and then compared in Section 2.4. Lastly, the face sketches, used by the GHT, are computed by Canny-edge detector which is presented in Section 2.5.

(a) Initial image



(b) Upper threshold=0.1

(c) Upper threshold=0.25

Figure 2.10: Applying Canny-edge detector to an image.

Chapter 3

FACE RECOGNITION USING THE GENERALIZED HOUGH TRANSFORM

3.1 Introduction

The GHT is based on the following voting schema: each feature from the input image votes for the possible locations of the reference point that might have generated this feature based on the R-table which is built using the template image. For example, a bike might be detected by the fact that many of its component parts, such as wheels, crank set, chain, seat post, etc., are casting many votes for a specific point such as the bike's center of gravity. As some parts might resemble each other, like the front and back wheels, their associated features increment multiple locations of the accumulator of the GHT. As a result, only the peaks of the accumulator are considered for further processing. These peaks represent the certainty of detecting the template object in the input image.

The power of the GHT for object detection stems from the fact that the used

features can be extracted by a multitude of methods, ranging from ones based on simple edge pixels or interesting points to the more elevated ones based on patches and regions that are somewhat compressed in the form of image descriptors. In cases in which training images are available, dictionaries of features can be built in order to speed up the task of object detection. Moreover, the GHT is robust to partial occlusions, small deformations and image noise.

Most of the algorithms used for the classification of objects from images are based on computing a set of descriptors for image patches around some points of interest or on sampling random patches across the images and applying a dimensionality reduction technique. Since many of the descriptors are not informative, some of them are discarded based on some metrics or on a previously learned dictionary. In order to ease the task of classification, the remaining descriptors are clustered.

The modified GHT selects interesting points by taking advantage of the image sketch computed with the Canny-edge detector, then it computes the descriptors corresponding to these points. The task of classification and the pruning of the descriptors are combined, so that only descriptors that meet the epsilon threshold are voting for the reference point of the object. The number of votes or hit rate indicate the degree of resemblance between the two objects and it helps in identifying the initial object.

The general structure of this system is depicted in Figure 3.1. The preprocessing block outputs a face sketch from the initial image by employing a Canny-edge detector. Then the feature extraction block computes the descriptors for the pixels along the edges. Based on the images stored in the database, the feature matching unit computes the similarity between the query image and all of the images from the database by applying a modified version of the GHT. Finally, the face identification block gives as output the image that most closely resembles the input image in terms of the highest hit rate.

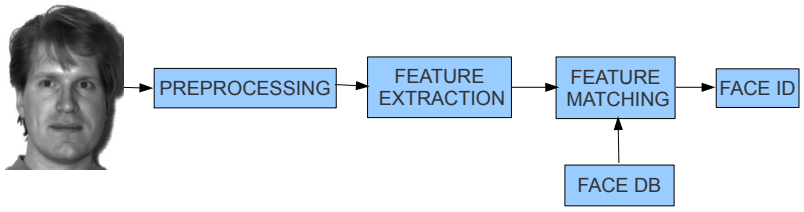


Figure 3.1: The structure of the face recognition system (original in colour).

This new approach to face recognition based on GHT is based on the fact that given two images, the system is able to compare them and give a score for their similarity using the highest hit rate stored in the accumulator of the GHT. Given a face image of a person from the database, the system compares this face image against all the other face images stored in the database and it comes up with the best possible match based on the similarity scores.

3.2 Computation of the Descriptors

The face sketches for both images are computed using the Canny-edge detector as they contain the major face traits of the faces. Then, for each point from the face sketch, a local descriptor is computed based on the values of the neighbourhood pixels.

Because of their popularity and the good results for face recognition, the descriptors used are DCT, SSLD, LARK and GDD. The comparison of their performance is done in Chapter 4 based on their results for face recognition using the modified GHT on the Yale database.

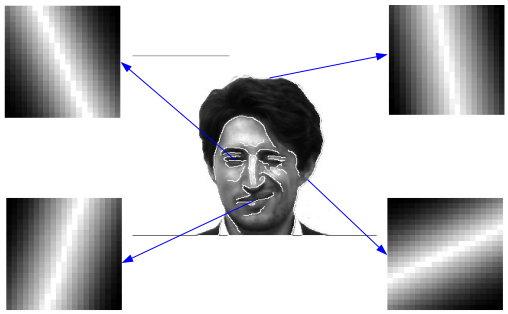


Figure 3.2: How GDD descriptors are computed along the face sketch (original in colour).

3.3 Matching Ensembles of Descriptors

One of the rudimentary ways of matching a set of descriptors extracted from a query image with a set of descriptors extracted from a template image is comparing each descriptor from the query image with the other descriptors from the template image. Although this has a high processing time and is prone to errors, as totally different descriptors can have a high similarity value or similar image descriptors might exist at different locations, it might yield good results if fine tuned for specific faces and small databases of images.

On the other hand, an approach based on the GHT, that clusters the descriptors, is more robust to errors and decreases the processing time substantially. It yields the good results when the thresholds are more relaxed as it counts on a higher number of points, not on an extremely high precision. It can yield good results even when conditions are very strict, if the number of matched descriptors is above a specific threshold.

Matching descriptors across images can also be performed by employing a probabilistic model called "star graph" [BI05], based on geometric relations between descriptors. It builds a pattern by connecting the descriptors from the template image

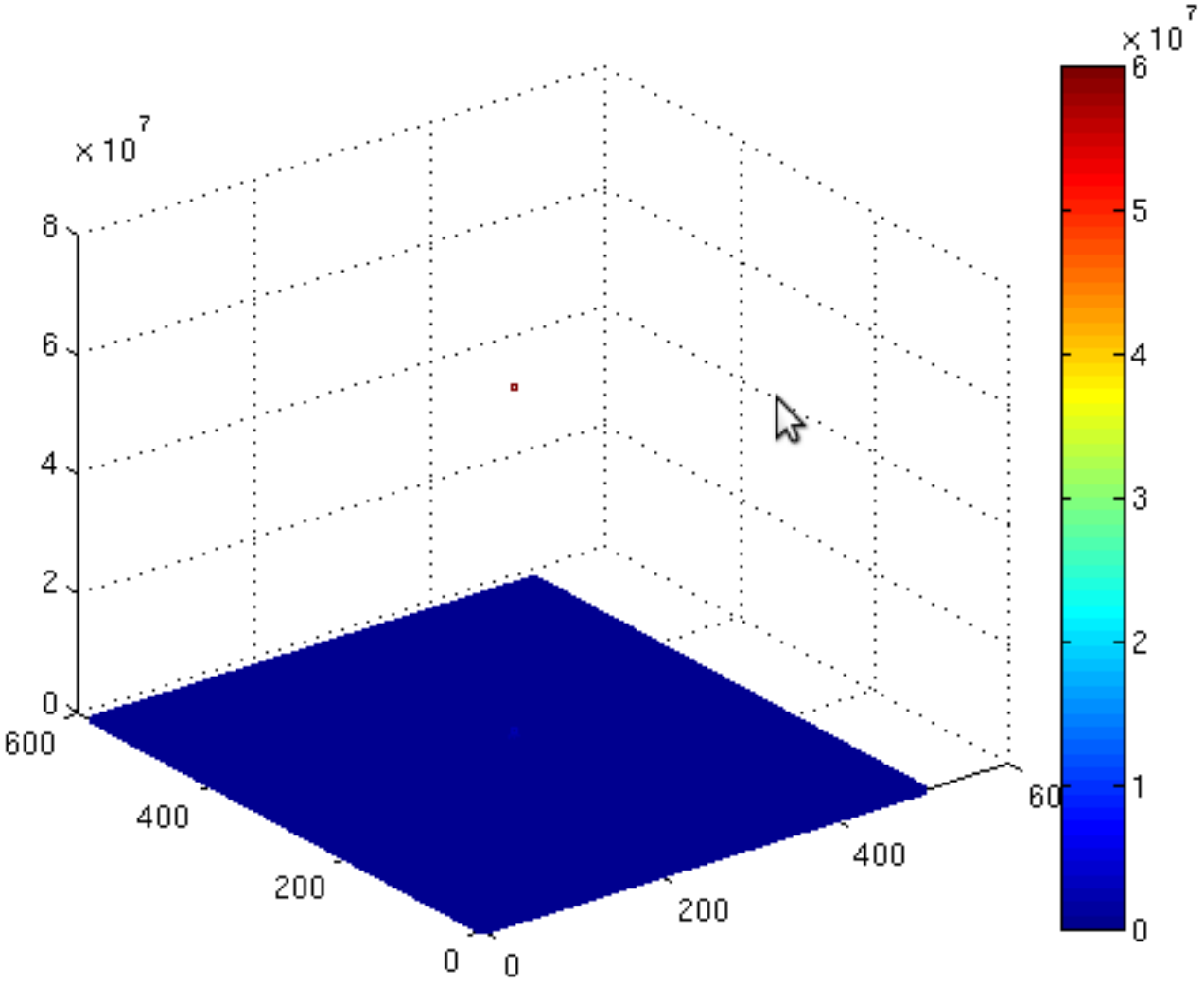
and then it searches for this pattern across the entire target image, having as a result a likelihood map. This process is similar to building a puzzle: trying to find the proper pieces, namely a set of descriptors, from the template image to build/match the regions from the target image [BI05]. Because the pattern might be found in the target image at a different scale, the template image is resized to different scales based on a Gaussian image pyramid and the pattern is extracted for each scale. Although this algorithm gives good results [SI07] in the case of object detection, it seems that its processing time is too high.

Another approach for matching descriptors across images is proposed in [SM11]. It reduces the size of the descriptors by PCA, then it simplifies their representation by using a non-linear mapping that stretches the values of the descriptors to the extreme ends so that the descriptors become more "binarized" [SM09], like the local binary patterns. The target image is then divided into a set of overlapping patches of the same size as the template image and their corresponding feature matrices are compared using MCS, as in [SM09]. If training data is available, the MCS measure can be replaced by the one shot similarity measure [SM11], especially in the case of face matching.

Whereas the approach based on the modified GHT handles by default small local deformations, noise from the image and occlusions, this is achieved in [SM11], [SM10] by adjusting the weights for the MCS in order to select only the interesting points from the face image. The adjustment of these weights is done by applying a non-linear mapping of features so that the histogram of the transformed features is more uniform and the features are more binary like.

In order to have a set of eigenvectors as a basis for PCA, 120 face images are employed in [SM11], whereas the modified GHT approach does not require any training data. On the other hand, reducing the dimensionality of the features, by retaining only the salient characteristics, makes the approach from [SM11] more suitable for

real-time application. The dimensionality reduction stems from the fact that the descriptors are redundant, their redundancy being a side effect of embedding a multitude of details. In order to deal with large out-of-plane rotations, a mirror reflected version of the features matrix is used in [SM11] and only the one that has a greater similarity is considered for further processing.

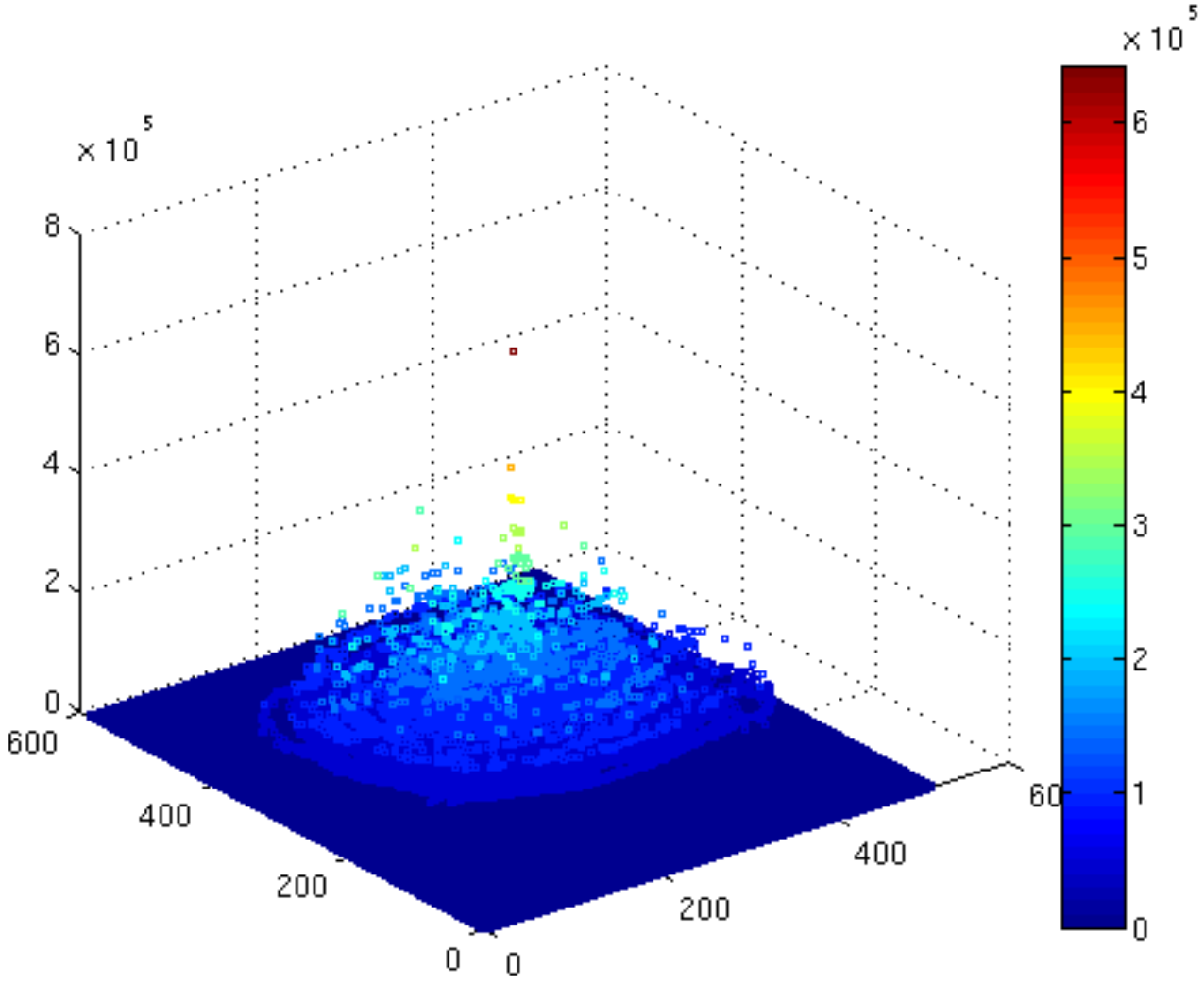


(a) A face image is compared against itself using GHT.

Figure 3.3

The representation of similarity between two images can be done by employing a salience map. Inspired by this, the accumulator of the modified GHT combined with

the GDD can be visualized in Figure 3.3 by using a scatter plot of the points from the accumulator and assigning them a colour based on their values and the colour map.

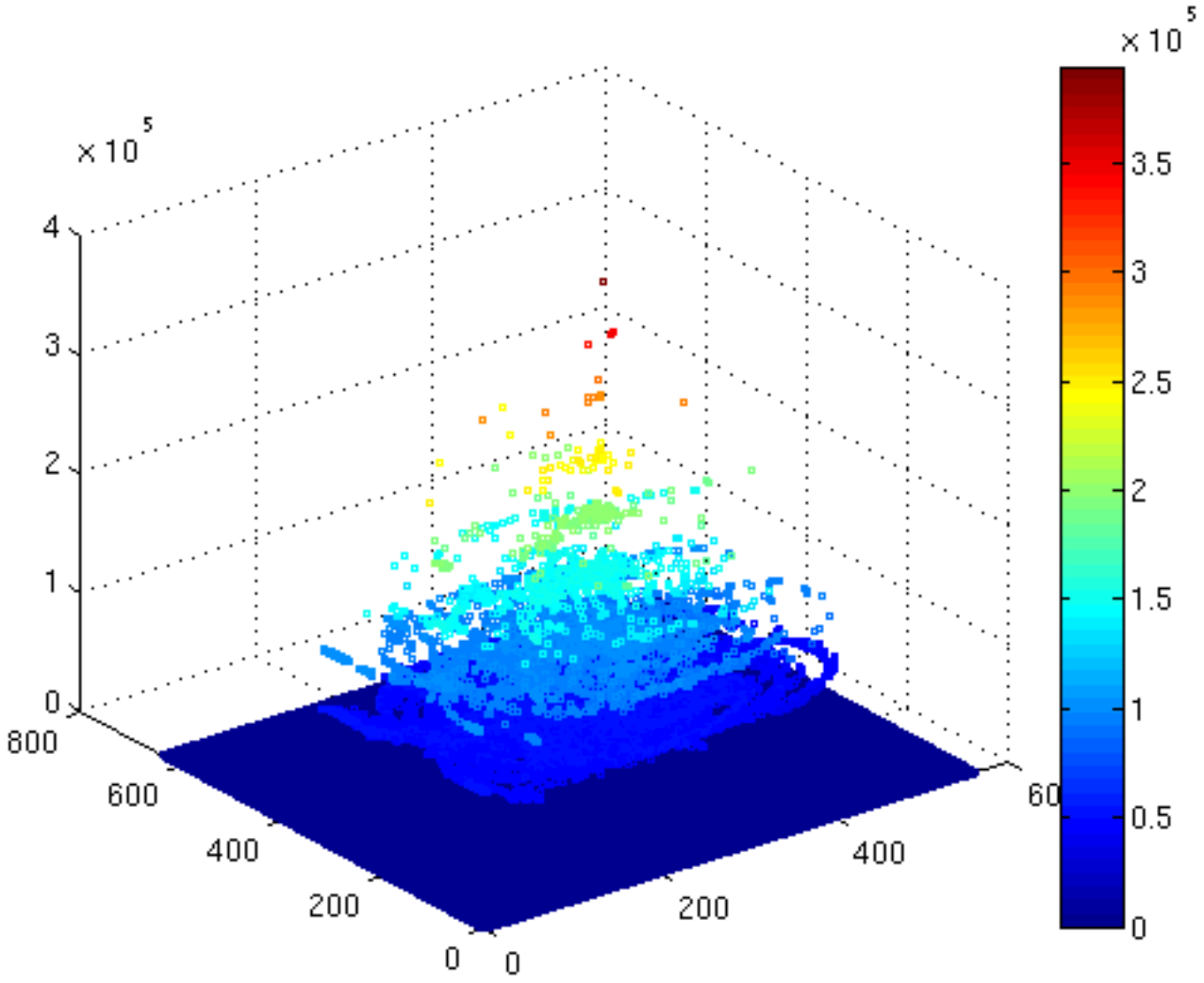


(b) Two distinct face images of the same person are compared using GHT.

Figure 3.3

In Figure 3.3a, the two face images are the same and it can be seen that there are no points in the accumulator having a value greater than zero, except for the reference point. Figure 3.3b represents the case when the images correspond to the same person but they are taken under different conditions. There are only a few other peaks having a value higher than 3×10^5 and less than 6×10^5, the last value being

the highest hit rate which sets the new location of the reference point inside the new image and also represents the certainty that the two faces are identical.



(c) Two face images corresponding to different individuals are compared using GHT.

Figure 3.3: Hough accumulator vizualization (original in color).

In Figure 3.3c, which represents the accumulator when the image of the same person is compared with a different one, it can be clearly seen that there are lots of values in the middle range, between 1.5×10^5 and 3×10^5, and as a result, the votes are more dispersed and do not produce a high hit rate for the reference point. Since the difference between the second and the third case is significant, namely, 2.5×10^5,

this metric can be used to discriminate between individuals.

3.4 Modified Generalized Hough Transform

This novel approach to face recognition uses the Hough transform to search not only for the face sketch, but also for regions which resemble each other in the new space defined by the local image descriptors. The transform takes into account not only the position of the points, but also the value of their corresponding descriptors, which are compared against one another using the matrix cosine similarity measure.

The R-table of the GHT is modified in order to accommodate the descriptors corresponding to different gradient orientations ϕ_i along the face sketch, as depicted in Table 3.1.

i	ϕ_i	$\vec{r}_{\phi_i}$	D_{ϕ_i}
0	0	$\vec{r} = y\vec{x}, x \in B, \phi(x) = 0$	$D_{\phi_i} = GDD(x), x \in B, \phi(x) = 0$
1	$\Delta\phi$	$\vec{r} = y\vec{x}, x \in B, \phi(x) = \Delta\phi$	$D_{\phi_i} = GDD(x), x \in B, \phi(x) = \Delta\phi$
2	$2\Delta\phi$	$\vec{r} = y\vec{x}, x \in B, \phi(x) = 2\Delta\phi$	$D_{\phi_i} = GDD(x), x \in B, \phi(x) = 2\Delta\phi$
$\vdots$	$\vdots$	$\vdots$	$\vdots$

Table 3.1: Modified R-table structure.

Given a query image Q and a target image T, one should determine how high is the similarity between them. First, the face sketches of the images are produced using a Canny edge detector and for each of the points in these face sketches, a descriptor is computed. Then the modified R-table is computed only for the query image based on the following parameters: the reference point of the face sketch y, chosen to be the center of gravity of the edge points, the vector from the reference point to each point from the face sketch $\vec{r}_{\phi_i}$, the gradient orientation ϕ_i of these points and the descriptors D_{ϕ_i} associated to the points from the face sketch having the gradient orientation ϕ_i.

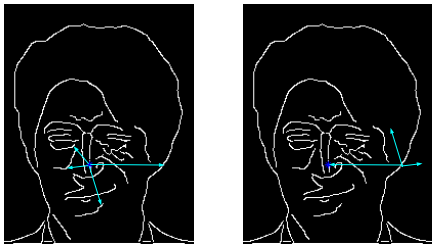


Figure 3.4: How modified GHT works: each point votes for the possible location of the reference point based on its position, orientation ϕ_i, descriptor D_{ϕ_i} and the corresponding vectors $\vec{r}^{j}_{\phi_i}$ and descriptors $D^{j}_{\phi_i}$ resulting from the R-table look-up (original in colour).

The descriptor of each point residing on the target face sketch in the cluster of gradient orientation ϕ_i is compared, based on the matrix cosine similarity and a set threshold, with the descriptors of the points from the query face sketch residing in the same cluster of gradient orientation ϕ_i. Based on the location of the points from the target image which have similar descriptors as those in the query image, and the modified R-table generated from the query image, the corresponding set of points from the accumulator array are incremented with the value of one plus a variable based on how close their descriptors are.

The highest hit rate from the accumulator array will correspond to the new reference point of the face sketch of the target image and it is used later to discriminate between folks: the higher it is, the better the image will be ranked. The idea behind this is that each point from the face sketch casts votes for possible positions of the reference point and the missing points, in case of occluded faces, do not matter as long as there are enough remaining points to agree on the reference point location.

As the number of edge points corresponding to each face sketch might vary significantly, a formal complexity analysis of Algorithm 17 is very difficult. The average processing time when comparing two face images from the Yale database is roughly

Input: Q(query image), R-table for T(template image)
Output: maximum value from the accumulator array
hitRate = 0; cst = 10^6; $\epsilon = 0.05$; padding = maximum distance from R-table
for all points $P(x_c, y_c)$ on the Q face sketch **do**
  compute its gradient orientation ϕ_i
  compute its descriptor D_c
  get all the vectors $\vec{r}^{j}_{\phi_i}$ and descriptors $D^{j}_{\phi_i}$ from the R-table row for ϕ_i
  for all points j resulting from the R-table look-up for orientation ϕ_i **do**
    compute the similarity δ between descriptors D_c and D_{ϕ_i}
    if $\delta < \epsilon$ **then**
      compute the estimated reference point $R_e(x0, y0) = P(x_c, y_c) - \vec{r}^{j}_{\phi_i} + padding$
      accum(x0,y0) += round($(\epsilon - \delta) \times$ cst)+1
      hitRate++
    end if
  end for
end for
return R = houghpeak(accum)-padding, the new reference point

Algorithm 1: Modified GHT algorithm.

20 seconds.

Consider the case of Figure 3.4 when both the template and target image are the same and the maximum value from the accumulator corresponds to the proper location of the detected face. Although most of the points along the face contour vote for this position from the accumulator array, it can be seen in the left picture in Figure 3.4 that these points might also vote for other positions in the accumulator array and the only thing that impedes them is their associated descriptor. The R-table for the modified GHT has additional information about the descriptors from the template image corresponding to different orientations and this helps to discard part of these incorrect votes, but this is not enough in some cases and additional filtering criteria have to be imposed, such as tuning of several parameters, using additional descriptors or preprocessing the input image so that only the relevant parts are extracted and the image is somewhat invariant.

3.5 Analytical Comparison Between Eigenfaces, Fisherfaces and Modified GHT

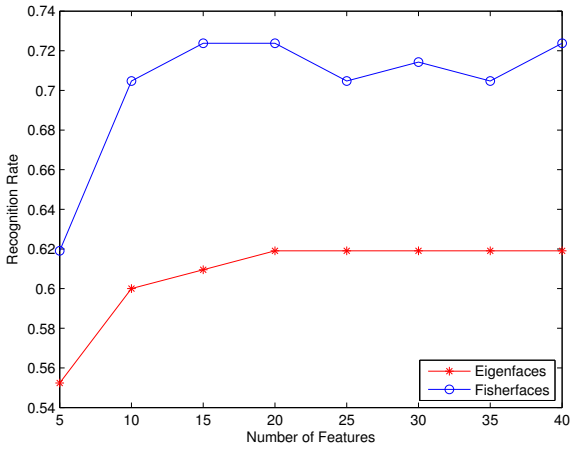


Figure 3.5: A comparison between Eigenfaces and Fisherfaces for different number of eigenvectors on a set of 105 test images and employing 60 training images from Yale database (original in colour).

The optimal method between Eigenfaces and Fisherfaces is Fisherfaces as it has several properties that intuitively suggest that it should fare well in a variety of circumstances, most notably the fact that it eliminates intra-class differences from its feature set. This suggests that it is close to optimal in deciding exactly what features are relevant to a particular class, given enough examples of that class. The implementation of the algorithms is based on [BHK97].

Fisherfaces method appears to be the best at simultaneously handling lighting variation, facial expression variation and the presence of glasses [BHK97]. As expected, Eigenfaces suffers when confronted with variation in facial expression and presence of glasses [BHK97].

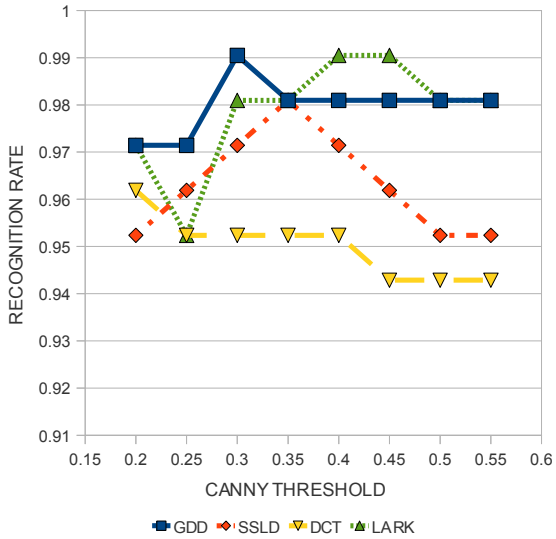


Figure 3.6: A performance comparison of different descriptors employed by the GHT for the task of face recognition on a set of 105 test images from Yale database (original in colour).

While Eigenfaces and Fisherfaces are two holistic face recognition methods based on PCA, the modified GHT takes into account not only the global shape of the face but also the local structure represented by the descriptors. The strength of PCA is mostly for face representation, whereas the modified GHT has a better face representation, based on different descriptors, and can also discriminate better between individuals, based on the maximum value of the accumulator. It can be seen in Figure 3.6 that the modified GHT leads to an increase in the recognition rate with at least 20% in comparison with Fisherfaces, no matter which of the descriptors is used.

One of the most significant advantages of the modified GHT is the fact that it requires no training data, whereas other methods, such as PCA, require a training set of images to build the new space of features. As the modified GHT is highly dependant on the descriptor used, one might improve it by employing more powerful descriptors, in the sense that they describe better the image content and are more discriminative, or combinations of descriptors.

Chapter 4

EMPIRICAL EVALUATION

The new approach to face recognition using the modified GHT is evaluated using the descriptors described in Section 2.3 for different tuning parameters. The used set of images is composed of 165 images, corresponding to 15 individuals, and each of the face images is compared against all the other face images from the database. Firstly, the effect of varying the size of the patch, for which descriptors are computed, is analyzed. Then, the influence of the epsilon threshold, which is needed when comparing two descriptors, is illustrated. The number of face traits that are taken into account is directly proportional with the upper threshold set for the Canny-edge detector and its variation effects on recognition rate are shown. Lastly, the number of gradient orientation bins, in which the GHT stores the descriptors, is evaluated. The gradient $\left[\frac{\partial I(x,y)}{\partial x}, \frac{\partial I(x,y)}{\partial y}\right]$ represents the variation of the grayscale intensity in the image and it is considered only for the points along the face sketch.

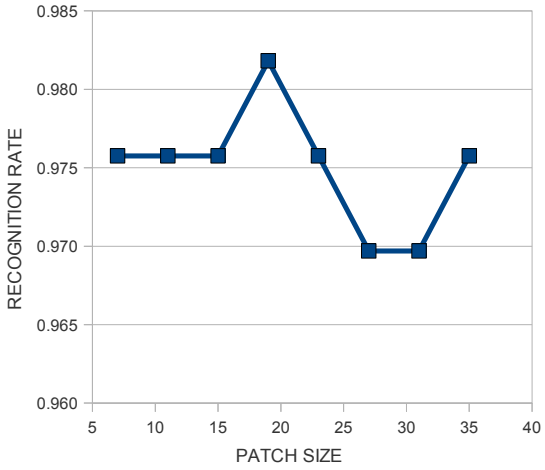


Figure 4.1: The influence of the GDD patch size on the recognition rate (original in colour).

4.1 Face Representation Based on Gradient Distance Descriptor

It seems that the recognition rate for the modified GHT employing GDD is quite steady for image patches of size between 7×7 and 35×35 pixels, accordingly to Figure 4.1. The best results are obtained when the size of the patches is close to 19×19 pixels wide. In case the size of the patch is increased too much, the processing time rises as well and the recognition rate usually drops after a specific patch size, which depends on the other parameters involved in the computation of the modified GHT.

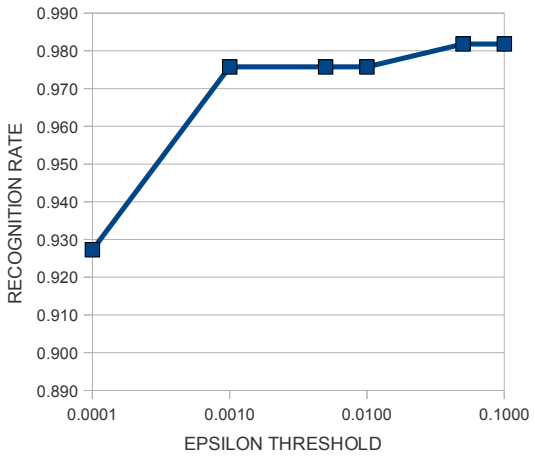


Figure 4.2: The influence of the GHT epsilon threshold on the recognition rate (original in colour).

The epsilon threshold is used when comparing two descriptors, based on a similarity metric, in order to determine whether they are similar or not. For example, in case a specific descriptor corresponding to the target image is different than all of the descriptors from the template image having a specific orientation, this descriptor is discarded and, as a result, does not vote for the location of the reference point.

It can be seen in Figure 4.2 that for values too small for the epsilon threshold, namely, 0.0001, the recognition rate drops substantially as most of the descriptors of the target image are discarded and only a few remain to vote for the possible positions of the reference point. As there are only a few descriptors, the weight of their voting is increased and any mismatched descriptor highly influences the final result. For visualization purposes, the epsilon thresholds are represented in Figure 4.2 along the horizontal axis using a logarithmic scale.

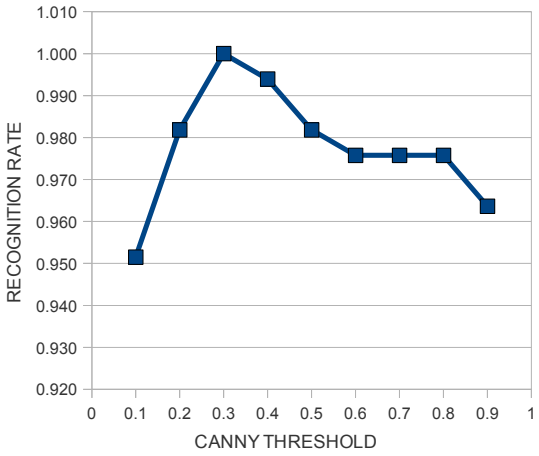


Figure 4.3: The influence of the Canny-edge detector threshold on the recognition rate (original in colour).

The upper threshold of the Canny-edge detector determines how many face traits are retained. The lower the threshold, the more details from the face are retained and the result looks like more of a detailed sketch made by an artist. Unfortunately, in this case the processing time increases and, because GDD is not a perfect hash function and MCS is not a perfect similarity measure, descriptors can have a high MCS value although they are produced by quite different image patches. This fact is illustrated in Figure 4.3 where the recognition rate drops to 157/165 for a threshold of 0.1. On the other hand, retaining only a few points from an image has a similar effect because there are only a few descriptors voting for the reference point and the mismatched descriptors highly influence the results. In case this critical mass of descriptors is not met, as in Figure 4.3, for a Canny upper threshold of 0.9, the recognition rate drops to 159/165.

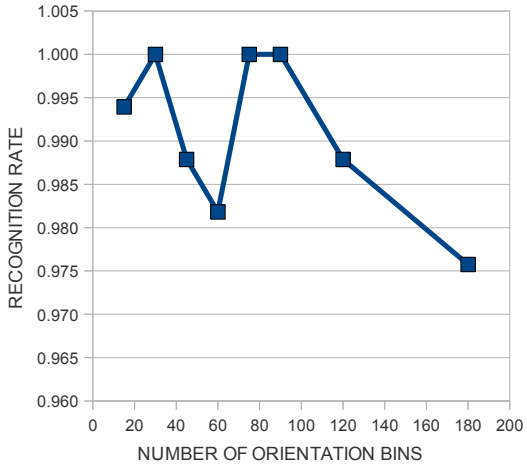


Figure 4.4: The influence of the number of gradient orientation bins on the recognition rate (original in colour).

The GHT clusters the descriptors of the points based on the gradient orientation of the points. If there are too many bins, then the gradient orientation plays a more important role and the face images have to be aligned for a good recognition rate. It can be seen in Figure 4.4 that without alignment the recognition rate drops to 161/165 for a number of 180 bins, whereas for a number of 20 bins, the recognition rate is almost perfect, 164/165. The number of bins represents how high the maximum difference between the gradient orientation of two points from the same bin can be, so for 180 bins, a maximum difference of 2 degrees is allowed, whereas for only 20 bins, the maximum allowed difference is of 80 degrees. Having fewer bins has the advantage of making the method more robust to deformations and affine transformations, but, on the other hand, it increases the processing time as more descriptors reside in the same bin and, as a result, more comparisons between descriptors have to be performed.

4.2 Face Representation Based on Locally Adaptive Regression Kernels

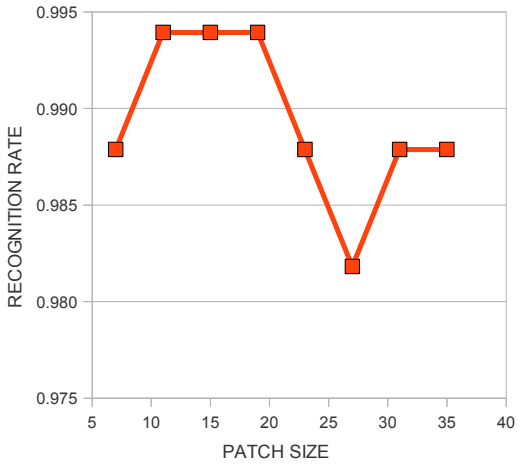


Figure 4.5: The influence of the LARK patch size on the recognition rate (original in colour).

The best recognition rates for the modified GHT employing the LARK descriptor are obtained for image patches of size between 11×11 and 23×23 pixels, according to Figure 4.5. As the LARK descriptor is a normalized descriptor, it performs well even for big image patches. The recognition rate drops significantly in Figure 4.6 when the epsilon threshold gets smaller, as the total number of matched descriptors is below a critical mass and the mismatched descriptors have a higher influence on the final values stored in the GHT accumulator.

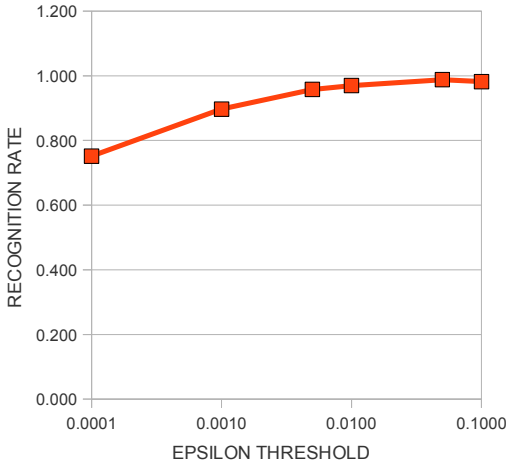


Figure 4.6: The influence of the GHT epsilon threshold on the recognition rate (original in colour).

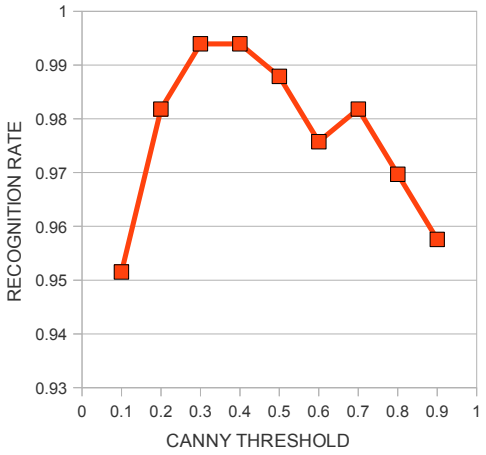


Figure 4.7: The influence of the Canny-edge detector threshold on the recognition rate (original in colour).

For a very low threshold for the Canny-edge detector, the recognition rate drops significantly in Figure 4.7 because similar LARK descriptors might result for quite different image patches or for image patches that overlap. If the imposed threshold is too high then less descriptors are computed and the recognition rate drops as well,

as the critical mass of descriptors needed for voting is not met and the mismatched descriptors have a higher influence.

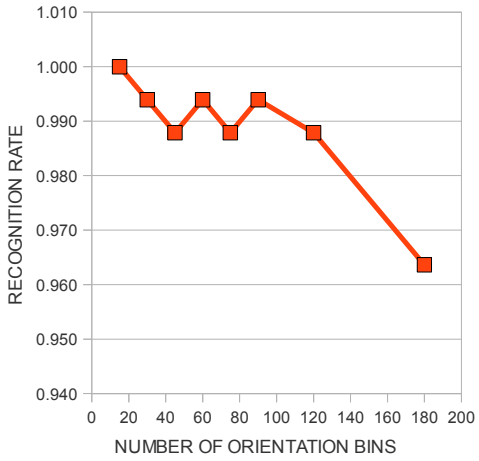


Figure 4.8: The influence of the number of gradient orientation bins on the recognition rate (original in colour).

As the number of gradient orientation bins decreases, more descriptors are located in each of the bins and this increases the chance of matching descriptors. It can be seen in Figure 4.8 that the overall classification accuracy is 100 % for only 15 bins and it drop to almost 96 % for 180 bins as it requires an alignment error smaller than 2 degrees.

4.3 Face Representation Based on Discrete Cosine Transform

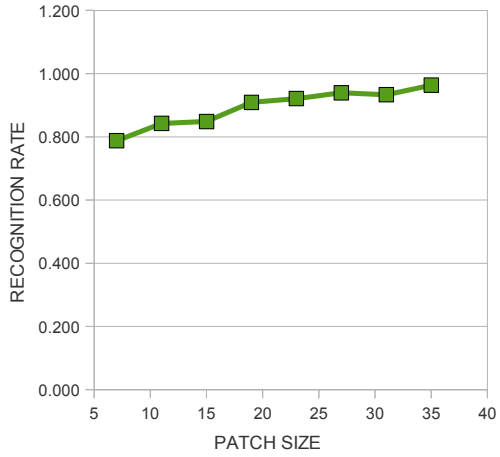


Figure 4.9: The influence of the DCT patch size on the recognition rate (original in colour).

According to Figure 4.9, the recognition rate is increasing as the size of patches expands because the DCT descriptor is a better representational descriptor but it is not discriminative enough when used in combination with the MCS metric. In order to make it more discriminative, a higher number of coefficients should be retained, but on the other hand, retaining too many of the high frequency coefficients might have a negative impact on the recognition rate. The high frequency coefficients of the DCT are associated with higher level of details or with image noise.

The threshold used when comparing the descriptors has a very high influence on the recognition rate, illustrated in Figure 4.10. As the threshold value decreases, less descriptors are voting for the reference point and, because the critical mass of descriptors is not met, any spurious matching is influencing the final result and it can go even to the extreme case when all the values from the accumulator are zero because all the descriptors are discarded.

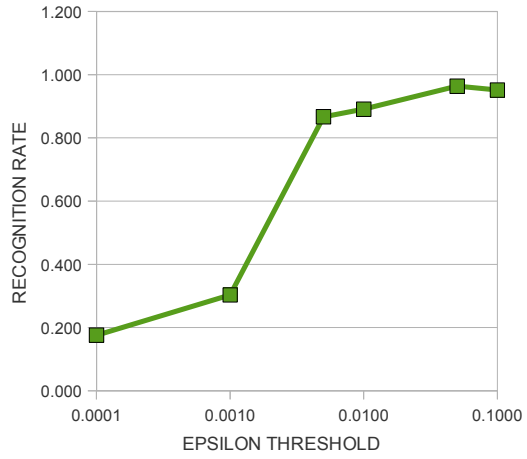


Figure 4.10: The influence of the GHT epsilon threshold on the recognition rate (original in colour).

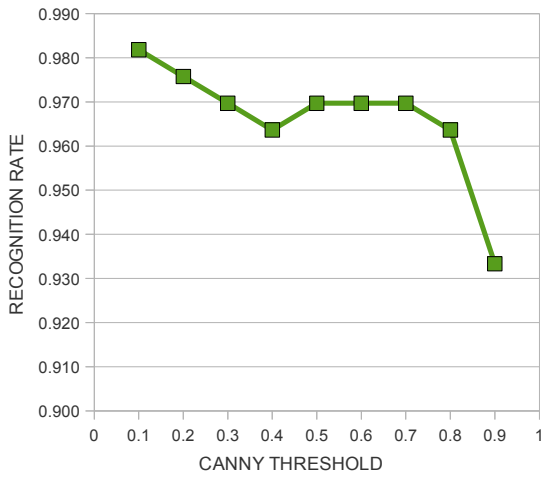


Figure 4.11: The influence of the Canny-edge detector threshold on the recognition rate (original in colour).

The DCT descriptors represent well the underlying structures of an image and, as a consequence, the recognition rate in Figure 4.11 is quite stable, around 160/165, for most of the thresholds imposed for Canny-edge detector, with the exception of the very high thresholds when the GHT has not enough points for the voting schema

and the recognition rate drops significantly.

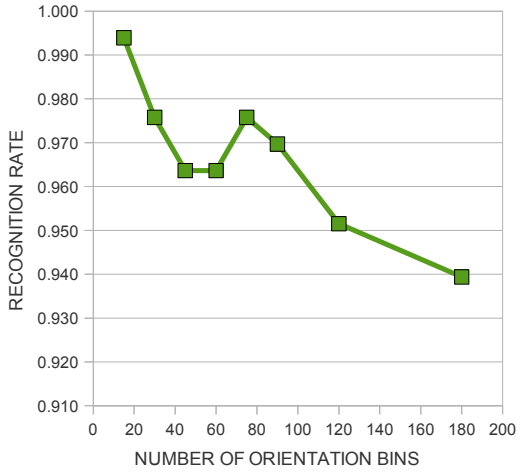


Figure 4.12: The influence of the number of gradient orientation bins on the recognition rate (original in colour).

As the DCT descriptor is not very discriminative, the number of descriptors that are matched is very important and can help increase the recognition rate as it can be seen in Figure 4.12. As the number of orientation bins decreases, there are more descriptors clustered in each of the bins and this increases the chance of matching descriptors by compensating for the imprecise alignment of the two face images.

As the number of DCT coefficients increases, the newly added harmonics to the DCT might encode either the finest details of the face or image noise. The experiment from Figure 4.13 shows that the added coefficients have a negative impact on the recognition rate because of an increased number in mismatched descriptors.

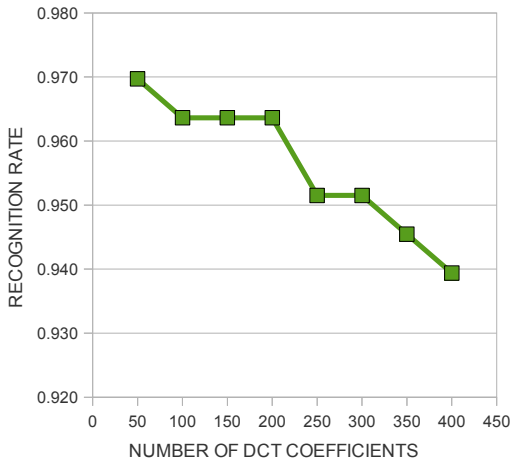


Figure 4.13: The influence of the number of DCT coefficients on the recognition rate (original in colour).

4.4 Face Representation Based on Self Similarities Local Descriptor

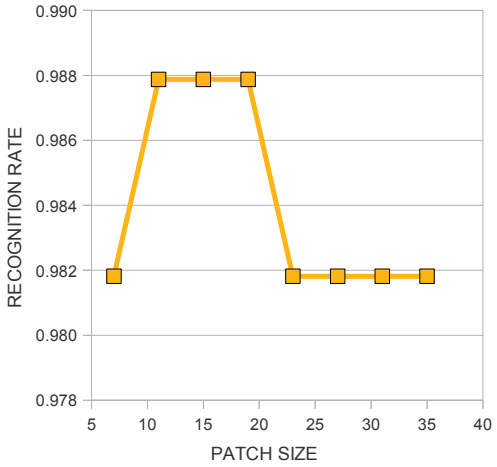


Figure 4.14: The influence of the SSLD patch size on the recognition rate (original in colour).

From Figure 4.14, it can be seen that the modified GHT is not significantly influenced by the variation of the image patch size that is employed to compute the SSLD descriptor. The recognition rate is the highest, 163/165, for image patch sizes between 11×11 and 23×23.

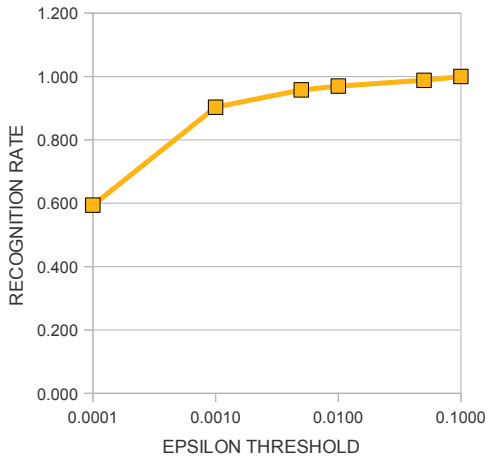


Figure 4.15: The influence of the GHT epsilon threshold on the recognition rate (original in colour).

The recognition rate in Figure 4.15 drops to 98/165 for a very small epsilon threshold, which is used when comparing descriptors using the MCS. By employing this small threshold of a value of 0.0001, too few descriptors are meeting this threshold constraint and this results in inaccurate results, as the influence of the remaining descriptors is substantially increased and any mismatched descriptor counts. In order to speed up the whole process and also increase the accuracy, a Canny upper threshold between 0.25 and 0.5 is recommended when using the SSLD, as shown in Figure 4.16.

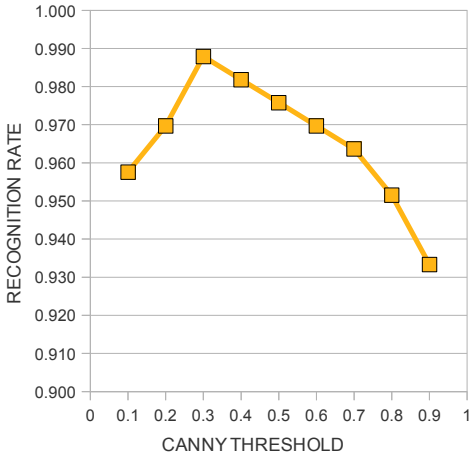


Figure 4.16: The influence of the Canny-edge detector threshold on the recognition rate (original in colour).

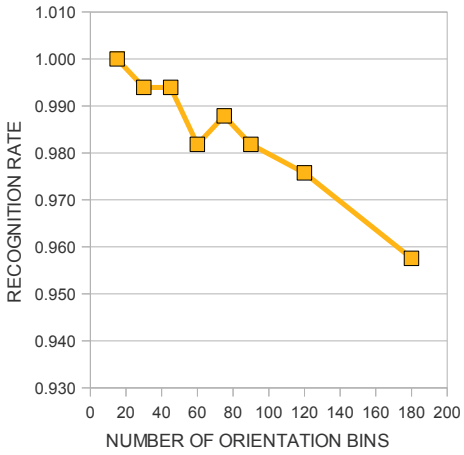


Figure 4.17: The influence of the number of gradient orientation bins on the recognition rate (original in colour).

The number of GHT orientation bins greatly affects the recognition rate of the modified GHT when employing the SSLD. It can be seen in Figure 4.17 that as the number of bins increase, the recognition rate decreases as the faces are not perfectly aligned and the method requires a higher localization precision, up to 2 degrees vari-

ation for 180 bins. In case the faces are not well aligned, the recognition rate drops because a lower number of matched descriptors vote for the reference point. The decreased number of matched descriptors is a results of the fact that the SSLD is not able to compensate for large angle rotations.

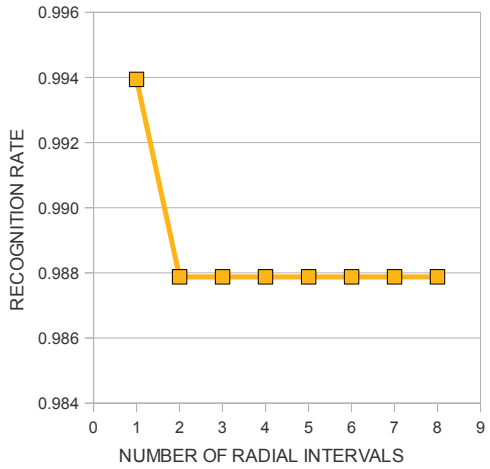


Figure 4.18: The influence of the number of SSLD radial bins on the recognition rate (original in colour).

It can be inferred from the Figure 4.18 that the recognition rate is not affected by the number of radial bins used in the log-polar transform when compressing the SSLD descriptor. As the number of radial bins increases, more details are retained by the descriptor, but too many details might be detrimental when comparing the descriptors using the MCS metric with an epsilon threshold that is too small.

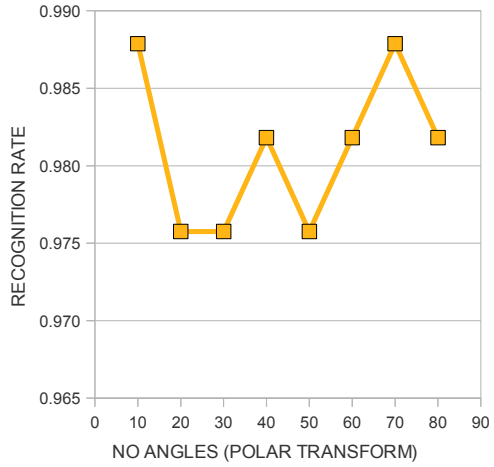


Figure 4.19: The influence of the number of SSLD angular bins on the recognition rate (original in colour).

4.5 Comparison of the descriptors performance

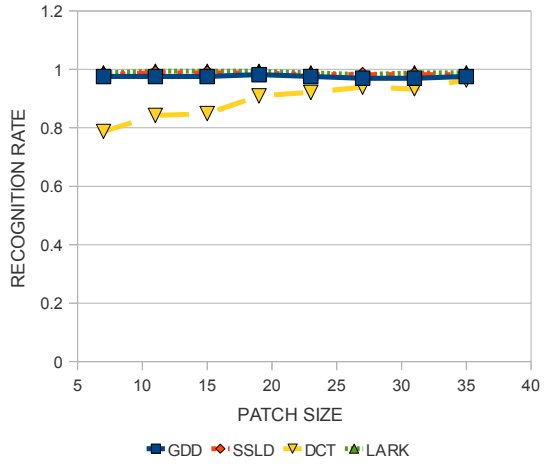


Figure 4.20: The influence of the patch size on the recognition rate for different descriptors (original in colour).

The patch size influence on the recognition rate is quite insignificant, except for the DCT descriptor. In order to improve the discriminative power of the DCT descriptor,

a higher number of coefficients should be retained, but on the other hand, retaining too many of the high frequency coefficients might have a negative impact on the recognition rate, as these coefficients are associated not only with high level of details but also with image noise.

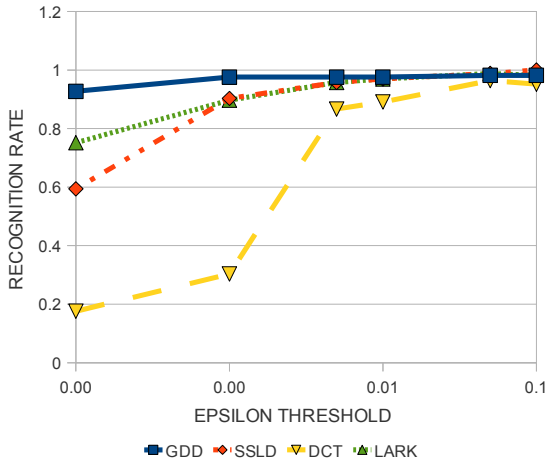


Figure 4.21: The influence of the epsilon threshold on the recognition rate for different descriptors (original in colour).

The performance of the descriptors for different epsilon thresholds reflects how discriminative they are. It can be seen in Figure 4.21 that GDD is the most discriminative descriptor, as it is quite steady under varying epsilon threshold, whereas the DCT is the least discriminative, mainly because of the low frequency coefficients.

The most important threshold is the one set for the Canny-edge detector, as it directly influence the number of face traits extracted and, therefore, the accuracy and processing time of the algorithm. It can be seen in Figure 4.22 that all descriptors perform well, with GDD having the best performance.

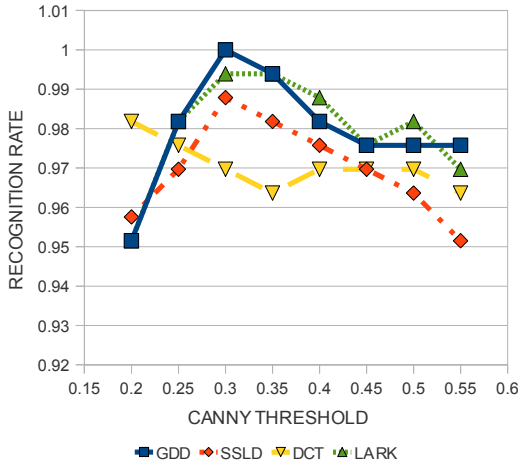


Figure 4.22: The influence of the Canny upper threshold on the recognition rate for different descriptors (original in colour).

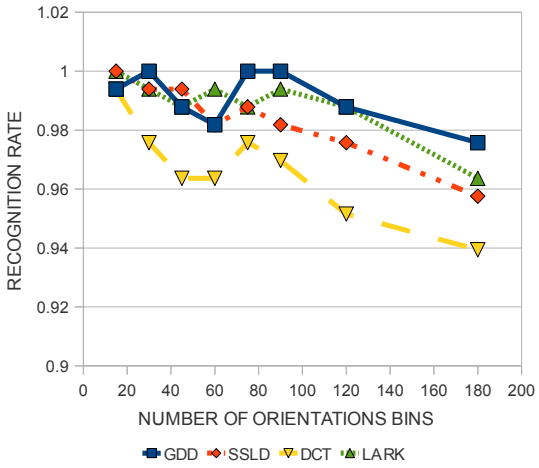


Figure 4.23: The influence of the number of gradient orientation bins on the recognition rate for different descriptors (original in colour).

The number of gradient orientation bins affects both the processing time and accuracy of the algorithm. The results from Figure 4.23 show again the superiority of the GDD descriptor and its high discriminative power.

Chapter 5

CONCLUSION

5.1 Summary

While the holistic approaches to face recognition, such as Eigenfaces and Fisherfaces, are mainly for face representation, the modified GHT comes up not only with a better local face representation, based on the local structure captured by descriptors, but also with a better model for discrimination between people based on the global shape of the face.

One of the most significant advantages of the modified GHT is the fact that it does not require any training data. For the task of face identification, the proposed method simply compares the template image with the other faces from the database and yields the highest score for the face image that resembles the template image the most. Moreover, the proposed approach to face recognition can handle partial occlusions, illumination changes and small deformations.

Based on the work of [SM11], a new descriptor is proposed, namely GDD, and its performance for the task of face identification on the Yale face database is compared the other descriptors in Figure 3.6 and a more in-depth analysis of it is done in Section 4.1, showing a slight improvement in recognition rate in comparison with the LARK

descriptor. The evaluation of its performance on the Yale face database in Section 4.1 proves that it performs better than the other descriptors and it increases the recognition rate with at least 20% in comparison with Fisherfaces for the task of face identification based on the new approach.

An interesting feature of the modified GHT approach to face recognition is that there is no need to select just one descriptor type and better results are likely to be obtained by using a couple of descriptors that provide good invariance to many transformations and also an increased discrimination power. As descriptors become more powerful, in the sense that they better describe the image content and are more discriminative, this method will improve accordingly. Moreover, it can be extrapolated for the task of object recognition and template image matching.

5.2 Future Work

A set of "attribute classifiers" [KBBN09] might reduce the number of comparisons between the query image and target images from the database and substantially decrease the searching time. These are binary classifiers trained to differentiate the individuals based on traits such as gender, race, age, hair color, flash, shiny skin etc. For example, in case the template image corresponds to a woman and the target image corresponds to a man, the classifier can detect it and discard this target image so that the modified GHT is computed for a lower number of images.

The task of identifying multiple faces in an image can be tackled by employing the Hough forests method described in [BLK12], but at an increased computation time in comparison with the GHT. Another way of handling multiple faces is by employing a face detector as a preprocessing method and then feeding these cropped faces to the modified GHT.

Instead of searching sequentially in each bin, corresponding to a specific gradient

orientation, a better approach is to index the descriptors from each of these bins in order to speed up the search by using a KD-tree as in [VCF07].

The modified GHT can take advantage of the similarity measure presented in [MNJ08] for comparing the descriptors. The basic idea behind it is to construct a tree structure that can sparsely represent the descriptors for the purpose of classification.

A good study that compares the performance of different similarity measures for image color and texture, with performance evaluated in the case of classification, image retrieval and segmentation, is presented in [PBRT99]. Based on these similarity measures and the set of predefined descriptors, one might come up with a modified GHT that is faster and has a higher recognition rate.

Alignment of the face images is required in the case of databases containing misaligned faces because most of the descriptors are not invariant to affine transformations applied to images. One of the best alignment techniques is based on mutual information and does not require information about the surface properties of the object, except its shape, and is robust to variations in illumination. The method is based on the mutual information between the model and the image, and it works well in domains where edge or gradient-magnitude based methods have difficulty, yet it is more robust than traditional correlation according to [VW95]. Mutual information is preferred to joint entropy because it better characterizes the regions of the image having a low contrast and, therefore, results in better alignment of the images.

Congealing is another method used for unsupervised image alignment which employs the raw intensity of the pixels. It is building a "distribution field" [HJLM07] that minimizes the entropy of the initial training set of images by applying different affine transformations. This new model is then used to align a new image with the existing ones. In order to deal with variable lighting, physical changes, occlusions and complex backgrounds, an improved solution is proposed by the same author in [HJLM07]. The novel idea arises from the fact that the SIFT descriptor can better

describe an image and it is more robust to changes than the raw intensities of the pixels.

A more efficient method for face alignment is presented in [TZP11] and is robust not only to non-uniform illumination, but also to occlusions. This algorithm is based on the maximization of gradient correlation coefficient, which represents the underlying image structures better than other approaches based on mere pixel intensities.

The local binary patterns (LBP) descriptor is a texture descriptor that extracts the features from different regions of the image and then concatenates them to build the descriptor, and it can be employed to speed up the computation of the descriptors.

Stemming from the fact that SIFT descriptor can be used for face recognition [BLGT06], the modified GHT might take advantage of this powerful descriptor which is invariant to scaling and rotation and can handle affine distortions, image noise, changes of the viewpoint and illumination [Low04].

Considering the good results for face detection from [CSH+10], *Weber Local Descriptor* (WLD) is a texture feature that is worth considering for the task of face recognition using the modified GHT. It is based on the fact that humans perceive patterns based not only on changes in the stimuli intensity, but also on the initial intensity of the stimuli which is taken as a reference point. This is similar to the fact that in a noisy atmosphere, one has to scream in order to be heard, whereas in a quiet environment, a whisper is enough.

Another powerful descriptor is presented in [CLZY08] and it is robust to a series of deformation such as fisheye lens, nonrigid, affine and other synthetic deformations. It is computed by combining the descriptors corresponding to regions with different sizes that are centred at the current pixel. Moreover, it has an associated similarity measure called local-to-global similarity which relies on regions of multiple sizes.

The accuracy of the modified GHT might be increased by combining multiple descriptors, as each descriptor encodes specific characteristics of an image and is

invariant to certain affine transformations, and by employing specific comparison metrics for each descriptor.

REFERENCES

[ACE07] M. Anelli, L. Cinque, and S. Enver. Deformation tolerant generalized Hough transform for sketch-based image retrieval in complex scenes. *Image and Vision Computing*, 25(11):1802–1813, November 2007.

[AHP06] T. Ahonen, A. Hadid, and M. Pietikainen. Face description with local binary patterns: Application to face recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(12):2037–2041, December 2006.

[APHM04] T. Ahonen, M. Pietikainen, A. Hadid, and T. Maenpaa. Face recognition based on the appearance of local regions. In *Proceedings of the 17th International Conference on Pattern Recognition*, volume 3, pages 153–156, August 2004.

[Bal81] D.H. Ballard. Generalizing the Hough transform to detect arbitrary shapes. *Pattern Recognition*, 13(2):111–122, 1981.

[BHK97] P.N. Belhumeur, J.P. Hespanha, and D.J. Kriegman. Eigenfaces vs. fisherfaces: recognition using class specific linear projection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(7):711–720, July 1997.

[BI05] O. Boiman and M. Irani. Detecting irregularities in images and in video. In *Tenth IEEE International Conference on Computer Vision*, volume 1, pages 462–469, October 2005.

[BLGT06] M. Bicego, A. Lagorio, E. Grosso, and M. Tistarelli. On the use of SIFT features for face authentication. In *Computer Vision and Pattern Recognition Workshop*, page 35, June 2006.

[BLK12] O. Barinova, V. Lempitsky, and P. Kohli. On detection of multiple object instances using Hough transforms. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PP(99):1, 2012.

[CLZY08] H. Cheng, Z. Liu, N. Zheng, and J. Yang. A deformable local image descriptor. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, June 2008.

[CSH+10] J. Chen, S. Shan, C. He, G. Zhao, P.M. Andinen, X. Chen, and W. Gao. WLD: A robust local image descriptor. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(9):1705–1720, September 2010.

[CWS95] R. Chellappa, C.L. Wilson, and S. Sirohey. Human and machine recognition of faces: a survey. *Proceedings of the IEEE*, 83(5):705–741, May 1995.

[GW06] R.C. Gonzalez and R.E. Woods. *Digital Image Processing (3rd Edition)*. Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 2006.

[HJLM07] G.B. Huang, V. Jain, and E. Learned-Miller. Unsupervised joint alignment of complex images. In *IEEE 11th International Conference on Computer Vision*, pages 1–8, October 2007.

[KBBN09] N. Kumar, A.C. Berg, P.N. Belhumeur, and S.K. Nayar. Attribute and simile classifiers for face verification. In *IEEE 12th International Conference on Computer Vision*, pages 365–372, October 2009.

[KG09] B. Kulis and K. Grauman. Kernelized locality-sensitive hashing for scalable image search. In *IEEE 12th International Conference on Computer Vision*, pages 2130–2137, October 2009.

[Kor07] P. Kort. The use of the Hough transform in shape-from-shading. Master's thesis, University of Regina, 2007.

[KSX06] B.V.K.V. Kumar, M. Savvides, and C. Xie. Correlation pattern recognition for face recognition. *Proceedings of the IEEE*, 94(11):1963–1976, November 2006.

[LD95] M.J. Li and R.W. Dai. A personal handwritten Chinese character recognition algorithm based on the generalized Hough transform. In *Proceedings of the Third International Conference on Document Analysis and Recognition*, volume 2, pages 828–831, August 1995.

[Low04] D.G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal for Computer Vision*, 60(2):91–110, November 2004.

[LT09] L. Ling and Ö.M. Tamer, editors. *Encyclopedia of Database Systems*. Springer US, 2009.

[LZ05] Q. Li and B. Zhang. Image matching under generalized Hough transform. In Nuno Guimares and Pedro T. Isaas, editors, *International Conference on Applied Computing*, pages 45–50. IADIS, 2005.

[MK01] A.M. Martinez and A.C. Kak. PCA versus LDA. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(2):228–233, February 2001.

[MM09] S. Maji and J. Malik. Object detection using a max-margin Hough transform. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 1038–1045. IEEE, June 2009.

[MNJ08] F. Moosmann, E. Nowak, and F. Jurie. Randomized clustering forests for image classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(9):1632–1646, September 2008.

[PBRT99] J. Puzicha, J.M. Buhmann, Y. Rubner, and C. Tomasi. Empirical evaluation of dissimilarity measures for color and texture. In *The Proceedings of the Seventh IEEE International Conference on Computer Vision*, volume 2, pages 1165–1172, 1999.

[PGM+03] P.J. Phillips, P. Grother, R. Micheals, D.M. Blackburn, E. Tabassi, and M. Bone. Face recognition vendor test 2002. In *IEEE International Workshop on Analysis and Modeling of Faces and Gestures*, page 44, October 2003.

[SB07] J.W. Schneider and P. Borlund. Matrix comparison, part 1: Motivation and important issues for measuring the resemblance between proximity measures or ordination results. *Journal of the American Society for Information Science and Technology*, 58(11):1586–1595, 2007.

[Sch00] A. Schubert. Detection and tracking of facial features in real time using a synergistic approach of spatio-temporal models and generalized Hough transform techniques. In *Fourth IEEE International Conference on Automatic Face and Gesture Recognition*, pages 116–121, 2000.

[SI07] E. Shechtman and M. Irani. Matching local self-similarities across images and videos. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, June 2007.

[SM09] H.J. Seo and P. Milanfar. Nonparametric detection and recognition of visual objects from a single example, September 2009.

[SM10] H.J. Seo and P. Milanfar. Training-free, generic object detection using locally adaptive regression kernels. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(9):1688–1704, September 2010.

[SM11] H.J. Seo and P. Milanfar. Face verification using the LARK representation. *IEEE Transactions on Information Forensics and Security*, 6(4):1275–1286, December 2011.

[TP91] M.A. Turk and A.P. Pentland. Face recognition using eigenfaces. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 586–591, June 1991.

[Tri09] S. Trivedi. Face recognition using eigenfaces and distance classifiers: A tutorial. URL: http://onionesquereality.wordpress.com/2009/02/11/face-recognition-using-eigenfaces-and-distance-classifiers-a-tutorial, February 2009.

[TZP11] G. Tzimiropoulos, S. Zafeiriou, and M. Pantic. Robust and efficient parametric face alignment. In *IEEE International Conference on Computer Vision*, pages 1847–1854, November 2011.

[VCF07] E. Valle, M. Cord, and S.P. Foliguet. Matching local descriptors for image identification on cultural databases. In *Ninth International Conference on Document Analysis and Recognition*, volume 2, pages 679–683, September 2007.

[VW95] P. Viola and W.M. Wells. Alignment by maximization of mutual information. In *Fifth International Conference on Computer Vision*, pages 16–23, June 1995.