# A Robust Face Recognition Algorithm for Real-World Applications

zur Erlangung des akademischen Grades eines

Doktors der Ingenieurwissenschaften

der Fakultät für Informatik
der Universität Fridericiana zu Karlsruhe (TH)

**genehmigte**

## Dissertation

von

**Hazım Kemal Ekenel**

aus Samsun, Türkei

Tag der mündlichen Prüfung: **02.02.2009**

Erster Gutachter: **Prof. Dr. A. Waibel**

Zweiter Gutachter: **Prof. Dr. J. Kittler**

# Abstract

Face recognition is one of the most challenging problems of computer vision and pattern recognition. The difficulty in face recognition arises mainly from facial appearance variations caused by factors, such as expression, illumination, partial face occlusion, and time gap between training and testing data capture. Moreover, the performance of face recognition algorithms heavily depends on prior facial feature localization step. That is, face images need to be aligned very well before they are fed into a face recognition algorithm, which requires precise facial feature localization. This thesis addresses on solving these two main problems —facial appearance variations due to changes in expression, illumination, occlusion, time gap, and imprecise face alignment due to mislocalized facial features— in order to accomplish its goal of building a generic face recognition algorithm that can function reliably under real-world conditions.

The proposed face recognition algorithm is based on the representation of local facial regions using the discrete cosine transform (DCT). The local representation provides robustness against appearance variations in local regions caused by partial face occlusion or facial expression, whereas utilizing the frequency information provides robustness against changes in illumination. In addition, the algorithm bypasses the facial feature localization step and formulates face alignment as an optimization problem in the classification stage. Therefore, the system is free from the misalignment problem due to erroneous facial feature localization.

The algorithm's robustness against partial face occlusion, expression, illumination, time gap, and uncontrolled data capture conditions is first tested on five well-known benchmark face databases, namely on the AR, CMU PIE, FRGC, Yale B, and extended Yale B face databases. Extensive experiments have been conducted to analyze the effects of the algorithm's parameters on the classification performance. Moreover, the algorithm's robustness against image compression and registration errors is also assessed and it is compared with well-known generic face recognition algorithms. On all the experiments the algorithm attains very high correct recognition rates. It is found to be significantly superior to generic face recognition algorithms. It also outperforms, or performs as well as the algorithms that are designed specifically for just one type of factor that causes facial appearance variation, such as illumination. Experimental results show that, in the case of upper face occlusion caused by sunglasses, the main problem for low performance is not mainly because of missing eye region information but because of misalignment due to erroneous manual labeling of eye center positions. Since the algorithm is free from this problem, it also achieves very high correct recognition rates on this type of data.

Several systems have been developed based on the proposed face recognition algorithm. In addition to the tests on the benchmark face databases, these systems are also evaluated on data collected under real-world conditions. One of the systems performs person identification in smart rooms and has been evaluated within the CLEAR evaluations. Other real-world applications, door monitoring, visitor interface, person identification in movies, have also been tested extensively. These evaluations show that the algorithm can work reliably under real-world conditions. The algorithm is also extended for a 3D face recognition scheme and found to perform successfully on the 3D data.

# Zusammenfassung

Gesichtserkennung ist eines der wichtigsten Probleme in den Bereichen Maschinensehen und Mustererkennung. Das Gebiet, das die intensivsten Anstrengungen in der Gesichtserkennungsforschung angetrieben hat, sind Sicherheitsanwendungen, von Authentifizierung, z.B. zur Zugangskontrolle für elektronische Transaktionen, Computer-Login oder Internet-Zugang, bis hin zu Videoüberwachung, z.B. in Banken, Kaufhäusern oder auch im öffentlichen Raum.

Zudem ist Personenidentifikation eine der wichtigsten Komponenten für intelligente Interaktions-Applikationen. Hierbei bedeutet *intelligente Interaktion*, dass perzeptuelle Technologien eingesetzt werden, um Mensch-Mensch- und Mensch-Maschine-Interaktionen zu erleichtern. Sowohl als Assistent in Mensch-Mensch-Interaktionen, z.B. als Gedächtnisstütze, die einem sagt, mit wem man gerade redet, als auch in Mensch-Maschine-Interaktionen, z.B. eine Maschine, die ihren Benutzer identifiziert und ihre Einstellungen entsprechend anpasst, liefert Personenidentifikation das wichtigste Merkmal natürlicher Interaktionen: Personalisierung. Weiterhin kann die Identität einer Person genutzt werden, um die Leistung anderer perzeptueller Technologien zu erhöhen, wie z.B. Analyse von Gesichtsausdrücken oder Kopfdrehungen, da es durch sie möglich wird, personenspezifische Modelle zu verwenden.

Gesichtserkennung und Sprecheridentifikation sind bekanntermaßen natürliche Identifikationsmethoden, da das Gesicht und die Sprache die Modalitäten sind, die wir im täglichen Leben benutzen, um Menschen zu identifizieren. Obwohl andere Methoden, wie z.B. die Identifikation anhand von Fingerabdrücken, bessere Identifikationsleistungen erreichen können, sind sie aufgrund ihrer intrusiven Natur ungeeignet für natürliche Interaktionen. Der größte Vorteil von Gesichtserkennung ist, dass sie die Möglichkeit der passiven Identifikation bietet, die zu identifizierende Person also nicht kooperieren oder eine bestimmte Aktion ausführen muss. Zum Beispiel kann ein intelligenter Supermarkt seine regelmäßigen Kunden wiedererkennen, wenn sie den Laden betreten. Die Kunden müssen nicht sprechen oder direkt in eine Kamera schauen, um erkannt zu werden. Dies macht Gesichtserkennung zu einer idealen Wahl für natürliche Interaktions-Applikationen, da sie unauffällig im Hintergrund laufen kann, ohne die zu identifizierenden Personen zu behindern oder zu unterbrechen.

Gesichtserkennung hat in einer Vielzahl von intelligenten Interaktionssystemen Anwendung gefunden. Die Anwendungsgebiete, auf die sich diese Arbeit konzentriert, können in drei Gruppen eingeteilt werden. Die erste Gruppe besteht aus Gesichtserkennung für intelligente Umgebungen. Diese Gruppe beinhaltet Identifikationsaufgaben an einem festen Ort, z.B. in einem intelligenten Haus, das Familienmitglieder automatisch identifiziert. Die zweite Gruppe verwendet Gesichtserkennung für intelligente Maschinen. In dieser Gruppe identifiziert

eine Maschine die Person, die mit ihr interagiert, z.B. ein Auto, das seinen Fahrer identifiziert, oder ein Roboter, der die Person, die ihn bedient, wiedererkennt. Die letzte Gruppe besteht aus Gesichtserkennung für intelligente Bild- oder Videosuche. In dieser Gruppe werden Gesichtsbilder als Hinweise zur Suche nach Personen benutzt.

Gesichtserkennung ist ein sehr anspruchsvolles Problem in den Bereichen Maschinensehen und Mustererkennung. Das Abbild eines Gesichtes kann aufgrund von Unterschieden in Gesichtsausdruck, Beleuchtung, Verdeckung, Kopfdrehung und Alterung stark variieren. Die Variationen, die durch diese Faktoren verursacht werden, sind oft stärker als die Variationen zwischen Gesichtsbildern unterschiedlicher Personen. Ein anderer wichtiger Faktor, der Gesichtserkennung erschwert, ist die Registrierung von Gesichtsbildern. Um Gesichtsbilder geeignet vergleichen zu können, müssen die Positionen lokaler Merkmale zueinander passend ausgerichtet sein. Dies erfordert die präzise Lokalisierung bestimmter Gesichtsmerkmale, was eine sehr schwierige Aufgabe ist.

Eine Vielzahl potenzieller Anwendungen hat zu ausgiebigen Forschungsaktivitäten im Bereich der Gesichtserkennung geführt. Viele Algorithmen wurden entwickelt, die einen einzelnen oder eine Kombination zweier Faktoren, die Variationen in der Ansicht von Gesichtern verursachen, zu behandeln versuchen. Besonders der Behandlung von Beleuchtungsveränderungen wurde große Aufmerksamkeit zuteil [AMU97, CWX+06, GBK01, GMB04, LHK05, SRR01, ZACJ07]. Alle diese Algorithmen werden nur gegen Ansichtsvariationen evaluiert, für die sie entwickelt wurden. Zum Beispiel werden Algorithmen, die entwickelt wurden um Beleuchtungsänderungen zu behandeln, mit Datensätzen evaluiert, die nur Beleuchtungsänderungen enthalten. Daraus resultierend existieren viele Gesichtsdatenbanken, die meist unter kontrollierten Bedingungen aufgenommen wurden und die Ansichtsvariationen enthalten, die von einem einzelnen Faktor oder einer Kombination zweier Faktoren verursacht wurden. Diese Studien haben wertvolle Einblicke in verschiedene Aspekte der Gesichtserkennung geliefert und die Datenbanken, die gesammelt wurden um die Algorithmen zu evaluieren, sind sehr nützlich, um die Robustheit eines Algorithmus gegen bestimmte Quellen von Ansichtsvariationen zu bestimmen. Sie geben jedoch keine Hinweise darauf, wie gut der getestete Algorithmus unter realen Bedingungen funktioniert. Es ist notwendig und wichtig, einen Gesichtserkennungsalgorithmus auf diesen Benchmark-Datenbanken zu testen. Dies ist aber nicht genug, um zu garantieren, dass er zuverlässig unter realen Bedingungen funktioniert, auch wenn er gute Ergebnisse auf allen Datenbanken erzielt. Die Hauptgründe hierfür sind zum einen, dass in den Benchmark-Datenbanken die Ansichtsvariationen durch eine einzelne Quelle oder eine Kombination zweier Quellen verursacht werden. Unter realen Bedingungen werden die Ansichtsvariationen jedoch durch zahlreiche Quellen gleichzeitig verursacht. Zum anderen enthalten die Benchmark-Datenbanken diskrete Variationen, z.B. Kopfdrehungen bestimmter Winkel. In Wirklichkeit sind jedoch alle Kopfdrehungen, Gesicht-

sausdrücke, Beleuchtungsänderungen, usw. in kontinuierlichen Intensitäten möglich. Zudem wurden die Benchmark-Datenbanken unter kontrollierten Bedingungen mit Kooperation der aufgezeichneten Personen aufgenommen. D.h. dass der/die Proband/in angewiesen wurde in die Kamera zu schauen und er/sie wusste, dass sein/ihr Bild aufgenommen wurde. Dies ist ein passendes Szenario für eine Authentifikationsaufgabe, bei der Kooperation erwartet werden kann. In anderen Anwendungsszenarien hingegen, wie z.B. in intelligenten Räumen, wird passive, unauffällige Identifikation benötigt.

Diese Arbeit hat daher zum Ziel, einen neuartigen, generischen Gesichtserkennungsalgorithmus zu entwickeln, der robust gegen Änderungen des Aussehens von Gesichtern ist, die durch Gesichtsausdruck, Beleuchtung, Verdeckung, Altern und unkontrollierte Aufnahmebedingungen verursacht werden.

Die Hauptschritte des vorgeschlagenen Gesichtserkennungsalgorithmus werden im Folgenden kurz dargestellt.

Die diskrete Cosinus-Transformation (DCT) wird benutzt um lokale Regionen zu repräsentieren. Die Verwendung der DCT hat mehrere Vorteile: Zum einen sind die datenunabhängigen Basisfunktionen der DCT sehr praktisch in der Anwendung, da z.B. keine repräsentative Menge von Trainingsdaten vorbereitet werden muss, um einen Unterraum zu berechnen. Zum anderen liefert die DCT Frequenzinformationen, was sehr nützlich für die Behandlung von Variationen des Aussehens von Gesichtern ist. Es ist zum Beispiel bekannt, dass manche Frequenzbänder gut geeignet sind um Beleuchtungsvariationen zu kompensieren. Außerdem wird in dieser Arbeit gezeigt, dass die DCT-basierte Repräsentation lokaler Regionen für Gesichtserkennung bessere Ergebnisse liefert als Repräsentationen basierend auf den Karhunen-Loève-, Fourier-, Wavelet- oder Walsh-Hadamard-Transformationen.

Im vorgeschlagenen, auf lokalen Ansichten basierenden, Ansatz zur Gesichtserkennung, wird ein detektiertes und registriertes Gesichtsbild in $8 \times 8$-Pixel große Blöcke aufgeteilt. Danach wird die DCT auf jedem Block ausgeführt. Die resultierenden DCT-Koeffizienten werden mit dem zig-zag-scanning-Verfahren sortiert. Basierend auf einer Strategie zur Merkmalsselektion werden $M$ Koeffizienten benutzt. Das Resultat ist ein $M$-dimensionaler lokaler Merkmalsvektor. Schließlich werden die lokalen Merkmalsvektoren konkateniert, was einen Merkmalsvektor für das gesamte Bild ergibt. Die Klassifikation wird von einem Nächster-Nachbar-Klassifikator durchgeführt, der die L1-Norm als Distanzmetrik benutzt.

Der Gesichtserkennungsalgorithmus hat zwei wesentliche Punkte. Zum einen wird das Frequenzband, das zur Klassifikation verwendet wird, automatisch ausgewählt. Dazu wird zunächst die Klassifikation mit mehreren Frequenzbändern durchgeführt, indem ein Fenster der Größe $M$ über die extrahierten DCT-Koeffizienten geführt wird, und für jede Fensterposition die Klassifikation mit

den selektierten Koeffizienten durchgeführt wird. Das Frequenzband, das die besten zwei Kandidaten optimal separiert, wird als zuverlässigstes Frequenzband angenommen und zur Klassifikation verwendet. Auf diese Weise kann sich der Algorithmus durch Verwendung des passenden Frequenzbandes automatisch an veränderte Beleuchtungsverhältnisse anpassen. Der zweite Punkt betrifft die Merkmalsnormalisierung. Um die Beiträge der Koeffizienten und Blöcke für die Klassifikation vergleichbar zu machen, werden die Koeffizienten jedes Merkmalsvektors zuerst durch ihre Standardabweichungen dividiert, und danach jeder lokale Merkmalsvektor zu einem Einheitsvektor normiert.

Im vorgeschlagenen Ansatz wird die Gesichtsregistrierung durchgeführt, indem die kleinste Distanz im Klassifikationsschritt minimiert wird. Da alle Menschen dieselbe Gesichtskonfiguration haben, können die Positionen der Gesichtsmerkmale mit Hilfe der Position und Größe des Gesichtes grob geschätzt werden. Danach wird eine Suche um die geschätzten Positionen herum nach den exakten Positionen der Gesichtsmerkmale durchgeführt. Die Kandidaten für die exakten Positionen der Gesichtsmerkmale werden benutzt, um für jeden Vergleich eines Test- mit einem Trainings-Gesichtsbild mehrere Registrierungen des Gesichtes vorzunehmen. Diejenige Registrierung, die zur minimalen Distanz zwischen Test- und Trainingsbild führt, wird zur Klassifikation verwendet. Daher werden die Augenpositionen eines Trainings-Gesichtsbildes für jeden Vergleich mit einem Test-Gesichtsbild neu ermittelt, wodurch Inkonsistenzen in den manuellen Annotationen der Augenpositionen der Trainingsbilder gehandhabt werden.

Diese Arbeit konzentriert sich darauf, einen generischen, robusten Gesichtserkennungsalgorithmus zu entwickeln, der zuverlässig in realen Applikationen eingesetzt werden kann. In Richtung dieses Zieles wurden die folgenden Beiträge geleistet:

Ein Gesichtserkennungsalgorithmus wurde entwickelt, der Variationen des Aussehens von Gesichtern behandeln kann, die von Verdeckungen, Gesichtsausdruck, Beleuchtung, zeitlichem Abstand und unkontrollierten Aufnahmebedingungen verursacht werden. Der Algorithmus wurde mit Hilfe von Standard-Benchmarks eingehend unter verschiedenen Bedingungen evaluiert, und es wurde festgestellt, dass er sowohl den bekannten generischen Algorithmen, als auch den spezifischen Algorithmen, die entwickelt wurden um einen Variationsfaktor zu behandeln, signifikant überlegen ist. Der Algorithmus erreichte eine korrekte Erkennungsrate von 98,5% und 96,2% auf den Bildern der face recognition grand challenge Datenbank (FRGC) [PFS+05], unter kontrollierten Bedingungen – in einem Studio mit kontrollierter Beleuchtung – und unkontrollierten Bedingungen – unter wechselnden Bedingungen, in Gängen, Hallen oder in Außenbereichen. Die Leistung auf der AR Gesichtsdatenbank [MB98] bei Verdeckung des oberen und unteren Gesichtsteils beträgt 97,3% und 98,2%. Die erzielten Resultate bei Beleuchtungsvariationen betragen 100% auf der CMU PIE

Gesichtsdatenbank [SBB03], auf allen Beleuchtungsuntergruppen der Yale Gesichtsdatenbank B [GBK01], sowie auf der zweiten und dritten Beleuchtungsuntergruppe der erweiterten Yale Gesichtsdatenbank B [LHK05], während die Resultate der vierten und fünften Beleuchtungsuntergruppe bei 98,7% und 99,0% liegen. Es ist das erste Mal in der Literatur, dass alle Variationen des Aussehens von Gesichtern von einem generischen Algorithmus behandelt werden, d.h. ohne individuelle Algorithmen für jede der Variationsquellen zu entwickeln.

Anders als bei konventionellen Gesichtserkennungssystemen benötigt der vorgeschlagene Algorithmus keinen zusätzlichen Schritt zur Lokalisierung von Gesichtsmerkmalen, um die Registrierung durchzuführen. Er führt die Lokalisierung implizit während des Klassifikationsschrittes durch. Weiterhin wurde gezeigt, dass der vorgeschlagene Registrierungsansatz sogar besser funktioniert als die Registrierung mit manuellen Annotationen. Z.B. betragen die Ergebnisse auf der AR Gesichtsdatenbank [Mar02] bei Verdeckung der unteren Gesichtshälfte bei Registrierung mit manuellen Annotationen 91,8%, während die Resultate mit dem vorgeschlagenen Ansatz bei 97,3% liegen. Es wurde ebenfalls gezeigt, dass das Hauptproblem bei Verdeckungen der oberen Gesichtshälfte Fehler in der Registrierung sind, und nicht die Verdeckung selbst. Wegen einer Sonnenbrille können Augenpositionen, die weithin zur Registrierung verwendet werden, selbst manuell nicht zuverlässig annotiert werden. Wenn nur die manuellen Annotationen zur Registrierung verwendet werden, liegen die Resultate bei Verdeckung der oberen Gesichtshälfte durch eine Sonnenbrille bei 38,2% auf der AR Gesichtsdatenbank [Mar02]. Wird jedoch die vorgeschlagene Registrierung verwendet, steigen die Resultate auf 97,3%. Die Optimierungsprozedur, die in den Klassifikationsschritt integriert ist, macht den Algorithmus also unempfindlich gegenüber falsch lokalisierten Gesichtsmerkmalen. Bis zu einer Distanzabweichung von 18% zwischen den Augäpfeln liefert der Algorithmus stabile Leistungen.

Der Algorithmus wurde in mehreren realen Systemen eingesetzt und arbeitet zuverlässig unter realen Bedingungen. Die entwickelten Systeme beinhalten ein Türüberwachungssystem, bei dem Personen, die einen Seminarraum betreten, unauffällig mit einer Kamera, die gegenüber der Eingangstür angebracht ist, identifiziert werden; ein Besucher-Interface, bei dem ein Besucher auf einem Monitor eine personalisierte Nachricht angezeigt bekommt, bevor er an die Tür klopft; ein System bei dem ein Roboter die Person identifiziert, mit der er gerade interagiert; und ein System zur Personensuche in Videos anhand von Gesichtsbildern. Zusätzlich wurde der Algorithmus in einem Gesichtserkennungssystem benutzt, das in den CLEAR Evaluationen [SBB+07] evaluiert wurde. Hierbei besteht der Datenkorpus aus realen Daten, die während vorlesungsartiger Seminare oder kleinerer Arbeitsgruppenseminare in intelligenten Räumen aufgezeichnet wurden. Das vorgeschlagene System erzielte die beste Leistung bei allen Kombinationen von Trainings- und Testdaten in den CLEAR 2007 Evaluationen [SBB+07].

Das entwickelte Gesichtserkennungssystem wurde mit einem Sprecheridentifikationssystem kombiniert [EJFS07], um multimodale Personenidentifikation durchzuführen. Ein adaptiver Ansatz zur Gewichtung der Modalitäten wurde eingeführt, der erfolgreich die beiden Modalitäten kombiniert. Im Rahmen der CLEAR 2007 Evaluationen stellte sich heraus, dass der vorgesehene Gewichtungsansatz auch bei variierenden Erkennungsleistungen in und zwischen den beiden Modalitäten signifikante Verbesserungen durch die Fusion erzielte. Zum Beispiel erreichte das Gesichtserkennungssystem für eine bestimmte Kombination auf den Testdaten eine korrekte Erkennungsrate von 94,6%, während das Sprecheridentifikationssystem eine Rate von 96,4% erreichte. Die Fusion der Modalitäten erhöhte die Erkennungsrate auf 99,1%. Bei einigen Training-Test-Kombinationen erzielte die Sprecheridentifikation eine wesentlich schlechtere Leistung als die Gesichtserkennung, z.B. 41,9% gegenüber 84,9% oder 69,6% gegenüber 90,8%. Das multimodale System erzielte mit 86,3% bzw. 93,5% trotzdem eine im Vergleich zu den Einzelmodalitäten verbesserte Leistung.

# Contents

# List of Figures

# List of Tables

# List of Abbreviations

| | |
|---|---|
| AFRD | Advanced face recognition descriptor |
| ARG | Attributed relational graph |
| DCT | Discrete cosine transform |
| EHMM | Embedded hidden Markov model |
| FT | Fourier transform |
| GMM | Gaussian mixture model |
| HMM | Hidden Markov model |
| HSV | Hue-saturation-value color space |
| LAFR | Local appearance-based face recognition |
| LDA | Linear discriminant analysis |
| KLT | Karhunen-Loève transform |
| MAP | Maximum a posteriori |
| PCA | Principal component analysis |
| RBF | Radial basis function |
| SOM | Self-organizing map |
| SIFT | Scale-invariant feature transform |
| SVM | Support vector machine |
| WHT | Walsh-Hadamard transform |
| WT | Wavelet transform |

# 1 Introduction

Face recognition is one of the most important problems of computer vision and pattern recognition. The main group of applications that has fueled intense efforts on face recognition research is security applications, ranging from authentication tasks, such as for access control that can be used in electronic transactions, desktop login, and Internet access, to surveillance tasks, such as bank/store and public area security.

In addition to security applications, person identification is one of the most crucial building blocks for smart interaction applications. Here, smart interaction refers to using perceptual technologies for improved human-human and human-machine interactions. Either as an assistant in human-human interactions, e.g. a memory aid that tells the person who he is talking to, or in human-machine interactions, e.g. a machine that recognizes its user and customizes the preferences accordingly, person identification provides one of the most important characteristic of natural interactions: *personalization*. Besides, the identity of a person can be used to improve the performances of other perceptual technologies, such as expression analysis systems or head pose estimation systems, by enabling the use of person-specific models.

Face recognition and speaker identification are known to be the most natural person identification methods, since face and voice are the modalities that we use to identify people in our daily lives. Although other methods, such as fingerprint identification, can provide better performance, they are not appropriate for natural interactions due to their intrusive nature. The most important advantage of face recognition is the passive identification that it can provide, that is, the person to be identified does not need to cooperate or take any specific action. For example, a smart store can recognize its regular customers while they are entering the store. The customers do not need to talk or look directly into the camera to be recognized. This makes face recognition technology a perfect match for natural interaction applications, since it can work unobtrusively in the background without disturbing or interrupting the subjects to be identified.

Face recognition has found a wide range of smart interaction applications. The application areas, which have been focused on in this thesis, can be classified into three groups. Face recognition for smart environments constitutes the first group. This application group corresponds to identification tasks at a fixed

location, for instance, a smart home that identifies the family members. The second group is face recognition for smart machines. In this application group, a machine identifies the subject that it interacts with, for example, a car that identifies its driver or a robot that recognizes the person it serves. The last application group is face recognition for smart image/video retrieval. In this group, face images are used as cues for identity retrieval.

## 1.1 Motivation

Face recognition is a very challenging computer vision and pattern recognition problem. Facial appearance can undergo severe variations due to changes in facial expression, illumination, occlusion, head pose, and aging. In fact, facial appearance variations caused by these factors often dominate the one caused by identity differences. Another important factor that causes difficulty in face recognition is face alignment. In order to have a proper comparison between face images, they need to be aligned precisely. This requires, in turn, precise facial feature localization, which is a challenging task.

A wide range of potential applications have motivated extensive research efforts on face recognition. Many algorithms have been developed that aim at handling a single factor or combination of two factors that cause facial appearance variations. Especially handling illumination changes has been one of the main point of interest [AMU97, CWX+06, GBK01, GMB04, LHK05, SRR01, ZACJ07]. All these algorithms are only evaluated against the facial appearance variations that they are developed for. For example, the algorithms that are developed to handle illumination variations are tested on data sets that contain only illumination variations. As a result, there exist many face databases that have been collected mainly under controlled settings and that contain facial appearance variations caused by a single factor or a combination of two factors. Sample images from some of these databases are given in Figure 1.1(a,b,c). These studies have provided valuable insights about different aspects of performing face recognition and the databases collected to test them are quite beneficial to find out the algorithms' robustness against specific sources of variations. However, they do not provide a cue about how the tested algorithm is going to perform under real-world conditions. That is, it is necessary and important for a face algorithm to be tested on these benchmark face databases, but it is not sufficient to guarantee that it will work reliably under real-world conditions, even if it performs successfully on all of them. The main reasons are:

- In the benchmark face databases, the variations on the facial appearance are produced by controlling only a single source or combination of two sources of variation, however, in real-world, these variations occur by the combinations of multiple sources.

- The benchmark face databases contain discrete variations, e.g. head poses at some specific angles. On the other hand, in real-world, all kinds of pose variations and all kinds of expression, illumination variations at different strength levels can be encountered.
- The benchmark face databases are collected in a cooperative setting. That is, the individual is informed to stay in front of the camera and he/she is aware that his/her image is being recorded. This data collection setup is reasonable for the authentication task, in which cooperation is needed. However, it is incapable to address the application scenarios that require passive, unobtrusive identification, such as face recognition in smart environments.

Sample images collected from some of the real-world applications are shown in Figure 1.1(d,e).

Keeping these facts in mind, this thesis aims at developing a novel, generic face recognition algorithm that performs robustly in spite of the facial appearance variations caused by expression, illumination, occlusion, aging, and uncontrolled recording conditions. The algorithm has been extensively tested under all these conditions on the benchmark face databases, and on each condition it has been found to perform robustly. Another important novel property of the developed algorithm is that it does not necessarily need a facial feature localization step for face alignment, which makes it insensitive against registration errors due to erroneous facial feature localization. It has also been deployed for several applications and was found to work reliably under real-world conditions. Furthermore, the algorithm has been shown to perform successfully on 3D face data.

## 1.2 Approach

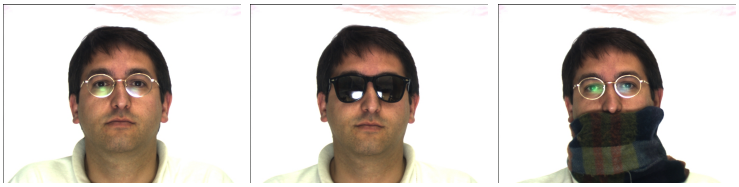
The main steps of the proposed face recognition algorithm are explained briefly in the following subsections.

### 1.2.1 Discrete Cosine Transform-based Local Facial Appearance Representation

The discrete cosine transform (DCT) is used to represent local regions. There are several advantages of using the DCT. Its data independent bases make it very practical to use. There is no need to prepare a representative set of training data to compute a subspace. In addition, it provides frequency information, which is very useful for handling changes in facial appearance. For instance, it

(a)

(b)



(c)

(d)

(e)

Figure 1.1: Sample images used to test the proposed face recognition algorithm. (a) The FRGC database. (b) The CMU PIE face database. (c) The AR face database. (d) CLEAR person identification evaluation corpus. (e) Door monitoring system evaluation corpus.

is known that some frequency bands are good for combating against illumination variations. Moreover, it has been found that the DCT-based local appearance representation is better than representations based on the Karhunen-Loève, Fourier, wavelet, and Walsh-Hadamard transforms in terms of face recognition performance.

### 1.2.2 Local Appearance-based Face Recognition

In the proposed local appearance-based face recognition (LAFR) approach, a detected and registered face image is divided into blocks of $8 \times 8$ pixels size. Afterwards, the DCT is performed on each $8 \times 8$ pixels block. The obtained DCT coefficients are ordered using zig-zag scanning. From the ordered coefficients, according to a feature selection strategy, $M$ of them are selected and normalized, resulting in an $M$-dimensional local feature vector. Finally, the DCT coefficients extracted from each block are concatenated to construct the overall feature vector. The classification is done using a nearest neighbor classifier with L1 norm as the distance metric.

### 1.2.3 Feature Selection and Feature Normalization

There are two main points in the algorithm. The first one is to decide automatically which frequency band to use for classification. In the algorithm, the classification is done using multiple frequency bands, that is, by selecting different DCT coefficients with a sliding window of size $M$ and performing classification with each frequency band. The band that provides the maximum separation between the closest two candidates is chosen as the most reliable band and its decision is used as the classification result. This way, by using the appropriate frequency band, the algorithm can adapt itself automatically to changing illumination conditions.

The second point is about feature normalization. In order to balance the coefficients' and blocks' contributions to the classification, the coefficients in each feature vector are first divided by their standard deviations and then the local feature vector is normalized to unit norm.

### 1.2.4 Face Registration by Minimizing the Closest Classification Distance

In the proposed approach, face registration is done by minimizing the closest distance at the classification step. Since all humans have the same facial feature configuration, once the face is located, positions of the facial features can be

roughly estimated. Search for the precise facial feature positions is conducted around the estimated positions. Various candidate facial feature positions are used to provide several aligned test face images, while comparing a test face image with an already aligned training face image. The facial feature positions, which lead to the aligned test face image that has the minimum distance to the training image, are selected as the facial feature locations. Thus for each training sample, separate eye center positions are determined for the test face image. In this way, inconsistencies across manual eye center labels of the training images are also handled.

## 1.3 Contributions

This thesis focuses on developing a generic, robust face recognition algorithm that can be used reliably for real-world applications. Towards this goal, the following contributions have been made:

- A face recognition algorithm that can handle facial appearance variations caused by facial occlusion, expression, illumination, time gap, and uncontrolled capture conditions, has been developed. The algorithm has been extensively evaluated on the standard benchmarks under different conditions and it is found to be significantly superior to both well-known generic face recognition algorithms and to specific face recognition algorithms that have been designed for each of the factors that cause facial appearance variations. The algorithm achieved 98.5% and 96.2% correct recognition rates on the images from the Face Recognition Grand Challenge face database (FRGC) [PFS+05] captured under the controlled —in a studio setting with controlled illumination— and uncontrolled —under changing conditions in hallways, atria or outdoors— conditions, respectively. The performance on the AR face database [MB98] against upper and lower facial occlusion is 97.3% and 98.2%, respectively. The obtained results under illumination variations are 100% on the CMU PIE face database [SBB03], on all illumination subgroups of the Yale face database B [GBK01] and on the second and third illumination subgroups of the extended Yale face database B [LHK05], whereas they are 98.7% and 99.0% on the fourth and fifth illumination subgroups of the extended Yale face database B [LHK05]. It is the first time in the literature that all facial appearance variations are handled with one generic algorithm, i.e. without devising individual algorithms for each variation.
- Different from traditional face recognition systems, the proposed algorithm does not need an additional facial feature localization step for face registration. It implicitly performs feature localization at the classification step. Moreover, it has been shown that the proposed registration

approach performs even better than doing registration with manual labels. For instance, on the AR face database [MB98], against lower facial occlusion, the obtained result when the test images are aligned using the manual labels was 91.8%, while with the proposed registration approach it has become 97.3%. It has also been shown that the main problem with the upper face occlusion is due to registration errors and not the occlusion itself. Due to the sunglasses, the eye center points that are widely used for face alignment cannot be reliably labelled even manually. When only the manual labels are used to align the test images, the achieved correct recognition rate against upper facial occlusion with sunglasses is 38.2% on the AR face database [MB98]. However, when the proposed registration approach is applied, the performance jumps to 97.3%. The optimization procedure integrated to the classification step makes it also insensitive to registration errors caused by mislocalized feature points. The algorithm can tolerate upto 18% of the interocular distance as localization error and up to this point it provides stable performance.

- The algorithm has been deployed for various real-world systems and is able to work reliably under real-world conditions. The developed systems are: a door monitoring system, where individuals entering a seminar room are identified unobtrusively with a camera located opposite to the entrance door; a visitor interface, where a visitor looks at a monitor to read the displayed message before knocking on the door and receives a personalized information according to his/her identity; a system where a robot identifies the person who interacts with it; and a cast retrieval system where the main characters are retrieved using their face images. In addition, the algorithm has been used in a face recognition system which has been evaluated in the CLEAR evaluations [SBB+07], where the data corpus consists of real-world data collected in smart rooms from lecture-like seminars and interactive small working group seminars. The proposed system was the best performing system on all training-testing combinations in the CLEAR 2007 evaluations [SBB+07].

- The developed face recognition system has been combined with a speaker identification system [EJFS07] to perform multimodal person identification. An adaptive modality weighting scheme that can successfully combine audio and visual modalities has been introduced. It has been shown that the proposed weighting scheme is robust even if the validation data is misleading in terms of recognition performance of individual modalities and even if the performances of the systems are not balanced. For example, the validation set provided in the CLEAR 2007 evaluations, was easier than the data set provided for testing. The face recognition algorithm achieved 100% correct classification rate in most cases and always outperformed the speaker identification on the validation set. However, on the testing set it turned out that at some training-testing combinations,

speaker identification is more successful. Nevertheless, this does not affect the performance of the proposed adaptive modality weighting scheme. For example, at one combination on the test set, face recognition achieved 94.6% correct recognition rate and speaker identification reached 96.4%. The combination of the modalities provided 98.7% correct classification. At some training-testing combinations speaker identification performed significantly worse than face recognition, for example, 84.6% versus 41.9% or 90.8% versus 69.6%. However, the multimodal system still reached an improved performance, with 86.3% for the former and 93.5% for the latter case.

## 1.4 Outline

The organization of this thesis is as follows:

In Chapter 2, an overview of the related work is given. Well-known generic face recognition algorithms, specialized face recognition algorithms developed for combating illumination variations and occlusion, component-based face recognition algorithms, and discrete cosine transform based face recognition algorithms are briefly presented.

In Chapter 3, the proposed local appearance-based face recognition algorithm is introduced. First, the advantages of using a local approach are explained. Then, possible local appearance representation methods are presented. Finally, the details of the proposed algorithm are given.

In Chapter 4, parameters of the local appearance-based face recognition algorithm are analyzed. At the beginning of the chapter, information about benchmark face databases and experimental setups is given. The effects of the following parameters are investigated: feature normalization, distance metric, block size, frequency band, image partitioning strategy, and local appearance representation method. In addition, in this chapter, the algorithm's robustness against compression is assessed.

In Chapter 5, the proposed face registration and automatic frequency band selection approaches are described. The chapter starts with showing appearance-based face recognition algorithms' sensitivity to misalignment. Then, the proposed registration technique is explained in detail and the obtained experimental results are discussed. Afterwards, the automatic frequency band selection method is introduced. The chapter ends with a comparison of the proposed face recognition algorithm with well-known face recognition approaches.

In Chapter 6, real-world face recognition systems are presented. In this chapter, detailed information about person identification in smart rooms, the door

monitoring system, the visitor interface, face recognition for humanoid robots and person identification in movies is provided. In addition, the extension of the proposed face recognition algorithm for 3D face recognition is described.

In Chapter 7, the outcomes of the thesis are discussed and conclusions are given.

# 2 Literature Review

This chapter gives an overview of the related work that has been conducted on face recognition. The chapter consists of five sections. In Section 2.1, generic face recognition algorithms are reviewed. Face recognition algorithms developed to handle illumination variations and facial occlusion are presented in Sections 2.2 and 2.3, respectively. In Section 2.4, algorithms that use local components for identification are explained. Finally, discrete cosine transform based face recognition algorithms are described in Section 2.5.

## 2.1 Generic Face Recognition Algorithms

There have been many generic face recognition algorithms proposed. However, three of them have had a large impact on the face recognition research community and they have inspired countless studies. These are eigenfaces [TP91], Fisherfaces [BHK97], and Bayesian face recognition [MJP00]. In this section, these approaches are described shortly.

The *eigenfaces* approach [TP91] is the most well-known face recognition algorithm. In the algorithm, first a face subspace is constructed from training face images using the principal component analysis (PCA). The face images are then represented in this subspace, which corresponds to the eigenvectors, also called eigenfaces, of the covariance matrix of the face images. Only a subset of the eigenvectors are used to represent the face images, in order to achieve dimensionality reduction. For an $M$-dimensional subspace, the selected eigenvectors are the ones that correspond to first $M$ eigenvalues, sorted in descending order according to their magnitudes. This way the face images can be reconstructed with the smallest mean-square error for any given subspace dimensionality. The classification is done by comparing the face images in this subspace.

As an extension to the eigenfaces approach, to handle head pose variations, the view-based eigenfaces approach is introduced in [PMS94]. In this method, for each different view of an individual an eigenspace is constructed. When a test image arrives, at first the eigenspace that can best represent the view of the face image is determined and then classification is done in that eigenspace. This method performs better than the universal eigenfaces approach in which

only one face space is constructed disregarding the different views of the individuals.

Similar to the idea in [PMS94], to represent the variations in the face space more efficiently, the mixtures of eigenfaces approach is proposed in [FCH98, KKB02, TC02]. While the ordinary eigenfaces algorithm represents the face images with only one subspace, these mixture methods use more than one subspace to represent them. The motivation behind these methods is the belief that the face space can possess clusters corresponding to the variations. Therefore, one hopes that representing each cluster by a local subspace is a more reasonable approach than representing the whole space with a single linear subspace.

Nonlinear extensions of the eigenfaces algorithm via kernel methods have been also studied [YAK00, KJK02, Yan02]. In these approaches, the input face space is first mapped into a higher dimensional space by using nonlinear functions such as a polynomial kernel. The PCA is performed in this higher dimensional new space.

*Fisherfaces* is another very well-known face recognition approach [BHK97]. It uses linear discriminant analysis (LDA) for subspace projection. The aim of the LDA is to extract the projection directions that are effective for discrimination. To achieve this goal, LDA utilizes class information and tries to find the best subspace where the ratio of between-class scatter to within-class scatter is maximized. Recall that PCA tries to maximize the total scatter across all classes. The Fisherfaces approach also has extensions. For example in [KKB03], a LDA mixture model is used, where for each cluster in the face space, a separate LDA is performed and in [Yan02] a nonlinear extension of it, the kernel Fisherfaces algorithm is proposed.

*Bayesian face recognition* approach [MJP00] is famous for its formulation of the multi-class face recognition problem as a two-class classification problem. In this method, intrapersonal and extrapersonal differences are used to exploit the knowledge of critical variations for discriminating the individuals.

In addition to these approaches, there are several other popular face recognition algorithms, such as face recognition by elastic bunch graph matching [WFKM97], Laplacianfaces [HYH$^+$05], local binary patterns [AHP06], etc. A survey on face recognition algorithms can be found in [ZCPR03].

## 2.2 Illumination

One of the most addressed problems in face recognition is illumination variations. It has attracted significant attention during the last decade and there have been many solutions proposed for this problem [AMU97, SK01, SRR01,

WLWZ03, CWX+06, GBK01, VT02, GMB02, LHK05, ZACJ07, LCLZ07]. These solutions can be classified as: invariant features, canonical forms, and variation modeling [SK01]. In the first approach, features insensitive to illumination variations are searched for [AMU97]. The second approach tries to remove the illumination variation either by an image transformation or by synthesizing a new image [SK01, SRR01, WLWZ03, CWX+06]. Finally in the third approach, illumination variation is learned and modeled in a suitable subspace [GBK01, VT02, GMB02, LHK05, ZACJ07]. Besides these solutions, in [LCLZ07] near-infrared lighting is proposed to achieve illumination invariant capture conditions.

### 2.2.1 Invariant Features

In [AMU97] different face representation approaches, such as edge maps, Gabor-like functions, derivatives of the gray-level, and log transformations, are evaluated under lighting direction changes. In total, 107 different operators are tested, but none of them provided insensitiveness against illumination. In this study, it is concluded that the variance in appearance of one person under different lighting conditions, the inner class variance, can be greater than the variance in appearance of different persons under the same lighting condition, the inter class variance.

### 2.2.2 Canonical Forms

The face shape is extracted from a single image using a statistical shape-from-shading model in [SK01]. After extracting the face shape, new face image samples are synthesized under different illumination conditions. In [SRR01], the quotient image method is introduced. In this method, first an illumination invariant signature image is obtained. Afterwards, using this image, face images with varying illumination are generated. A face normalization algorithm that transforms the lighting of one face image to that of another face image is presented in [WLWZ03]. The logarithmic total variation model is proposed in [CWX+06]. Illumination invariant facial structure is obtained from a single face image with this model.

### 2.2.3 Variation Modeling

The illumination cones method is introduced in [GBK01], in which facial appearance variations caused by different illumination conditions are modeled using a small set of face images taken under different lighting directions. The tensorfaces approach is presented in [VT02]. In contrast to eigenfaces [TP91], in which only

the space of face images is spanned, in this method, multilinear analysis is used to decompose the facial image data tensor into five different matrices that span, in addition to face images, the space of people parameters, viewpoint parameters, illumination parameters, and expression parameters. Light-fields theory is utilized and used in a linear discriminant analysis framework in [GMB02]. In [LHK05], it is shown that the subspace that can model the appearance variations caused by changing lighting conditions, can be established by directly using the face images that are captured under pre-arranged lighting conditions.

## 2.3 Occlusion

Partial face occlusion is one of the most challenging problems in face recognition. In this section, related work about this topic will be briefly presented.

In [Mar02], face images are analyzed locally in order to handle partial face occlusion. The face image is first divided into $k$ local regions and for each region an eigenspace is constructed. If a region is occluded, it is automatically detected. Moreover, weighting of the local regions is also proposed in order to provide robustness against expression variations. A similar approach is presented in [TCZZ05], where a self-organizing map (SOM) is used to model the subspace instead of Gaussians or mixtures of Gaussians as in [Mar02].

A face is represented by the face attributed relational graph (ARG) structure in [PLL05]. This representation contains a set of nodes and binary relations between these nodes. In testing, first the correspondences between the ARG representations of the training and testing samples are established. According to the distance between these representations, the classification is performed.

In [FSL06], robustness against occlusion is provided by combining the subspace methods that aim at best reconstruction, such as principal component analysis, with the subspace methods that aim at discrimination, such as linear discriminant analysis.

A sparse signal representation is used to analyze partially occluded face images in [WGYM07]. Another representation based approach is proposed in [JM08]. Different from the studies [FSL06, Mar02, PLL05, TCZZ05, WGYM07] in which occluded images are only included in the testing set, in [JM08] they are included both in the training and testing sets. The occlusion problem is handled as a reconstruction problem and the classification is done according to the obtained reconstruction error on a test image.

A common point of the studies listed in this section is the use of the AR face database [MB98] for the face recognition experiments.

## 2.4 Component-based Approaches

Face recognition based on local facial regions has attracted a significant amount of interest. Approaches that utilize local regions either use salient regions or they just partition the face image into rectangular blocks. The approaches that exploit salient regions can also be further divided into two subgroups as the ones using predefined regions, such as eyes, and the ones using automatically learned regions. In this section, first brief information about the salient region-based studies is given, then the methods that perform generic partitioning are overviewed.

In [BP93], whole face template as well as eyes, nose, and mouth are used for face recognition (See Figure 2.1.). The relative location of these regions with respect to the eye position is the same for each face image. Template matching is employed for identification. Normalized cross correlation is used as the matching score. The resulting scores from each representation, that is from the whole face and local regions, are accumulated and the test face image is assigned with the identity of the training sample that attains the highest matching score. According to the classification performance of the individual representations, the eye region is found to be the most discriminative region. The second best result is obtained by the nose region. The mouth region takes the third place. Interestingly, the entire face image is found to have the least discrimination power.

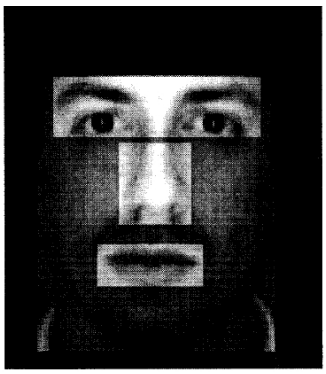


Figure 2.1: Local regions used for face recognition in [BP93] (From [BP93]).

Automatically detected facial features are utilized for identification in [PMS94]. The eigenfeature-based approach is used for facial feature detection. In this method, first an eigenspace, which is named eigenfeature, is built for each facial feature. Afterwards, at each pixel, the distance-from-feature-space is computed, which indicates how well the eigenfeature can represent the region under examination. Low distance values correspond to good representation capability, whereas high distance values correspond to poor representation capability. This way a distance map is obtained and the global minimum of this distance map

is selected as the facial feature location. The detectors are developed for the left eye, right eye, nose, and mouth regions. Sample facial feature templates are shown in Figure 2.2. However, the mouth region is not used for identification, since it is sensitive to expression variations. The performances of three different representation schemes are assessed. These are whole face, combined facial features and combined whole face and facial features representation. Feature extraction is done by projecting the whole face and the facial features onto the corresponding eigenspaces. It is observed that in the lower feature dimensions, combined facial feature representation outperforms the whole face representation. The fusion of these representations improves the performance slightly.

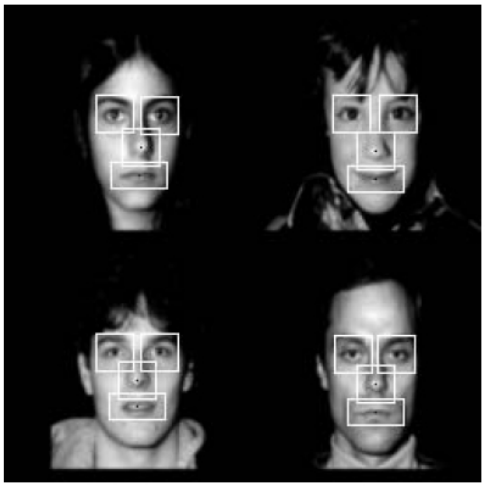


Figure 2.2: Local regions used for face recognition in [PMS94] (From [PMS94]).

A normalized face image is divided into four predefined regions in [LKK05]. These regions are the left eye region, the right eye region, the nose region, and inner face region excluding the mouth (see Figure 2.3). Similar to the approach in [PMS94], the mouth region is excluded due to its sensitivity to expression variations. So called DCT/LDA based features are extracted from these regions, as well as from the whole face and both are then used for classification.

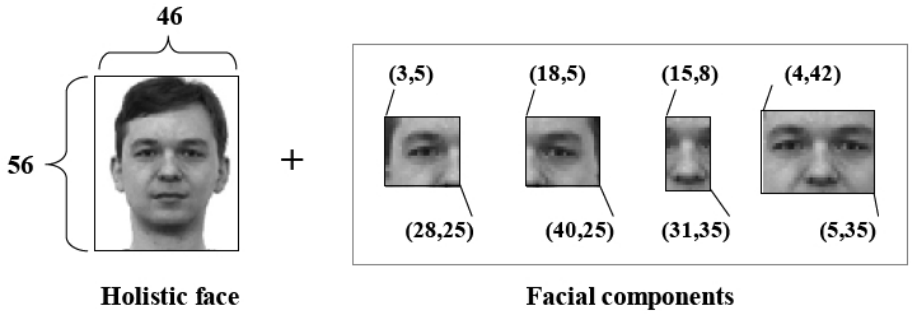
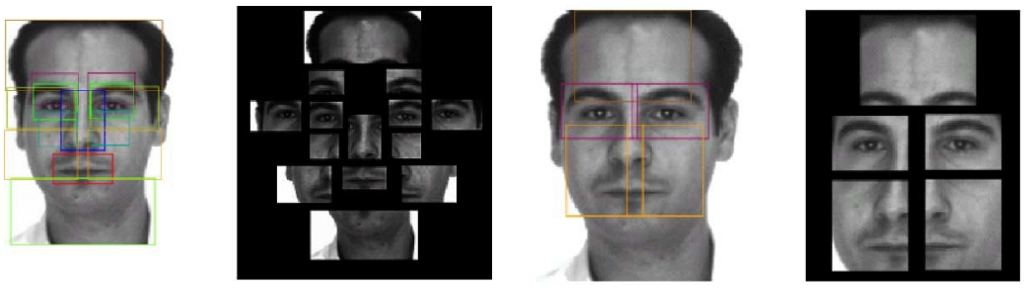



Figure 2.3: Local regions used for face recognition in [LKK05] (From [LKK05]).

In [KKHK05], two different partitioning schemes are considered. Similar to [BP93], the relative locations of these regions with respect to the eye position are the same for each face image. The first scheme has 14 overlapping components consisting of small and large regions (See Figure 2.4(a).). Small regions are located around the salient points such as eyes, nose or mouth. The large regions are located around the forehead, cheeks, and neck. The second scheme has five overlapping components (See Figure 2.4(b).). These components are obtained by combining two or three components of the first partitioning scheme into one. This is done by exhaustively searching for the best performing combinations on the training set. The resulting components contain the forehead including the eyebrows, the left eye region, the right eye region, the lower left face region, and the lower right face region both including most of the nose. In this method, LDA is applied both on the whole face and on the components. The extracted feature vectors are combined and another LDA is applied on the resulting combined feature vector.

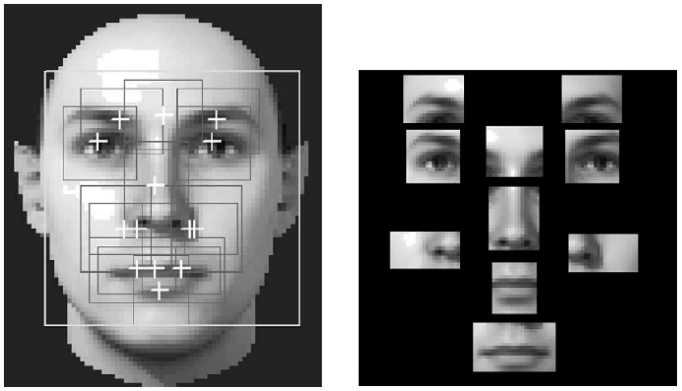


(a) (b)

Figure 2.4: Local regions used for face recognition in [KKHK05] (From [KKHK05]).

Facial features are detected by a two level, component-based face detector in [HHWP03]. The first level corresponds to detection of facial features, whereas the second level checks for the geometrical configuration to verify whether there exists a face or not in the scene. Facial feature detection is based on support vector machine (SVM) classifiers. In the study, 14 facial components, which are shown in Figure 2.5(a), are detected. From these 14 components, the components around the cheeks and the highly overlapped ones are discarded and 10 of them are kept for face recognition. These components can be seen in Figure 2.5(b). Face recognition is also done with SVM classifiers. The one-vs-all strategy is employed, where for each subject an SVM classifier is trained discriminating the person from the other subjects.

14 automatically learned components are used for face detection and identification in [HSP07]. These components are depicted in Figure 2.6. Similar to [HHWP03], SVM classifiers are used to detect facial features and to classify faces.



(a) (b)

Figure 2.5: Local regions used for (a) face detection, (b) face recognition in [HHWP03] (From [HHWP03]).

(a) (b)

Figure 2.6: Local regions used for (a) face detection, (b) face recognition in [HSP07] (From [HSP07]).

Generic partitioning has also been employed in several studies. In [AHP06, CEW05, GA04, PB99, PZ96, Sco03, TLV04, ZR04], the face is divided into non-overlapping blocks, whereas in [EMR00, KD98, LC04, Nef99, SP03a] overlapping blocks are used. In [AHP06], local binary pattern histograms are extracted from $18 \times 21$ pixels resolution blocks. The DCT is applied on image blocks in [CEW05, PB99, PZ96, Sco03, TLV04, ZR04]. $8 \times 8$ pixels blocks are used in [CEW05, TLV04]. Different block sizes are tried in [PB99, PZ96, Sco03, ZR04]. PCA is utilized in [GA04], where the input face image is divided into 4 to 4096 non-overlapping blocks. In all the works that use overlapping image blocks, the DCT is used for feature extraction and the extracted features are fed into parametric classifiers. In [EMR00, KD98, Nef99], the extracted features are given as input to a HMM classifier, whereas in [LC04, SP03a] they are given as input to a Gaussian mixture model (GMM) classifier.

## 2.5 DCT-based Face Recognition Algorithms

The discrete cosine transform has been used as a feature extraction step in several studies on face recognition. In [PZ96], block-based DCT is used to extract the features. A subset of the obtained DCT coefficients are then vector quantized and the classification is done using the nearest neighbor classifier. The experiments are conducted on a database of 25 people. Two different block sizes, $16 \times 16$ and $32 \times 32$ pixels resolution, and two different local feature dimensions, 8 and 16, have been tested with respect to different codebook sizes, 16, 32, 64, and 128. It has been observed that larger block size and codebook size result in better performance. However, no significant performance improvement is obtained when increasing the feature dimension from 8 to 16 in the case of $32 \times 32$ pixels block size. The performance of the DCT-based feature vectors is also compared with the one of raw pixel intensity values. It has been found that 16-dimensional DCT-based feature vectors perform as well as 256-dimensional pixel intensity values-based feature vectors, which implies that using the DCT, feature dimensionality can be reduced greatly without losing performance.

The DCT-based features are classified with a one-dimensional hidden Markov model (HMM) in [KD98]. DCT is applied on overlapping image blocks. The obtained DCT coefficients are ordered using the zig-zag scan pattern and only the first few coefficients are selected as features. Image blocks are processed sequentially from top left part of the face image to bottom right and the extracted observation sequence is fed into the one-dimensional HMM. The algorithm is evaluated on the ORL face database [SH94] with respect to image block size, percentage of overlap between the neighboring blocks, and the number of used DCT coefficients. The best result is attained with an image block size of $16 \times 16$ pixels, having 75% overlap and using ten DCT coefficients.

In [PB99] the DCT coefficients of the entire image or its blocks are computed and from the obtained coefficients, only a subset of them is selected by diagonally scanning the upper-left part. The coefficients are then given as an input to a multi-layer perceptron. The algorithm's performance is extensively tested on the ORL face database [SH94] by varying the parameters of the number of coefficients, number of hidden neurons, and block size. The best performance is attained on the entire image, with 35 coefficients and 75 hidden neurons. The results obtained on $8 \times 8$ and $16 \times 16$ pixels resolution blocks are shown to be slightly inferior than the ones obtained on the entire image. The main reason for this could be attributed to the fact that the location information of the local features are lost while feeding them into a neural network structure.

DCT coefficients of image blocks are used as observation vectors of an embedded hidden Markov model (HMM) in [Nef99]. In this approach, each face region — forehead, eyes, nose, mouth, and chin— are modeled with a *super state*. Each *super state* contains *embedded states*. In the algorithm, the face image is scanned with a window size of $8 \times 10$ pixels resolution. There exists overlap between the neighboring blocks. For each block only six DCT coefficients are used. The proposed method has been evaluated comparatively with the eigenfaces [TP91] and other HMM-based face recognition approaches [Sam94] on the ORL face database [SH94]. The results show that use of embedded HMM scheme with DCT coefficients as the observation vectors provides improved performance.

Another method that uses DCT coefficients as observation vectors of an HMM scheme is proposed in [EMR00]. DCT is performed on overlapping $8 \times 8$ pixels blocks and 15 DCT coefficients are extracted by diagonally scanning the upper-left part of the DCT coefficient block. The proposed method is tested on the ORL face database [SH94]. 100% correct recognition rate is obtained on this database.

In [HL01], the DCT is performed on the entire image and a square subset of the DCT coefficients from the top-left part is used as the feature vector. The nearest neighbor classifier, which uses the L2 norm as the distance metric, is used in the study. The proposed algorithm's performance is assessed with respect to varying parameters such as number of training images per person, number of used DCT coefficients, and geometric normalization on the Achermann database from the University of Bern, the ORL database [SH94], the MIT database, and the CIM database, which was collected at the Center for Intelligent Machines (CIM) in McGill University. The algorithm is also compared with the eigenfaces approach [TP91] and found superior to it.

A derived coefficient set, called mod 2 feature set, from the DCT coefficients are proposed in [SP03a] for face-based identity verification. $8 \times 8$ pixels blocks, having 50% overlap between horizontally and vertically neighboring blocks, are used. Each block is represented with the DCT-mod 2 feature set, which are obtained by replacing the first three DCT coefficients according to the zig-zag scan

pattern with their horizontal and vertical *deltas*. The *deltas* correspond to the difference between the horizontally or vertically neighboring blocks' DCT coefficients and they are interpreted as representing transitional spatial information. In addition to DCT-mod 2 feature set, three feature sets, namely the DCT-delta, DCT-mod, and DCT-mod-delta feature sets are also analyzed. DCT-delta feature set is extracted by replacing the DCT coefficients with their horizontal and vertical deltas. DCT-mod feature set is extracted by removing the first three DCT coefficients from the DCT-based feature vector. DCT-mod-delta is derived again by removing the first three coefficients from the DCT-based feature vector and combining the remaining coefficients with the corresponding DCT-delta feature vector. Gaussian mixture models (GMM) are used for modeling the distribution of extracted feature vectors and verification is done by comparing the average log-likelihood value of the claimant being genuine and the average log-likelihood value of the claimant being an impostor. For the GMM, eight mixtures are used and 15 coefficients are selected for the DCT. The experiments are conducted on the VidTIMIT audio-visual database [SP03b] with artificially varying illumination conditions and on Weizmann database [AMU97]. The DCT-mod 2 feature set is shown to be the most suitable feature extraction method compared to the DCT-delta, DCT-mod, and DCT-mod-delta. The proposed method is also evaluated comparatively with the eigenfaces approach [TP91], ordinary DCT and Gabor wavelets. The experimental results indicate that the DCT-mod 2 feature set is more discriminative and more robust to illumination variations than these approaches.

In [Sco03], a network of networks (NoN) model is fed by the DCT coefficients extracted from image blocks. The DC coefficient, the DCT coefficient at the top-left, is excluded from the feature computation and five different feature vectors are computed from the remaining DCT coefficients. The first method computes the squared sum of the DCT coefficients at each block, whereas the second method sums the absolute values of the DCT coefficients. The third method calculates the mean squared value of the DCT coefficients. Similarly, the fourth method calculates the mean absolute value of the DCT coefficients in a block. Finally, the fifth method first subtracts the average DCT value of the image block from the values of the DCT coefficients and then calculates the mean absolute value of resulting differences. For all of the five methods, the best recognition rates are obtained using $8 \times 8$ pixels block size on the ORL face database [SH94].

The idea of having separate image partitioning schemes and fusing their classification outcomes at the decision level is proposed in [ZR04]. In this method, three different approaches, Bayesian face recognition [MJP00], Fisherfaces [BHK97], and DCT are performed both on the entire face image and on the partitioned face images. Three different image partitioning schemes are employed. The first one divides the face image horizontally into four equal pieces. The second and third ones segment the face image both vertically and horizontally, the former leading

to four equal pieces and the latter leading to twelve equal pieces. The classification outputs of each partitioning scheme and the original image are fused at the decision level via the sum rule [KHDM98]. In the DCT-based method, the top-left square subset of DCT coefficients are used for classification. From the results on the Yale [BHK97] and ORL [SH94] face databases, it is observed that fusing the information coming from different partition schemes improves the correct recognition rates. In the experiments, the DCT-based method outperforms Bayesian face recognition [MJP00] and Fisherfaces [BHK97].

The energy histogram of the DCT coefficients is used for face recognition in [TLV04]. In this approach, the DCT is performed on $8 \times 8$ image blocks and four different feature sets are constructed with the obtained DCT coefficients. The first feature set contains only the top-left DCT coefficient; the second, third, and fourth ones contain square subsets of DCT coefficients from the top-left part having the size of $2 \times 2$, $3 \times 3$, and $4 \times 4$, respectively. Energy histograms of these feature sets are used as feature vector for classification. The nearest neighbor classifier is used. The L2 norm is employed as the distance metric. The experiments are conducted on the Yale face database [BHK97]. The effects of bin size and use of different feature sets are analyzed. The best recognition performance is attained by using the second feature set with a histogram bin size of 30.

In [CEW05] PCA and LDA are performed in the DCT domain. First, DCT is applied on the $8 \times 8$ pixels resolution blocks. The resulting DCT coefficients are quantized and then ordered according to the zig-zag scan pattern. Only a number of DCT coefficients containing high magnitudes is kept. The obtained DCT coefficients from each block are concatenated. PCA and LDA are applied on this combined vector. The FERET database [PMRR00] is used for the experiments. It is shown that PCA and LDA can be applied in the DCT domain without losing performance, while having reduced storage requirements and computational load.

A component-based DCT/LDA approach is proposed in [LKK05]. In this approach, the DCT and LDA are applied successively both on the entire face image and on four predefined facial components. These components are, left and right eye regions, nose region, and inner face region excluding the mouth. The DCT is applied both on the intensity and edge images. The Sobel operator is used to extract the edge image. From the resulting DCT coefficients, the ones that have a high ratio of between class variance to within class variance are kept and fed into the LDA. The classification is done by finding the minimum weighted Euclidean distance between the feature vector of the test image and the feature vectors of the training images. The magnitudes of feature vectors extracted from each component are normalized before calculating the distance. Separate weights are assigned to the components. To speed up the classification process, a representative feature scheme is introduced, which is defined as the median

features of each class. During classification, the feature vector of the test face image is first compared with the representative features. Then, it is compared with all the training samples of the top candidate identities found in the first step. The proposed method is evaluated on the MPEG-7 data set and a data set from the Korean Broadcasting System and its performance is compared with the one of MPEG-7 advanced face recognition descriptor (AFRD). Component-based DCT/LDA approach is found to be more successful than the MPEG-7 AFRD.

In [ECW05], the DCT is applied on the entire face image. Low frequency DCT coefficients are discarded from the feature representation to provide robustness against illumination variations. Feature representations from each class are then clustered in order to model the nonlinear face manifold with multiple linear manifolds, hence improving the performance of the LDA, which takes the clustered feature representations as input. Finally, the output of the LDA is fed into radial basis function (RBF) neural networks. Several experiments have been conducted on the ORL [SH94], FERET [PMRR00], and Yale [BHK97] databases. The proposed method is found to achieve high correct recognition rates on each of these databases.

The DCT is applied on logarithm images in order to normalize illumination variations in [CEW06]. The transform is performed on the entire image. Values of the coefficients are normalized according to the DC coefficient's value, that is, the coefficients are scaled so that each face image have the same DC coefficient value. Low frequency DCT coefficients, which are sensitive to the illumination variations, are removed from the DCT representation. The face image is reconstructed by using the remaining DCT coefficients. Logarithm images are used for classification, no inverse logarithm is taken. Correlation and the eigenfaces [TP91] method are applied on the normalized face images. The nearest neighbor classifier with the L2 distance metric is employed. The experimental results on the CMU PIE [SBB03] and Yale face database B [GBK01] show that the proposed approach is robust against illumination variations.

# 3 Local Appearance-based Face Recognition

The local appearance-based face recognition algorithm is a generic, practical, and robust face recognition algorithm that utilizes representations of local facial regions and combines them at the feature level which provides conservation of the spatial relationships. The underlying ideas for preferring a local appearance-based approach over a holistic appearance-based approach are as follows:

- In a holistic appearance-based face recognition approach, a change in a local region can affect the entire feature representation, whereas in local appearance-based face recognition it affects only the features that are extracted from the corresponding block while the features that are extracted from the other blocks remain unaffected.
- A local appearance-based algorithm can facilitate weighting of local regions. It can put more weight to the regions which are found to be more discriminant. Moreover, this can also improve robustness against occlusion, by giving less weight to the regions where an occlusion is detected.

A diagram indicating the feature extraction via local appearance representation is shown in Figure 3.1. In the approach, a detected and aligned face image is first divided into local regions. For example, in Figure 3.1 it is done without considering any salient regions, such as eyes. Then, a transform can be used to represent local facial regions. Afterwards, the extracted representation coefficients from each block are combined in order to provide the feature vector that represents the entire face image.

In this chapter, first, the representation approaches that can be used to model the local facial regions are overviewed. According to the experimental results, the discrete cosine transform (DCT) has been found to be the optimal representation method. Therefore, in the second section local appearance-based face recognition using the DCT is explained in detail. In the last section a brief information about the nearest neighbor classification method, which has been used as the classifier in the face recognition algorithm, is given.

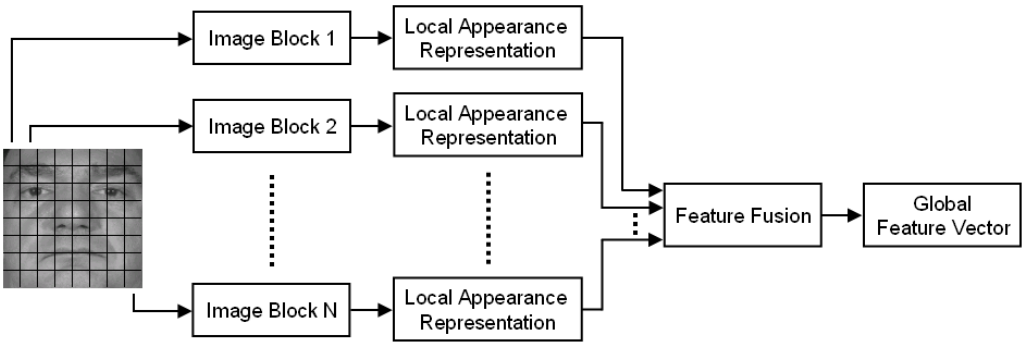



Figure 3.1: Local appearance representation.

## 3.1 Local Appearance Representation Methods

In the following subsections, the image transforms that can be used to represent local facial appearance are briefly explained.

### 3.1.1 Discrete Cosine Transform

The discrete cosine transform (DCT) is a well-known signal analysis tool used in compression standards due to its compact representation power, which is superior to that of the other widely used input independent transforms, e.g. discrete Fourier transform and Walsh-Hadamard transform [GW01]. Although Karhunen-Loève transform (KLT) is known to be the optimal transform in terms of information packing, its data dependent nature makes it infeasible for use in some practical tasks. Furthermore, DCT closely approximates the compact representation ability of the KLT, which makes it a very useful tool for signal representation both in terms of information packing and in terms of computational complexity due to its data independent nature.

The 2-D discrete cosine transform of an $m \times m$ image block is defined as

$$C(u,v) = \alpha(u)\alpha(v) \sum_{x=0}^{m-1} \sum_{y=0}^{m-1} (f(x,y) cos[\frac{(2x+1)u\pi}{2m}] cos[\frac{(2y+1)v\pi}{2m}]) \quad (3.1)$$

for $u, v = 0, 1, , m-1$ where,

$$\alpha(u) = \begin{cases} \sqrt{1/m} & for \quad u = 0 \\ \sqrt{2/m} & for \quad u = 1, 2, ..., m-1 \end{cases} \quad (3.2)$$

and the 2-D inverse discrete cosine transform is defined as

$$f(x,y) = \sum_{u=0}^{N-1}\sum_{v=0}^{N-1}(\alpha(u)\alpha(v)C(u,v)cos[\frac{(2x+1)u\pi}{2N}]cos[\frac{(2y+1)v\pi}{2N}]). \quad (3.3)$$

The DCT basis functions can be seen in Figure 3.2. As can be seen from the top-left part of the basis functions and also from Equation 3.1, the $(0,0)$ component represents the average intensity value of the image, which is directly affected by illumination variations. From the figure, it can also be noticed that the $(0,1)$ and $(1,0)$ components represent the vertical and horizontal intensity changes, respectively.

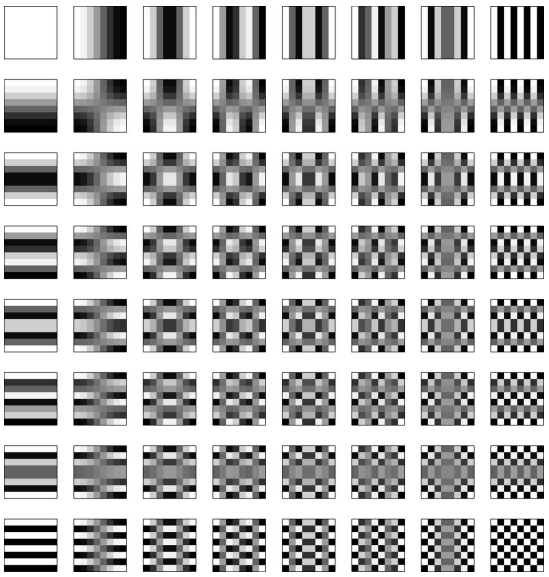
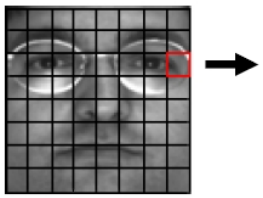


Figure 3.2: DCT basis functions for $m = 8$.

A sample DCT output is depicted in Figure 3.3. As one can observe, the coefficients that account for a greater degree of the representation capability are located at the top-left block of the matrix. To construct the feature vector from the 2D DCT coefficients, the coefficients are ordered using the zig-zag scanning pattern (see Figure 3.4). In this way, the coefficients containing the most information are preserved when the vector is truncated.

### 3.1.2 Principal Component Analysis

Principal component analysis (PCA) is one of the most well-known dimensionality reduction techniques. It is an unsupervised method and it tries to



| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 861 | -158 | 28 | -21 | 8 | -2 | -1 | -1 |
| 87 | -47 | 63 | -68 | 18 | 0 | -5 | 4 |
| 85 | 17 | 73 | -19 | -1 | -2 | -2 | 3 |
| -1 | -27 | 10 | 20 | -10 | 15 | -15 | 6 |
| -7 | 6 | -13 | 12 | -12 | 15 | -6 | 1 |
| -10 | 16 | -15 | 23 | -15 | 12 | -6 | 3 |
| -7 | 10 | -13 | 8 | -7 | 9 | 1 | 3 |
| -5 | 3 | -7 | 6 | -5 | 2 | 2 | 2 |

Figure 3.3: Sample DCT output of a face image block. The processed block is marked with a red bounding box.

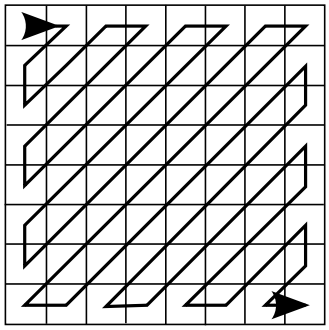


Figure 3.4: Zig-zag scan pattern.

find the best representation subspace that minimizes the reconstruction error. The method has been widely used for holistic face recognition, for example for eigenfaces [TP91]. It can be also used for representing local facial regions [PMS94, GA04].

In principal component analysis, first, the subspace that represents the input data is learned from the available training samples. Let $\boldsymbol{B_i}$ represent an $m \times n$ resolution image block. By concatenating the rows or columns, the two-dimensional image block can be converted to a $N = m \times n$ dimensional vector $\boldsymbol{y_i}$. Let $\boldsymbol{y_{i,1}}, \boldsymbol{y_{i,2}}, \boldsymbol{y_{i,3}}, \ldots, \boldsymbol{y_{i,K}}$ be the set of the $i^{th}$ blocks from $K$ images. The average $i^{th}$ block, $\boldsymbol{y_{i,m}}$, is calculated as

$$\boldsymbol{y_{i,m}} = \frac{1}{K} \sum_{k=1}^{K} \boldsymbol{y_{i,k}}. \tag{3.4}$$

The difference between the training face image block and the average face image block is

$$\bar{\boldsymbol{y}}_{\boldsymbol{i,k}} = \boldsymbol{y_{i,k}} - \boldsymbol{y_{i,m}}. \tag{3.5}$$

The $N \times N$ covariance matrix, $\boldsymbol{C}$, can be calculated as

$$\boldsymbol{C} = \boldsymbol{F}\boldsymbol{F}^T, \tag{3.6}$$

where $\boldsymbol{F}$ is an $N \times K$ matrix, containing the $\bar{\boldsymbol{y}}_{\boldsymbol{i,k}}$s in its columns. If the dimension of the image block is less than the number of training images, $N < K$, then the eigenvectors of $\boldsymbol{C}$ can be calculated directly. If the dimension of the image block is higher than the number of training images, $K < N$, then first $K \times K$ matrix $\boldsymbol{L}$ is computed

$$\boldsymbol{L} = \boldsymbol{F}^T\boldsymbol{F}. \tag{3.7}$$

The eigenvectors $\boldsymbol{v_L}$ and eigenvalues $\lambda_L$ of $\boldsymbol{L}$ can be calculated as

$$\boldsymbol{L}\boldsymbol{v_L} = \boldsymbol{\lambda_L}\boldsymbol{v_L}. \tag{3.8}$$

By substituting $\boldsymbol{L}$ with Equation 3.7

$$\boldsymbol{F}^T\boldsymbol{F}\boldsymbol{v_L} = \boldsymbol{\lambda_L}\boldsymbol{v_L}. \tag{3.9}$$

Multiplying both sides by $\boldsymbol{F}$

$$\boldsymbol{F}\boldsymbol{F}^T\boldsymbol{F}\boldsymbol{v_L} = \boldsymbol{\lambda_L}\boldsymbol{F}\boldsymbol{v_L}. \tag{3.10}$$

Substituting $\boldsymbol{F}\boldsymbol{F}^T$ with $\boldsymbol{C}$

$$\boldsymbol{C}\boldsymbol{F}\boldsymbol{v_L} = \boldsymbol{\lambda_L}\boldsymbol{F}\boldsymbol{v_L}. \tag{3.11}$$

As it can be seen from Equation 3.11, the eigenvectors $\boldsymbol{v_C}$ of the covariance matrix $\boldsymbol{C}$ can be calculated as

$$\boldsymbol{v_C} = \boldsymbol{F}\boldsymbol{v_L}. \tag{3.12}$$

A face image block can be represented in this space without loss of information, by weighted sum of these eigenvectors. However, generally, a small set of eigenvectors is enough to represent the face image blocks properly. The first $M$ eigenvectors, $M < N$, corresponding to the highest first $M$ eigenvalues, are chosen to construct the subspace. The face image blocks are represented with $M$-dimensional feature vectors $\boldsymbol{r_k}$ by projecting them onto the subspace

$$\boldsymbol{r_k} = \boldsymbol{V}^T\bar{\boldsymbol{y}}_{\boldsymbol{k}}, \tag{3.13}$$

where $\boldsymbol{V}$ is the $N \times M$ dimensional matrix that contains the $M$ eigenvectors in its columns.

### 3.1.3 Wavelet Transform

Wavelet transformation is a powerful signal analysis tool, widely used for feature extraction, compression, and denoising. As its name implies, wavelet transform represents the signal with small waves of limited durations, which are called wavelets. It provides examination of the signal both in frequency and time domains.

The two-dimensional wavelet transform is performed by applying the one-dimensional wavelet transform to the rows and columns of the input image block consecutively. Tree representation of one level, two-dimensional wavelet decomposition is shown in Figure 3.5. In this figure, $H$ represents high-pass filtering, $L$ represents low-pass filtering, and $\downarrow 2$ represents downsampling by a factor of 2. The input image block $B_i$ of resolution $m \times m$ is first filtered along the rows and downsampled by 2 producing two $m \times m/2$ images $B_{i,H}$ and $B_{i,L}$ that have high and low frequency contents, respectively. After this decomposition, the wavelet transform is applied to the columns of these $m \times m/2$ resolution

images. In the final stage of the decomposition, there are four $m/2 \times m/2$ resolution subband images: $A_1$, the scaling component containing low-pass global information obtained by low-pass filtering the rows and columns, $H_1$, the horizontal details obtained by low-pass filtering the rows and high-pass filtering the columns, $V_1$, the vertical details obtained by high-pass filtering the rows and low-pass filtering the columns, $D_1$, the diagonal details obtained by high-pass filtering the rows and columns. This process is illustrated in Figure 3.5. The scaling component can be decomposed further to obtain higher order wavelet transform.

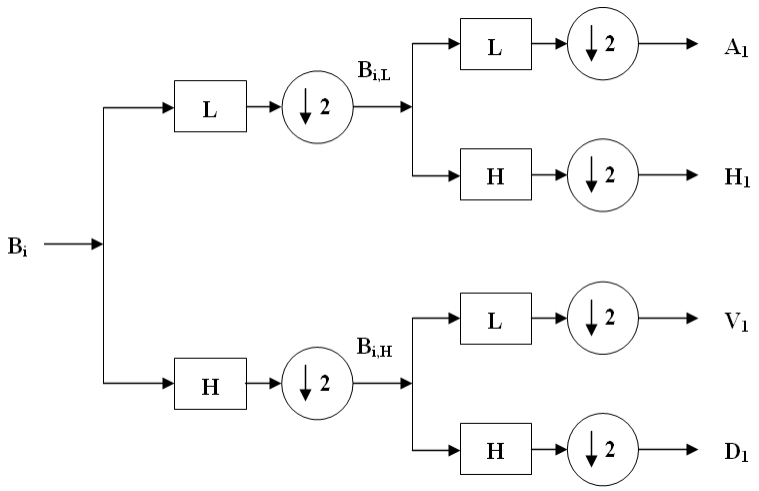



Figure 3.5: Tree representation of one-level 2-D wavelet decomposition.

### 3.1.4 Fourier Transform

Fourier transform is one of the widely used signal analysis tools. It transforms the input image from spatial domain to frequency domain, where the image is represented as weighted sum of sines and cosines at different frequencies. At some signal processing tasks, the features that are not observable in the spatial domain can be easily attained in the frequency domain. Therefore, analyzing the signal in the frequency domain can be beneficial for feature extraction.

The two-dimensional discrete Fourier transform of the input image block $\boldsymbol{B_i}$ of resolution $m \times n$ pixels can be calculated as

$$F(u,v) = \frac{1}{mn} \sum_{x=0}^{m-1} \sum_{y=0}^{n-1} B_i(x,y) e^{-j2\pi(ux/m+vy/n)}, \tag{3.14}$$

for $u = 0, 1, 2, \ldots, m-1$, $v = 0, 1, 2, \ldots, n-1$.

As can be seen from Equation 3.14, $F(0,0)$ corresponds to the average intensity value of the image block, whereas $F(m-1,n-1)$ corresponds to the highest frequency content.

### 3.1.5 Walsh-Hadamard Transform

Another image transform that can be used to represent local regions is Walsh-Hadamard transform (WHT). The Walsh-Hadamard kernel forms a matrix of $\pm 1$s and its rows and columns are orthogonal to each other. 1-D transformation kernel for $m = 8$ can be seen in Equation 3.15.

$$\mathbf{H_8} = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & -1 & 1 & -1 & 1 & -1 & 1 & -1 \\ 1 & 1 & -1 & -1 & 1 & 1 & -1 & -1 \\ 1 & -1 & -1 & 1 & 1 & -1 & -1 & 1 \\ 1 & 1 & 1 & 1 & -1 & -1 & -1 & -1 \\ 1 & -1 & 1 & -1 & -1 & 1 & -1 & 1 \\ 1 & 1 & -1 & -1 & -1 & -1 & 1 & 1 \\ 1 & -1 & -1 & 1 & -1 & 1 & 1 & -1 \end{bmatrix} \tag{3.15}$$

Given an image block $B_i$ of size $m \times m$, the discrete transform can be represented as:

$$T(u,v) = \sum_{x=0}^{m-1} \sum_{y=0}^{m-1} B_i(x,y) g(x,y,u,v) \tag{3.16}$$

The 2-D Walsh-Hadamard kernel is formulated as

$$g(x,y,u,v) = \frac{1}{m} (-1)^{\sum_{i=0}^{n-1} [c_i(x) r_i(u) + c_i(y) r_i(v)]}. \tag{3.17}$$

where

$$r_0(u) = c_{m-1}(u) \tag{3.18}$$
$$r_1(u) = c_{m-1}(u) + c_{m-2}(u) \tag{3.19}$$
$$r_2(u) = c_{m-2}(u) + c_{m-3}(u) \tag{3.20}$$
$$\vdots$$
$$r_{m-1}(u) = r_1(u) + r_0(u) \tag{3.21}$$

The summations are performed in modulo 2 arithmetic and $c_k(z)$ is the $k^{th}$ bit representation of $z$. For instance, if $n = 3$ and $z = 4$ (100 in binary), $c_0(z) = 0$, $c_1(z) = 0$, and $c_2(z) = 1$.

Walsh-Hadamarad basis functions for $m = 8$ can be seen in Figure 3.6

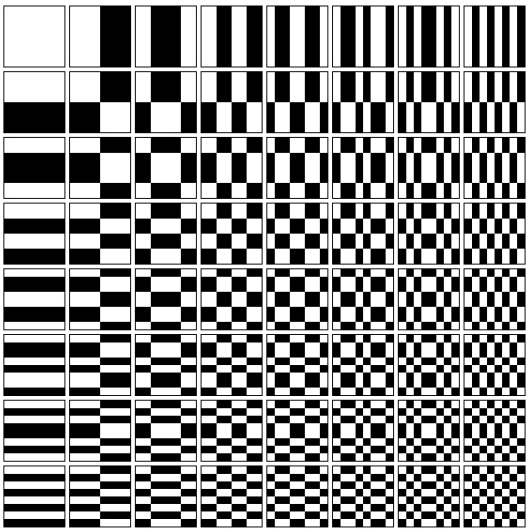


Figure 3.6: Walsh-Hadamard basis functions for $m = 8$.

## 3.2 Local Appearance-based Face Recognition Using Discrete Cosine Transform

In the proposed local appearance-based face recognition (LAFR) approach, a detected and registered face image is divided into local regions. Then, on each local region, the DCT is performed. The obtained DCT coefficients are ordered using zig-zag scanning. From the ordered coefficients, according to a feature selection strategy, $M$ of them are selected and normalized resulting in an $M$-dimensional local feature vector. Finally, the DCT-based feature vectors extracted from each block are concatenated to construct the overall feature vector. The classification is done using a nearest neighbor classifier. As one can notice there are several parameters involved in such an approach. These are:

- Feature normalization: Since the DCT coefficients have a different magnitude range and since local blocks with different brightness levels lead to DCT coefficients with different value levels, it is important to balance the contribution of each coefficient and each block to the classification.

- Block size: Applying DCT on large local regions provides more compactness in representation, however, it provides a poor statistical representation of the region. A better statistical representation can be provided using small local regions, however, this time the representation become less compact.
- Image partitioning: The face image can be partitioned by considering some salient regions, such as eyes, or by putting a rectangular grid without considering any salient region.
- Feature selection: Different types of facial appearance variations can be handled by using different frequency bands. Therefore, an automatic frequency band selection method is required to determine the appropriate frequency band to be used for classifying the test image.

In the next chapters, these points will be explained and analyzed in detail.

## 3.3 Nearest-neighbor Classification

Due to the non-parametric distribution of face data and due to the small sample size available for training, the nearest neighbor classification method has been widely used for face recognition. It is an easy and efficient, so called lazy classification algorithm, where there is no work done in training stage, such as density estimation, and all the work is conducted during testing. The classification is done by comparing a test sample with all the training samples in the database and by finding the training sample that has the closest distance. Several distance metrics can be used for the nearest neighbor classification, such as L1, L2 norms, and normalized correlation. The choice of the distance metric is very important and the classification performance can change dramatically according to the used distance metric.

# 4 Experiments

In this chapter, parameters of the proposed local appearance-based face recognition (LAFR) approach are analyzed through extensive experiments. The chapter starts with describing the benchmark face databases and experimental setups. Following sections convey detailed assessment of the effects of the parameters — feature normalization, distance metric, block size, frequency band, image partitioning strategy, and representation method— on the classification performance. Finally, in the last section, the proposed algorithm's robustness against compression is tested.

## 4.1 Benchmark Databases

Due to tremendous interest in face recognition research, many face databases have become publicly available to comparatively evaluate the performance of the face recognition algorithms. Five of them have been chosen in order to test the robustness of the proposed face recognition algorithm against facial appearance variations caused by partial face occlusion, expression, illumination, time gap, and uncontrolled conditions. The used databases are:

- The face recognition grand challenge (FRGC) database [PFS+05],
- The CMU pose, illumination, and expression (CMU PIE) database [SBB03],
- The AR face database [MB98],
- The Yale face database B [GBK01], and
- The Extended Yale face database B [LHK05].

Overview of the data sets and experimental setups are given in Table 4.1.

In the experiments, all the images were aligned with respect to the eye centers and scaled to $64 \times 64$ pixels resolution.

In the following subsections information about these databases, the derived data sets, and the experimental setups are given.

| Name of the database | Number of subjects | Number of training images per subject | Number of testing images per subject | Contained variations |
|---|---|---|---|---|
| FRGC | 120 | 10 / 10 | 10 / 10 | Controlled, uncontrolled conditions, expression, time gap |
| CMU PIE | 68 | 1 | 20 | Illumination |
| AR | 110 | 1 / 1 / 1 / 1 | 1 / 1 / 1 / 1 | Occlusion (sun glasses and scarf), time gap |
| Yale | 10 | 7 / 7 / 7 / 7 | 12 / 12 / 14 / 19 | Illumination with different strength levels |
| Extended Yale | 38 | 7 / 7 / 7 / 7 | 12 / 12 / 14 / 19 | Illumination with different strength levels |

Table 4.1: Overview of the data sets and experimental setups. Multiple numbers in the cells indicate the number of training, testing samples used for each training-testing condition from the same database.

### 4.1.1 The Face Recognition Grand Challenge Database

The FRGC database was collected at the University of Notre Dame during 2002-2004 academic years. It is the database that has been used in the face recognition grand challenge experiments which has been conducted by the National Institute of Standards and Technology (NIST) from May 2004 to March 2006 [PFS+05]. The database contains high resolution still images and 3D images. The still images are collected both in a controlled and uncontrolled way. The images collected in a controlled way contain frontal faces that are captured in a studio setting. These images have two different lighting conditions caused by use of two or three studio lights, and two different facial expressions which are neutral and smiling. The average distance between the centers of eyes is 261 pixels. The images that are collected in an uncontrolled way also contain frontal face images, but this time instead of having a studio setting, the data is captured under changing illumination conditions in hallways, atria, or outdoors. They also contain two different facial expressions —neutral and smiling. The average distance between the centers of eyes in this case is 144 pixels. For detailed information about the database please see [PFS+05]. From the FRGC database, two different data sets were derived that consist of still images. One of the data sets contains images taken under controlled settings and the other contains

images taken under uncontrolled settings. These images are selected from the fall 2003 and spring 2004 recordings. There are 120 subjects in the created data sets, who have at least ten images both in fall 2003 and spring 2004 recordings. The images from fall 2003 are used for training and the ones from spring 2004 are used for testing. Table 4.2 shows the setups of the experiments on the FRGC database. Sample input images and registered images are given in Figure 4.1.

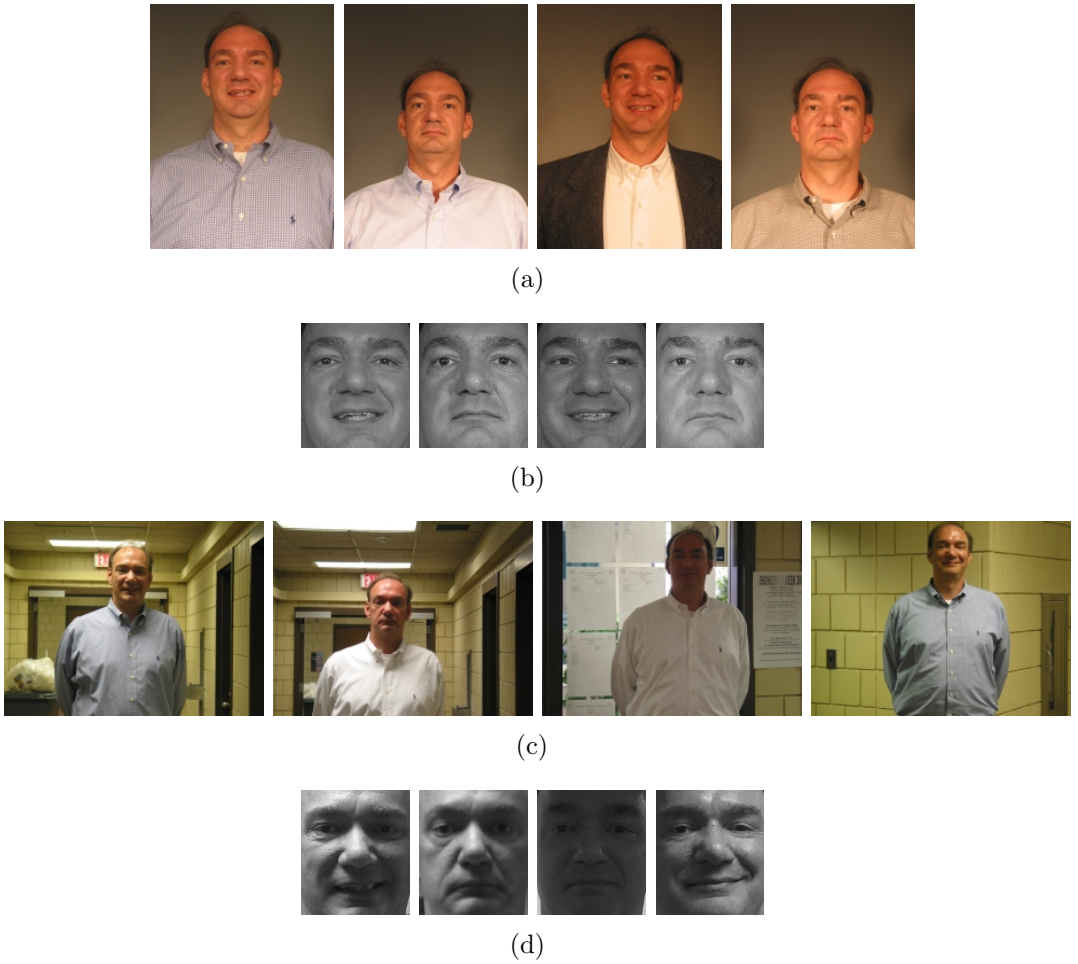


Figure 4.1: Sample images from the FRGC database. (a) Sample input images collected under controlled conditions from fall 2003. (b) Corresponding registered images from fall 2003. (c) Sample input images collected under uncontrolled conditions from spring 2004. (d) Corresponding registered images from spring 2004.

| Label of the experiment | Number of subjects | Training set | Number of training images per subject | Testing set | Number of testing images per subject |
|---|---|---|---|---|---|
| FRGC1 | 120 | Controlled face images from fall 2003 | 10 | Controlled face images from spring 2004 | 10 |
| FRGC4 | 120 | Uncontrolled face images from fall 2003 | 10 | Uncontrolled face images from spring 2004 | 10 |

Table 4.2: Experiments on the FRGC database.

### 4.1.2 The CMU Pose, Illumination, and Expression Database

The CMU PIE face database was collected at Carnegie Mellon University between October and December 2000. As the name implies, the database contains face images with varying head pose, illumination, and facial expression. There are 68 subjects in the database. In this work, frontal images from the illumination subset of the CMU PIE database are used, where each subject has 21 images captured under varying illumination conditions. Changing illumination conditions are provided by controlling 21 flashes during the image capture. For details about the capture conditions please see [SBB03]. The average distance between the centers of eyes is 82 pixels. In the experiments, for each subject a single image, where the face is frontally illuminated, is used for training and the remaining 20 face images, which are taken under varying illumination conditions, are used for testing. The setup of the experiment is given in Table 4.3. Sample input images and registered images are shown in Figure 4.2.

| Label of the experiment | Number of subjects | Training set | Number of training images per subject | Testing set | Number of testing images per subject |
|---|---|---|---|---|---|
| CMUPIE | 68 | Frontally illuminated face image | 1 | Face images under varying illumination | 20 |

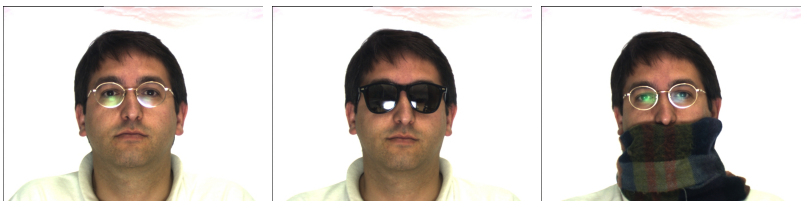
Table 4.3: Experiment on the CMU PIE face database.

(a)

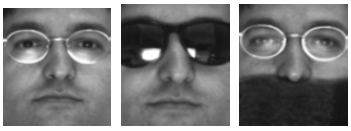
(b)

(c)

Figure 4.2: Sample images from the CMU PIE face database. (a) Sample input images. (b) Registered training image of a subject. (c) Registered testing images of the same subject.



(a)



(b)

Figure 4.3: Sample images from the AR face database. (a) Sample input images. (b) Corresponding registered images.

### 4.1.3 The AR Face Database

The AR face database was collected at the Computer Vision Center at the Universitat Autonoma de Barcelona (UAB), in Spain, in 1998 [MB98]. The database contains frontal face images with different facial expression, illumination, and occlusion. There are 126 subjects in the database. Each subject was recorded in two separate sessions. There is a time gap of two weeks between the sessions. The average distance between the centers of eyes is 105 pixels. From the database, one image per subject is used from the first session for training. This image is annotated as "1: neutral expression". For testing, two images per subject are used from each session, which are annotated as "8/21: wearing sunglasses", "11/24: wearing scarf", where the first number corresponds to the label in the first recording session and the second one corresponds to the label in the second recording session. From these two images, the ones with annotations "8/21: wearing sunglasses" are used for testing against upper face occlusion and the ones with annotations "11/24: wearing scarf" are used for testing against lower face occlusion. In the data set, there are 110 subjects who have all these samples in both of the sessions. Four separate experiments are conducted on this data set. Two of them are trained and tested within the first session and two of them are trained with the images from the first session and tested with the images from the second session. Setups of the experiments are presented in Table 4.4. Sample input images and registered images are shown in Figure 4.3.

| Label of the experiment | Number of subjects | Training set | Number of training images per subject | Testing set | Number of testing images per subject |
|---|---|---|---|---|---|
| AR1scarf | 110 | Face images without occlusion from session 1 | 1 | Face images with scarf from session 1 | 1 |
| AR1sun | 110 | Face images without occlusion from session 1 | 1 | Face images with sunglasses from session 1 | 1 |
| ARinterscarf | 110 | Face images without occlusion from session 1 | 1 | Face images with scarf from session 2 | 1 |
| ARintersun | 110 | Face images without occlusion from session 1 | 1 | Face images with sunglasses from session 2 | 1 |

Table 4.4: Experiments on the AR face database.

### 4.1.4 The Yale Face Database B / The Extended Yale Face Database B

The Yale face database B and the extended Yale face database B were collected at Yale University. The database contains pose and illumination variations. There are 10 subjects in the Yale face database B and 38 subjects in the extended Yale face database B. The Yale face database B is a subset of the extended Yale face database B. Separate experiments are conducted on the Yale face database B and the extended Yale face database B in order to compare the results with the ones in the literature that are obtained on these databases. Illumination variations are obtained by using a geodesic lighting rig with 64 computer controlled strobes. This way, for each person, in each pose, 64 images with different illumination conditions have been captured. These 64 images are divided into five subsets according to the angle between light source direction and camera's optical axis. Subset 1, with the angles less than 12 degrees, contains seven images. Subset 2, with the angles between 20 and 25 degrees, contains 12 images. Subset 3, with the angles between 35 and 50 degrees, contains 12 images. Subset 4, with the angles between 60 and 77 degrees, contains 14 images and finally subset 5, with the angles larger than 77 degrees, contains 19 images. From the database frontal face images under all illumination variations were selected. The average distance between eye centers is 92 pixels. For training, the first subset that has close to frontal illumination is used. For testing, subsets 2, 3, 4, and 5 are used. With increasing subset number, illumination variations become stronger as can be observed from the sample images in Figure 4.4. Setups of the experiments are presented in Table 4.5.

(a)

(b)

(c)

(d)

(e)

(f)

Figure 4.4: Sample images from the Yale face database B. (a) Sample input images. (b) Sample registered images from Subset 1. (c) Sample registered images from Subset 2. (d) Sample registered images from Subset 3. (e) Sample registered images from Subset 4. (f) Sample registered images from Subset 5.

| Label of the experiment | Number of subjects | Training set | Number of training images per subject | Testing set | Number of testing images per subject |
|---|---|---|---|---|---|
| Yale2 | 10 | Face images from subset 1 | 7 | Face images from subset 2 | 12 |
| Yale3 | 10 | Face images from subset 1 | 7 | Face images from subset 3 | 12 |
| Yale4 | 10 | Face images from subset 1 | 7 | Face images from subset 4 | 14 |
| Yale5 | 10 | Face images from subset 1 | 7 | Face images from subset 5 | 19 |
| ExtYale2 | 38 | Face images from subset 1 | 7 | Face images from subset 2 | 12 |
| ExtYale3 | 38 | Face images from subset 1 | 7 | Face images from subset 3 | 12 |
| ExtYale4 | 38 | Face images from subset 1 | 7 | Face images from subset 4 | 14 |
| ExtYale5 | 38 | Face images from subset 1 | 7 | Face images from subset 5 | 19 |

Table 4.5: Experiments on the Yale face database B / Extended Yale face database B.

## 4.2 Feature Normalization

Feature normalization is a very important processing step in local appearance-based face recognition using discrete cosine transform. There are two main points that should be taken into consideration. The first point is the total magnitude of each block's DCT coefficients. Since DCT is an orthonormal transformation and conserves all the energy of the processed input blocks, blocks with different brightness levels lead to DCT coefficients with different value levels. Because of this reason, the blocks with brighter content have more impact on the classification results. The other main point is the value range of the DCT coefficients. The first coefficients have higher magnitudes than the later ones. Therefore, they contribute more to the calculated distance in a nearest neighbor classification scheme, hence have more importance in the classification. However, it is known that, being able to represent more energy does not imply having more discriminative power [ES04].

In order to prevent the problems that may occur due to the imbalance between the blocks' impact to the classification, the feature vector extracted from each block is normalized to unit norm. Let $\boldsymbol{f_i}$ be a set of DCT coefficients used as the feature vector of the $i^{th}$ block in the image, then the normalized feature vector $\boldsymbol{f_i^B}$ becomes:

$$\boldsymbol{f_i^B} = \boldsymbol{f_i} / \left\| \boldsymbol{f_i} \right\| . \tag{4.1}$$

In order to solve the second problem —the imbalance between the coefficients' impact to the classification—, the coefficients are divided by their standard deviations that are learned from the training samples of all the used data sets. The standard deviation is calculated over all blocks, that is, there are no block specific values for the coefficients. The standard deviations obtained this way are illustrated in Figure 4.5. It can be seen from this figure that the first coefficients have a higher magnitude range. Let $f_{i,j}$ be the $j^{th}$ DCT coefficient from the $i^{th}$ block in the image and $\sigma\left(f_j\right)$ be the standard deviation of the $j^{th}$ DCT coefficient. The normalized coefficient, $f_{i,j}^C$, is calculated as

$$f_{i,j}^C = f_{i,j} / \sigma\left(f_j\right), \tag{4.2}$$

and the normalized feature vector $f_i^C$ consists of these normalized coefficients

$$\boldsymbol{f_i^C} = \left[ f_{i,1}^C, f_{i,2}^C, \ldots, f_{i,M}^C \right], \tag{4.3}$$

where $M$ denotes the local feature vector dimension.

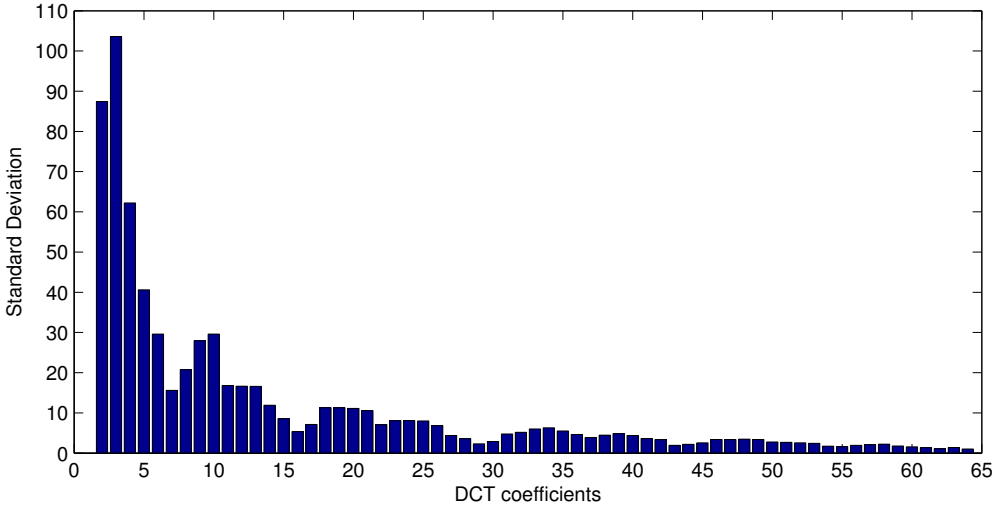



Figure 4.5: Standard deviations of the DCT coefficients. DCT coefficients are ordered according to the zig-zag scan pattern. The $0^{th}$ coefficient is excluded.

In order to balance the contributions of both the blocks and their coefficients to the classification at the same time, a combined normalization is performed by dividing the coefficients by their standard deviations and normalizing the local feature vector to unit norm:

$$\boldsymbol{f_i^{B,C}} = \boldsymbol{f_i^C} / \left\| \boldsymbol{f_i^C} \right\|. \tag{4.4}$$

In the feature normalization experiments, the $64 \times 64$ pixels resolution input face image is divided into $8 \times 8$ pixels resolution blocks. On each block DCT is applied. From the ordered DCT coefficients according to the zig-zag scan pattern, ten of them are selected by omitting the first DCT coefficient and selecting the following ten. The overall feature vector is constructed by concatenating the local feature vectors. It is then classified by using a nearest neighbor classifier. In the experiments, the effect of using different distance metrics to the classification results is also analyzed. Three different distance metrics, the L1 norm, the L2 norm, and the normalized correlation are compared,

$$d_{L1} = \sum_{k=1}^{K} \left| f_{training,k} - f_{test,k} \right|, \tag{4.5}$$

$$d_{L2} = \sum_{k=1}^{K} \left( f_{training,k} - f_{test,k} \right)^2, \tag{4.6}$$

$$d_{ncorr} = \frac{\boldsymbol{f_{training}} * \boldsymbol{f_{test}}}{|\boldsymbol{f_{training}}| * |\boldsymbol{f_{test}}|}, \tag{4.7}$$

where $K$ denotes the dimension of the overall feature vector, $f_{training,k}$ is the $k^{th}$ $(k = 1, 2, \ldots, K)$ component of the training feature vector, and $f_{test,k}$ is the $k^{th}$ component of the test feature vector.

The comparative results of different feature normalization methods are shown in Figures 4.6, 4.7, and 4.8. The comparison of using different distance metrics at the combined feature normalization is given in Figure 4.9. In each figure, the x-axis contains the experimental setups and the y-axis shows the corresponding correct recognition rates. Due to limited space, the labels of the experiments are shortened. In the shortened form, *F1* corresponds to *FRGC1*, *F4* corresponds to *FRGC4*, *CP* corresponds to *CMUPIE*; *A1sc*, *A1sun*, *Aisc*, *Aisun* correspond to *AR1scarf*, *AR1sun*, *ARinterscarf* and *ARintersun*; *Y2*, *Y3*, *Y4*, *Y5* correspond to *Yale2*, *Yale3*, *Yale4* and *Yale5*; *EY2*, *EY3*, *EY4*, *EY5* correspond to *ExtYale2*, *ExtYale3*, *ExtYale4*, *ExtYale5*, respectively.

From Figures 4.6, 4.7, and 4.8, it can be observed that at each experimental setup, the combined normalization has the best performance, except for the experiments *Yale2*, *Yale3*, and *ExtYale2*, where the correct recognition rate is already 100% without doing any normalization. Among the distance metrics, L1 norm is found to be the one which provides the highest correct recognition rate at each experimental setup as can be seen from Figure 4.9, except for the experiments *Yale2*, *Yale3*, *ExtYale2*, and *ExtYale3*, where with each distance metric 100% correct recognition rate is achieved. For example, in the experiment with the label *FRGC4*, with L1 norm as the distance metric, combined feature normalization achieves 90.8% correct classification rate, whereas unit norm feature normalization achieves 84.8% and feature normalization by dividing the DCT coefficients to their standard deviations achieves 62.8%. Without doing any feature normalization the obtained correct classification rate is 63.2%. The significant improvement justifies the necessity of feature normalization in local appearance-based face recognition using DCT. Interestingly, unit norm feature normalization provides much better results than standard deviation-based feature normalization. Moreover, applying standard deviation-based feature normalization alone does not improve the resuls over applying no normalization in most of the cases. However, using it in combination with unit norm feature normalization contributes positively to the performance. This results indicates that in the local appearance-based face recognition approach, it is more important to equalize the impacts of the blocks. Without doing this, thus having block importance directly proportional to the blocks' brightness levels causes a large drop in the performance. The benefit of equalizing the impact of the DCT coefficients is more visible, when the blocks' contributions to the classification are balanced.

The performance increase is higher when the experiment is difficult, that is, when there is a large difference between the appearances of the training and testing face images. The absolute performance improvement with combined

feature normalization with respect to applying no normalization is 4.1% in the experiment with the label *FRGC1* and 0.2% in the *ExtYale3*. In *Yale2*, *Yale3*, and *ExtYale2*, it is 0% since in both cases the correct recognition rate achieved in these experiments is 100%. On the other hand, the absolute increase is 27.6% in *FRGC4*, 38.7% in *CMUPIE*, 33.6% in *AR1scarf*, 10.9% in *AR1sun*, 30.9% in *ARinterscarf*, 20.9% in *ARintersun*, 48.6% in *Yale4*, 83.6% in *Yale5*, 66.8% in *ExtYale4*, and 89.5% in *ExtYale5*. These values are calculated using the L1 norm as the distance metric. However, as can be seen from Figures 4.7 and 4.8, similar observations are also valid when the L2 norm or normalized correlation is used. The main reason for having more improvement in the difficult classification experiments is the higher imbalance between the blocks' brightness levels in these cases. When there is a large illumination variation, occlusion or uncontrolled conditions, some of the blocks on the face image contain very high intensity values compared to some of the other blocks on the same face image, thus dominating the classification decision. This can be seen, for example, from the Figure 4.4 (e) and (f). Therefore, in these cases, it is more important to balance the contribution of each block to the classification decision.

The following correct recognition rates are attained when the L1 distance is used as the distance metric and combined feature normalization is applied: 97.9% in *FRGC1*, 90.9% in *FRGC4*, 99.8% in *CMUPIE*, 91.8% in *AR1scarf*, 37.3% in *AR1sun*, 80.9% in *ARinterscarf*, 38.2% in *ARintersun*, 100% in *Yale2* and *Yale3*, 95.7% in *Yale4*, 95.2% in *Yale5*, 100% in *ExtYale2* and *ExtYale3*, 93.1% in *ExtYale4* and *ExtYale5*. As one can notice, except the upper face occlusion —the *AR1sun* and *ARintersun* experiments— the performance is very high even under very difficult illumination conditions. This shows that upper face occlusion is a bigger problem than the changes in expression, illumination variations, uncontrolled conditions, and lower face occlusion. As the illumination variation becomes stronger, e.g. in *ExtYale4* and *ExtYale5* experiments, the performance drops. However, it still remains high.

According to the findings in the feature normalization experiments, in the following experiments combined normalization will be applied on the utilized DCT coefficients and the L1 norm will be used as the distance metric for nearest neighbor classification.

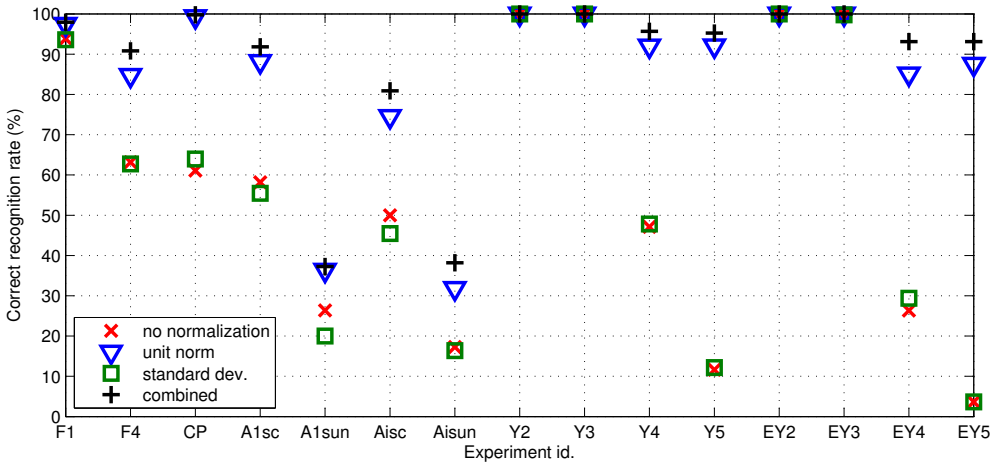



Figure 4.6: Comparison of feature normalization methods. The classification is done with a nearest neighbor classifier using the L1 norm as the distance metric.

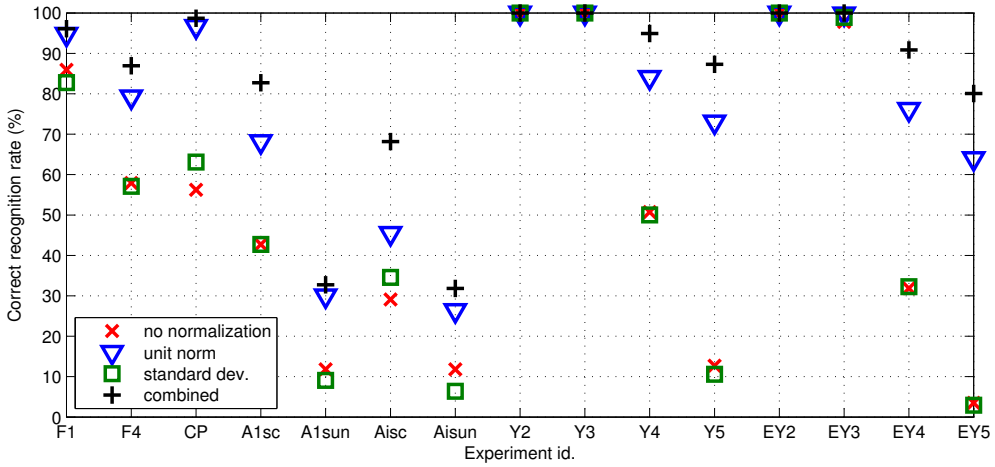



Figure 4.7: Comparison of feature normalization methods. The classification is done with a nearest neighbor classifier using the L2 norm as the distance metric.

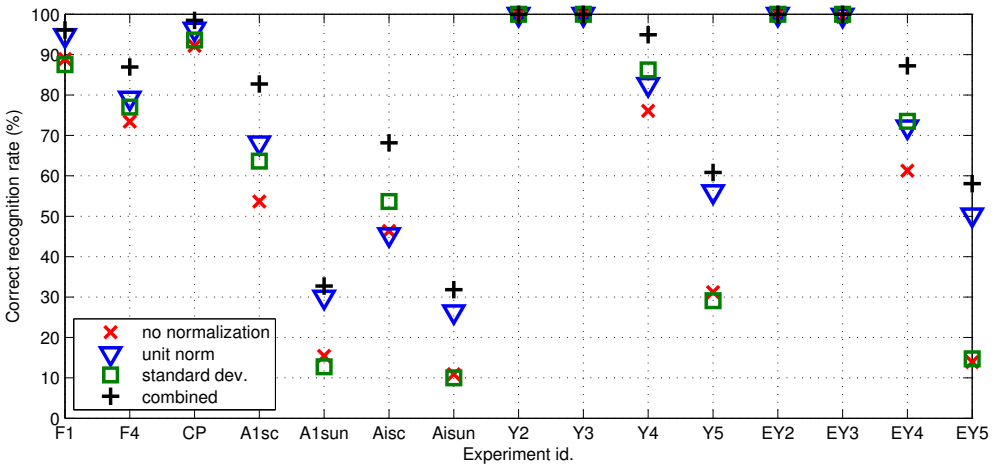



Figure 4.8: Comparison of feature normalization methods. The classification is done with a nearest neighbor classifier using the normalized correlation as the distance metric.

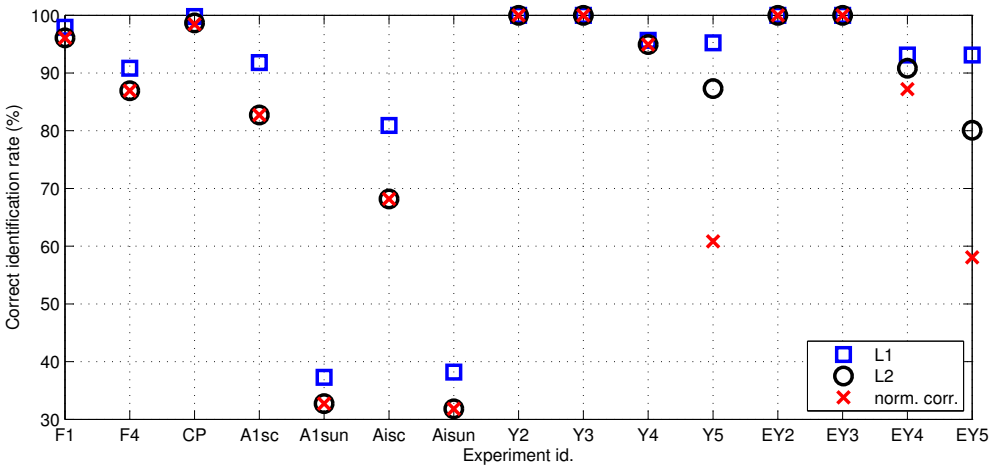



Figure 4.9: Comparison of distance metrics. Combined feature normalization is performed.

## 4.3 Block Size

Block size is one of the important parameters in local appearance-based face recognition. Applying DCT on large blocks provides more compactness in representation, however, it provides poor statistical representation of the block. On the other hand using small blocks as an input to DCT provides better statistical representation but less compactness. Hence, determining the block size is a trade-off between compactness and representation capability.

In order to observe the impact of block size on the face recognition performance, on each experimental setup, local appearance-based face recognition is performed with varying block sizes. Six different block sizes with the following pixel resolutions, $2 \times 2$, $4 \times 4$, $8 \times 8$, $16 \times 16$, $32 \times 32$, $64 \times 64$, are compared. The partitionings with different block resolutions are shown in Figure 4.10. The block resolution of $64 \times 64$ pixels corresponds to the entire face image. So, at this block size, the approach is based on whole appearance rather than the local appearance.

Figure 4.10: Face image partitioning with different block resolutions. The block resolutions are $2 \times 2$, $4 \times 4$, $8 \times 8$, $16 \times 16$, $32 \times 32$, $64 \times 64$, from left to right, respectively. The block resolution of $64 \times 64$ pixels corresponds to the entire face image.

In the block size experiments, the $64 \times 64$ pixels resolution input face image is divided into blocks with a certain block size. On each block DCT is applied. From the ordered DCT coefficients according to the zig-zag scan pattern, $M$ of them are selected by omitting the first DCT coefficient and selecting the following first $M$ of them. The selected coefficients are divided by their standard deviations that have been calculated for that block size from the training samples. Afterwards, the local feature vector is normalized to the unit norm. The overall feature vector is constructed by concatenating the local feature vectors. The global feature vector is then classified by using a nearest neighbor classifier. The number of obtained blocks can be calculated by simply dividing the image resolution to the block resolution. For example, in the case of using $2 \times 2$ pixels resolution blocks, there are 1024 blocks and in the case of using $64 \times 64$ pixels resolution blocks there is only one block which is the face image itself. The global feature dimension is calculated as the number of blocks times the dimension of the local feature vector. Therefore, the global feature dimension

can be only the multiples of number of blocks. For instance, the global feature dimension can be 1024, 2048, and 3072, when the block size of $2 \times 2$ is used.

The results of the block size experiments at different global feature vector dimensions can be seen from Figures 4.11-4.25. At dimensions that are higher than 2000, the blocks with resolution $2 \times 2$, $4 \times 4$, $8 \times 8$, $16 \times 16$ pixels outperform the blocks with resolution $32 \times 32$ and $64 \times 64$ pixels. The performance increases rapidly with the increasing global feature dimension till some point. After that point, depending on the experiment, the correct recognition rate either remains the same, increases slightly or decreases. For the experiments with the face images that contain strong illumination variations, such as *CMUPIE*, *Yale4*, *Yale5*, *ExtYale4*, and *ExtYale5*, it remains the same or it increases even slightly at some block sizes. The reason is that, in DCT-based representation, having higher dimensional feature vector implies adding the DCT coefficients that correspond to higher frequency content to the feature vector. Since it is known that these coefficients are less sensitive to the illumination changes, adding them to the feature vector improves the performance. In the experiments that do not contain strong illumination changes, the performance decreases by the increasing feature dimensionality. Overall, as can be derived from the Figures 4.11-4.25, using only a portion of the DCT coefficients would suffice to have high correct recognition rates. Moreover, using lower dimensional feature vectors decreases the computational load, which facilitates real-time processing. In order to analyze the results in detail at low dimension, the lowest common global feature vector dimension, which is 1024, is chosen and the correct classification rates obtained with different block sizes are compared. Note that, one can use even lower dimensional global feature vector at different block sizes. Nevertheless, since this is the lowest possible global feature size when using $2 \times 2$ pixel resolution blocks, this dimension is used to be able to compare all the block sizes. The results attained at this dimension are given in Table 4.6. In most of the cases, the best results are obtained with the $8 \times 8$ block size. The best result is obtained with $4 \times 4$ block size in *CMUPIE* and with $16 \times 16$ block size in *Yale5*. However, if one looks carefully at Figures 4.13 and 4.21, it can be observed that the block size of $8 \times 8$ outperforms the block sizes of $4 \times 4$ and $16 \times 16$ on these databases at dimensions lower than 1024. In addition to having better performance with respect to other block sizes, the $8 \times 8$ block size provides a good compromise between compactness and representation power. It is also the block size that is used in JPEG image compression standard which is based on DCT. Using DCT and having the same block size as in JPEG, makes the local appearance-based face recognition approach less sensitive to the problems that may arise due to compression. The only difference in representation between local appearance-based face recognition and JPEG is that there is no quantization step in the former, while there exists one in the latter. In the remaining experiments, only the block size of $8 \times 8$ pixels will be used.

| Experiment | $2 \times 2$ | $4 \times 4$ | $8 \times 8$ | $16 \times 16$ | $32 \times 32$ | $64 \times 64$ |
|---|---|---|---|---|---|---|
| FRGC1 | 97.2% | 98.2% | **98.5%** | 95.9% | 92.5% | 91.9% |
| FRGC4 | 67.6% | 89.8% | **91.4%** | 89.9% | 80.6% | 71.6% |
| CMUPIE | 96.1% | **100%** | 99.9% | 99.7% | 98.5% | 96.5% |
| AR1scarf | 87.3% | 89.1% | **90.9%** | 79.1% | 71.8% | 62.7% |
| AR1sun | 30.9% | 35.5% | **37.3%** | 25.5% | 15.5% | 12.7% |
| ARinterscarf | 69.1% | 79.1% | **82.7%** | 73.6% | 67.3% | 56.4% |
| ARintersun | 30.9% | 33.6% | **35.5%** | 26.4% | 14.5% | 10.9% |
| Yale2 | 100% | 100% | 100% | 100% | 100% | 100% |
| Yale3 | 100% | 100% | 100% | 100% | 100% | 100% |
| Yale4 | 89.9% | 95.7% | **98.6%** | 96.4% | 94.9% | 93.5% |
| Yale5 | 71.4% | 96.3% | 96.8% | **98.9%** | 84.% | 51.3% |
| ExtYale2 | 100% | 100% | 100% | 100% | 100% | 100% |
| ExtYale3 | 100% | 100% | 100% | 100% | 100% | 100% |
| ExtYale4 | 74.8% | 95.2% | **96.6%** | 94.1% | 83.2% | 82.4% |
| ExtYale5 | 43.5% | 96.4% | **97.1%** | 94.1% | 62.4% | 28.6% |

Table 4.6: Correct recognition rates obtained with different block sizes at the global feature dimension of 1024.

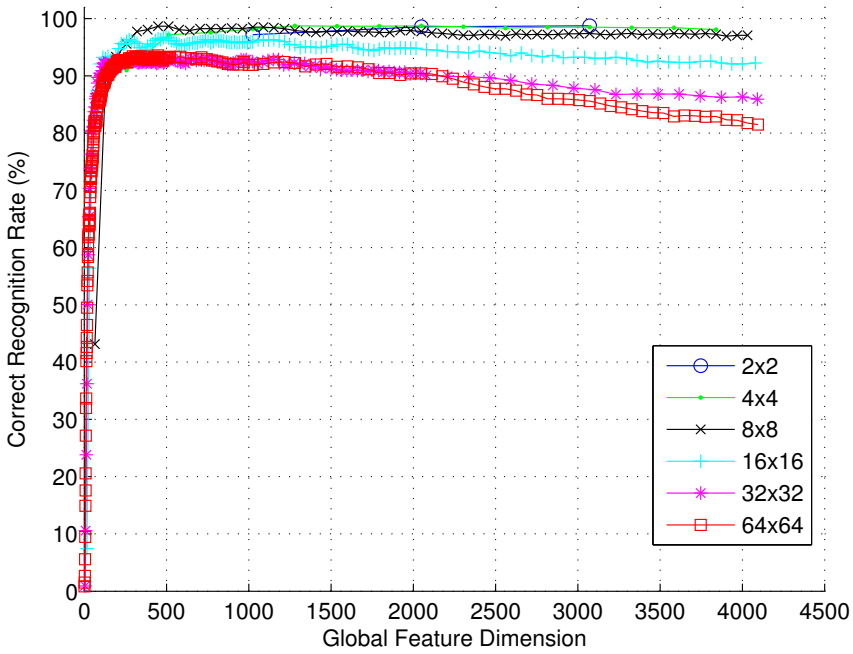



Figure 4.11: Comparison of different block sizes on *FRGC1*.

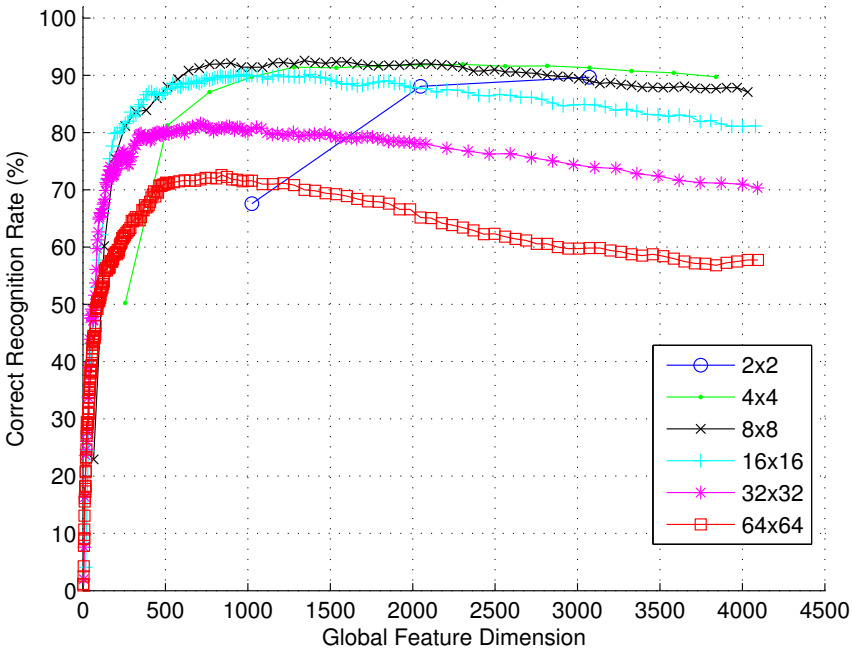



Figure 4.12: Comparison of different block sizes on *FRGC4*.

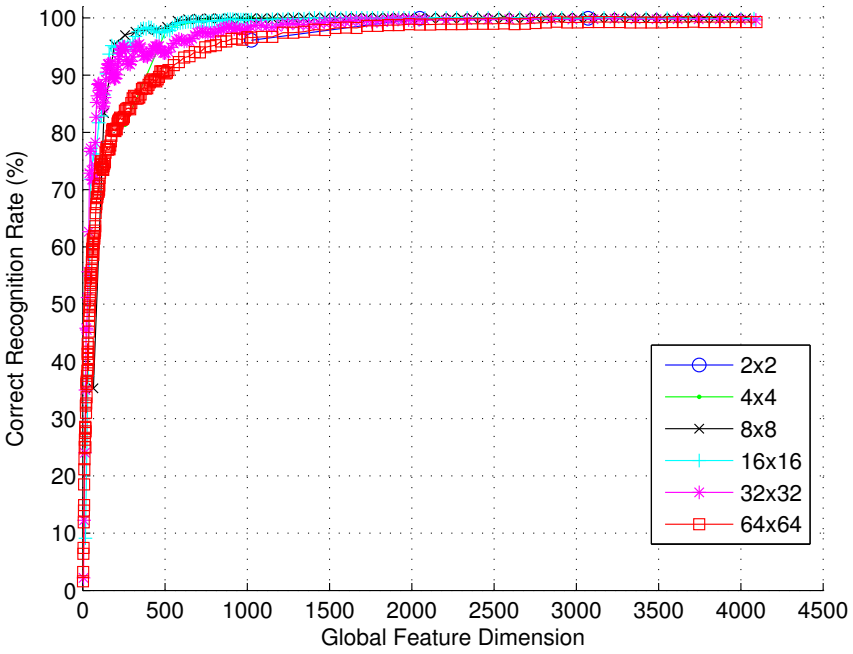



Figure 4.13: Comparison of different block sizes on *CMUPIE*.

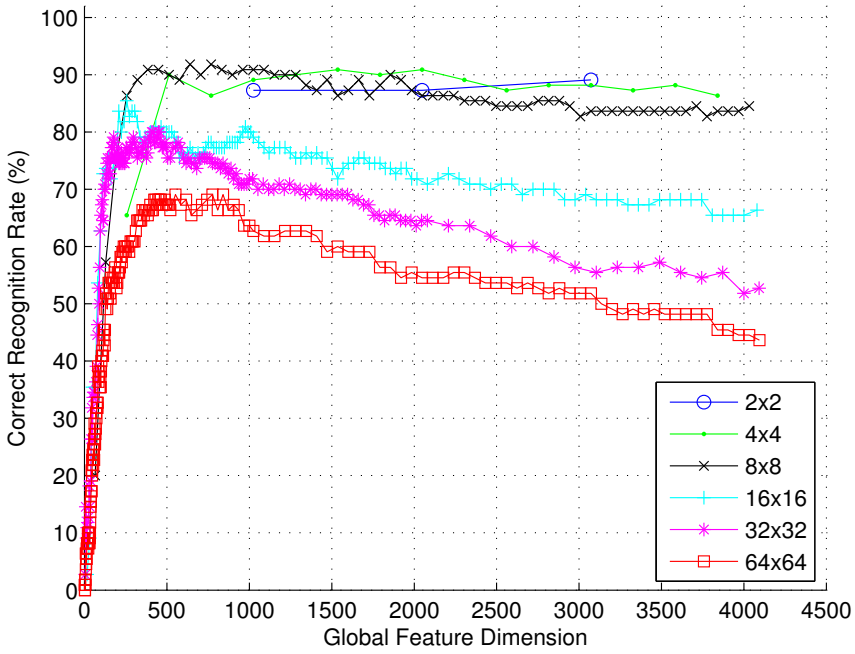



Figure 4.14: Comparison of different block sizes on *AR1scarf*.

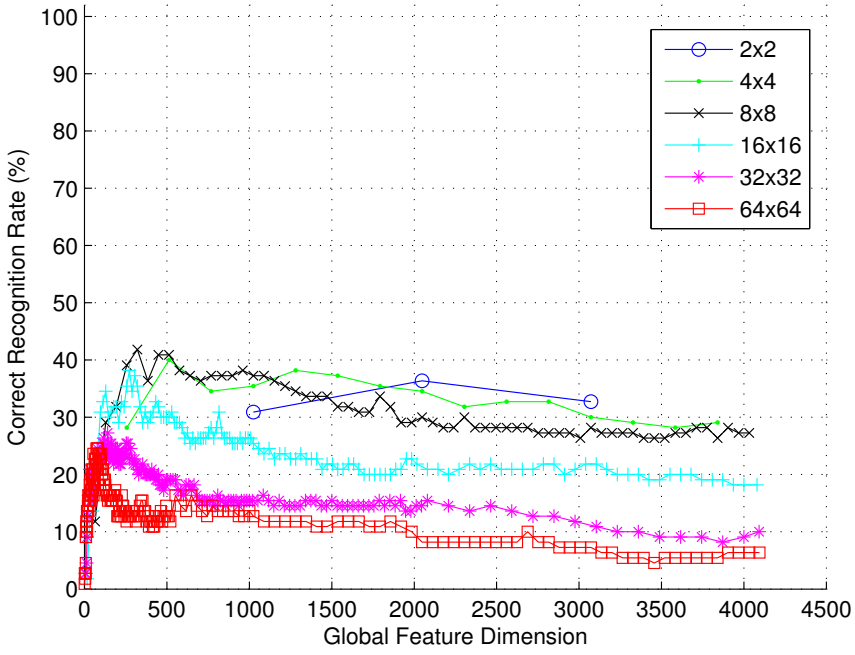



Figure 4.15: Comparison of different block sizes on *AR1sun*.

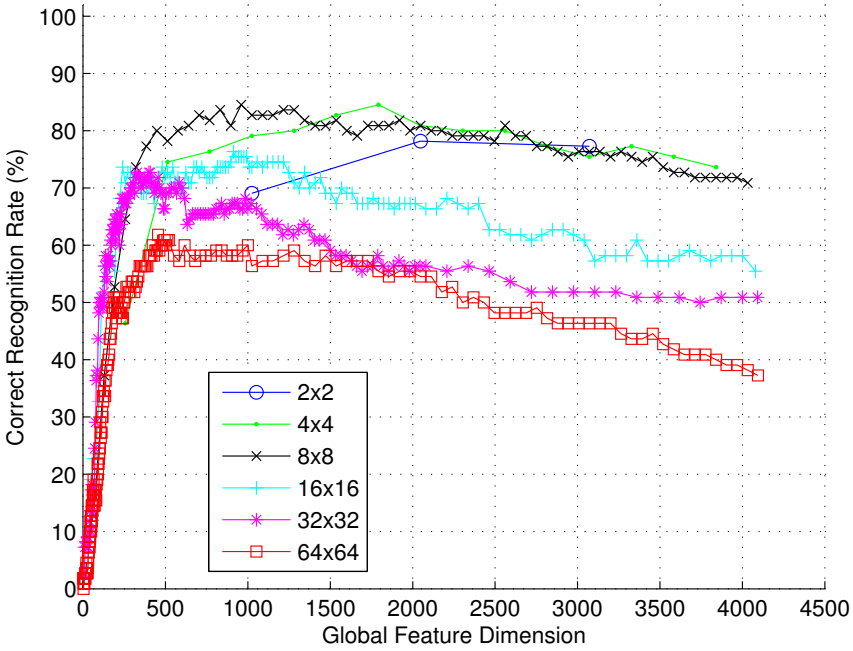



Figure 4.16: Comparison of different block sizes on *ARinterscarf*.

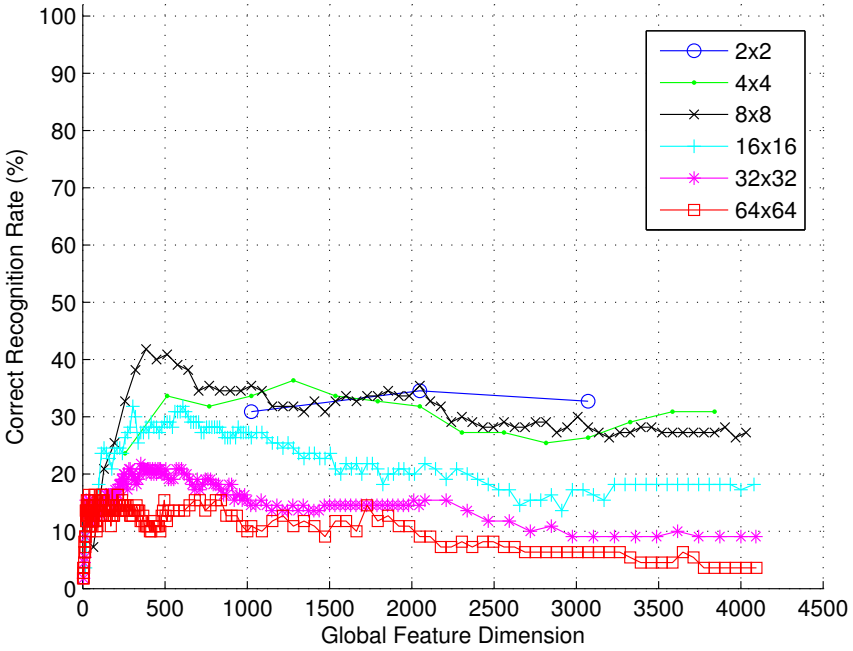



Figure 4.17: Comparison of different block sizes on *ARintersun*.

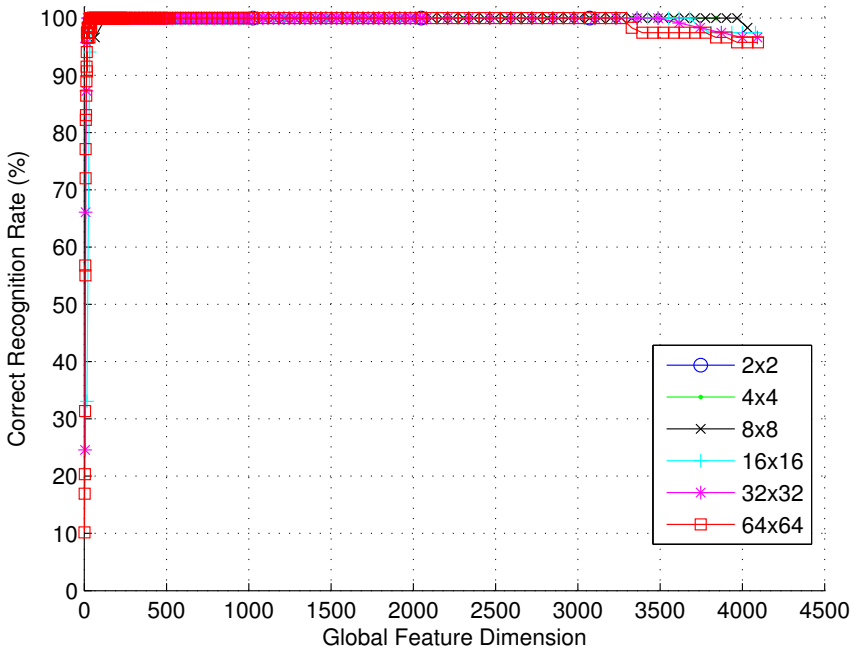



Figure 4.18: Comparison of different block sizes on *Yale2*.

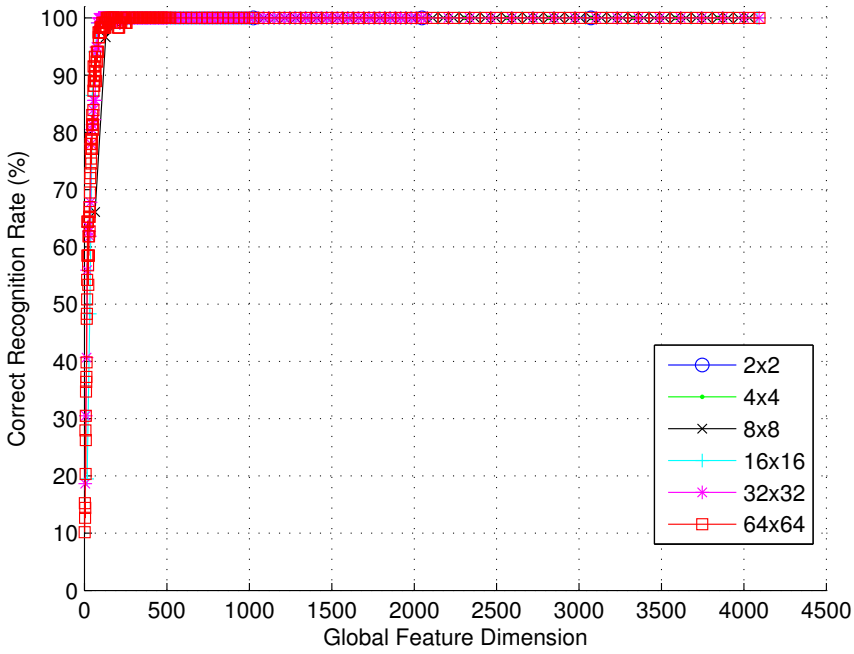



Figure 4.19: Comparison of different block sizes on *Yale3*.

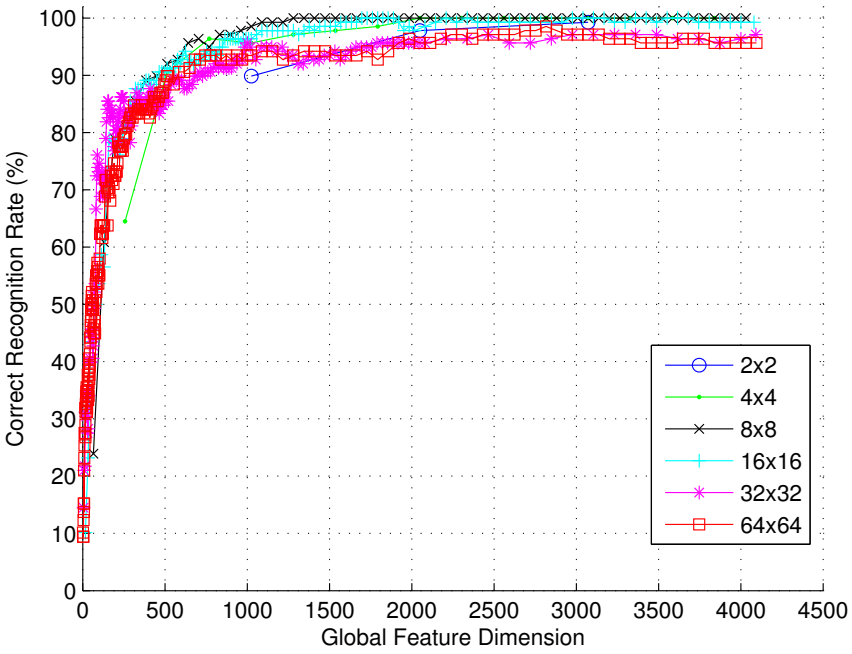



Figure 4.20: Comparison of different block sizes on *Yale4*.

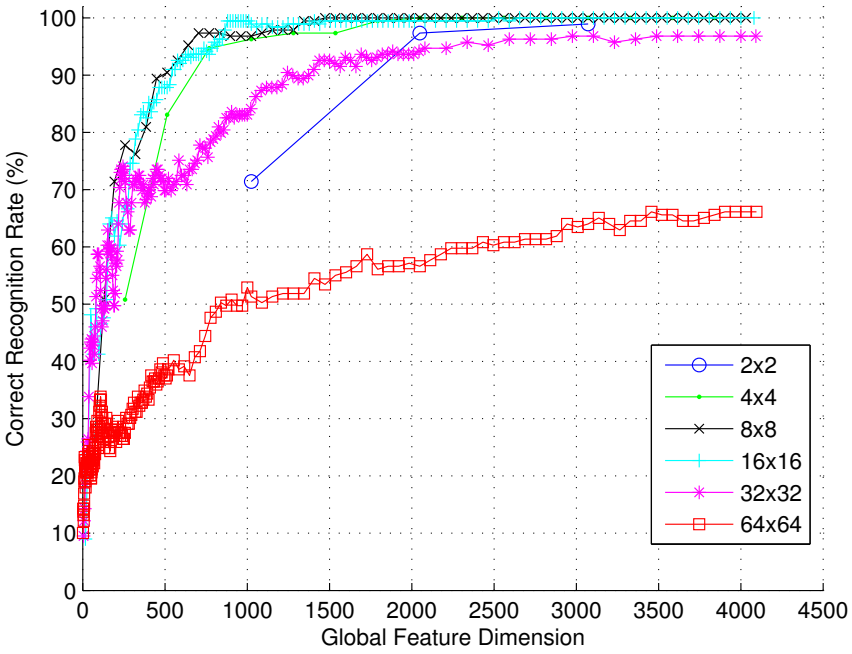



Figure 4.21: Comparison of different block sizes on *Yale5*.

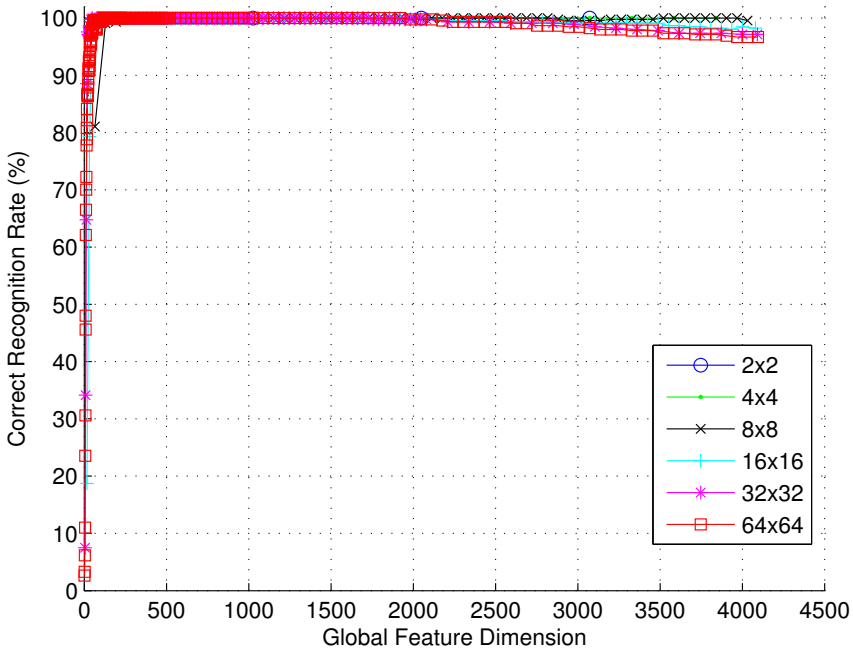



Figure 4.22: Comparison of different block sizes on *ExtYale2*.

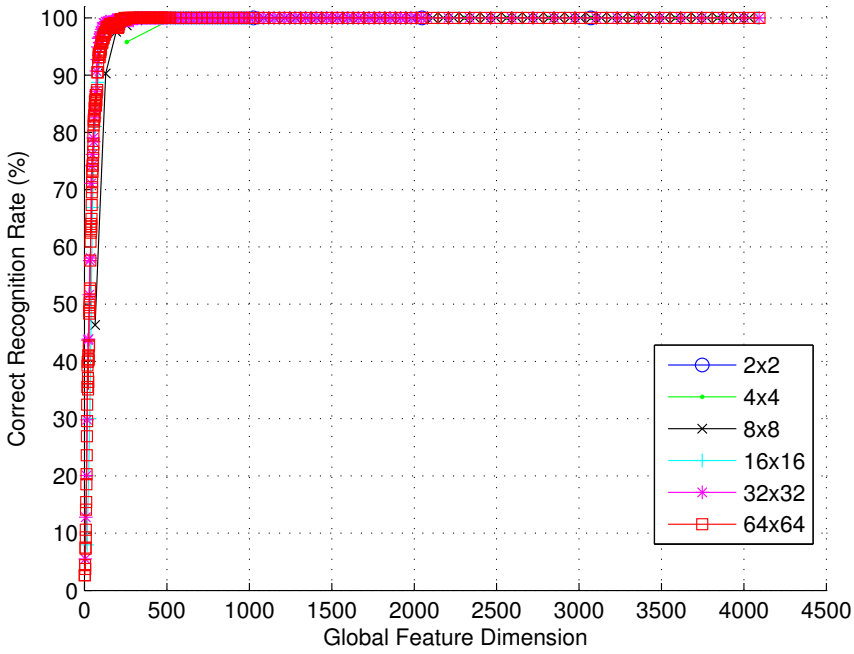



Figure 4.23: Comparison of different block sizes on *ExtYale3*.

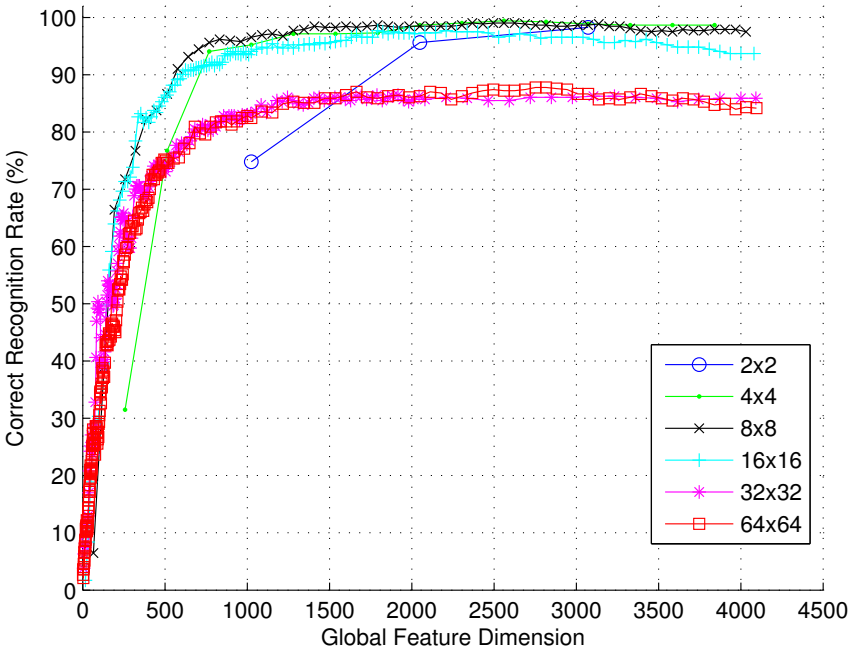



Figure 4.24: Comparison of different block sizes on *ExtYale4*.

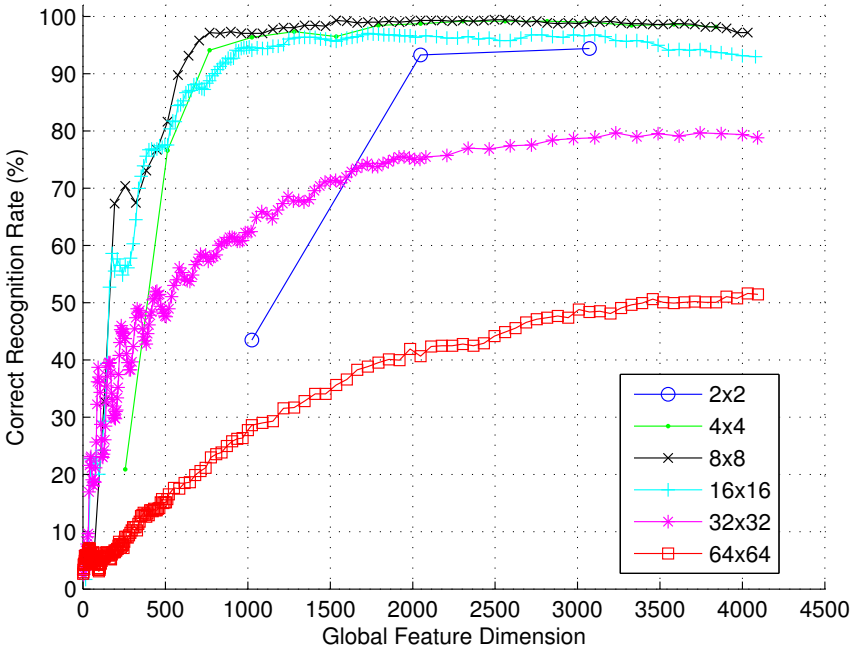



Figure 4.25: Comparison of different block sizes on *ExtYale5*.

## 4.4 Analysis of Frequency Bands

Another aspect in local appearance-based face recognition using the DCT is the selection of frequency content to be used for classification. That is, to determine the width and location of the window that the DCT coefficients are chosen from. As illustrated in Figure 4.26, each DCT basis has a different response which goes from coarser to finer as the basis index increases. Figure 4.26 shows each basis' output separately for an input image. If these outputs are analyzed in detail, it can be noticed that they depict how strong a specific basis pattern is observed in the corresponding block. For instance, the top-left output image contains the average values of the blocks, the one next to it shows the amount of vertical details, whereas the one below it shows the amount of horizontal details occurring in the blocks. The low frequency coefficients represent most of the input block's energy, whereas the higher frequency coefficients correspond to finer details. However, neither conserving more energy nor having finer details guarantees better discrimination. Depending on the identification task, the required local feature dimensionality and the frequency band may change. In order to observe the effect of feature dimensionality and frequency content simultaneously, a sliding window scheme is employed where windows with varying sizes are moved from the beginning to the end of the ordered DCT coefficients. The coefficients obtained this way are divided by their standard deviations and the local feature vector is normalized to unit norm. The results of the experiments can be seen from Figures 4.27-4.41. In the figures, the x-axes show how many DCT coefficients are removed from the beginning, while y-axes show the local feature dimension. The number of possible shifts depends on the dimensionality of the local feature vector. For example, when the local feature dimension is two, there are 63 possible shifts and when 63-dimensional local feature is used only two shifts are possible. The upper diagonals in the figures are padded with zeros, since at that region there exists no local feature dimension and shift combination. Dark red color indicates high correct recognition rates, whereas dark blue color corresponds to low correct recognition rates. Since the main goal in this experiment is to observe the relative performance of each frequency band, the color ranges are stretched for the illustration purposes. So that the best result gets the dark red color, even though it is not 100% correct classification.

On the *FRGC1* experiment (Figure 4.27), it is observed that the best result zone is the region where the number of removed coefficients is low and the local feature dimension is higher than a certain value. Quantitatively speaking, having more than four-dimensional local feature vectors by removing up to ten DCT coefficients from the beginning provides high correct classification rates on the *FRGC1* experiment. The same observation holds for the *FRGC4* experiment (Figure 4.28). However, it is better to use higher dimensional local feature vectors, that is the ones with more than seven or eight dimensions. On the *CMUPIE* experiment, a larger high performance zone is obtained. Besides

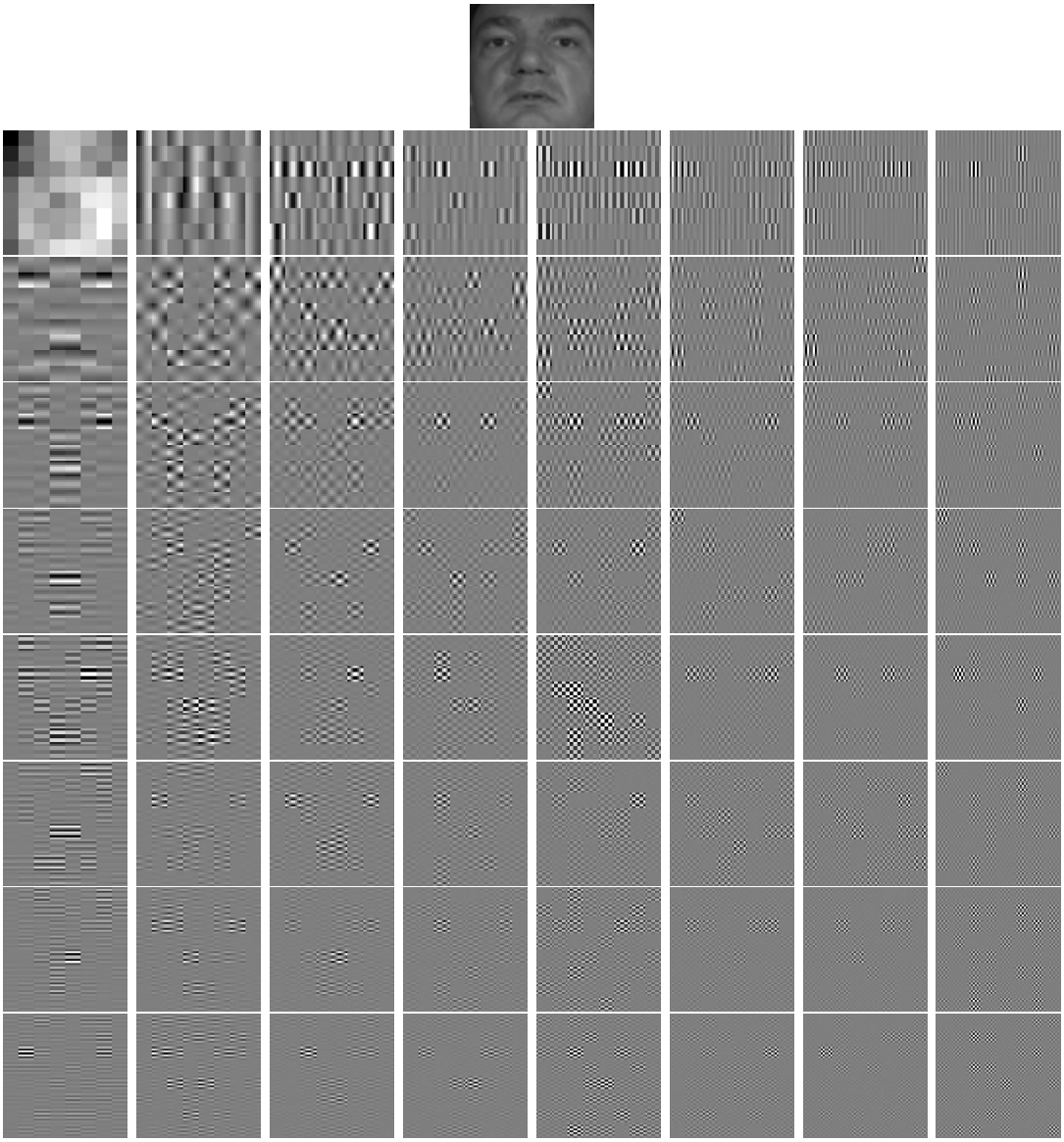


Figure 4.26: A sample frequency output of a face image. The face image at the top is the input image. The other images show the DCT outputs. Each image corresponds to an output transformed with a different basis. The order of the DCT outputs are the same as the order of the bases in Figure 3.2

some frequency bands with low dimensional feature vectors, the results are high. Even on some certain frequencies, with low dimensional feature vectors, the correct recognition rates remain high. The high performance zones from lower face occlusion experiments (Figures 4.30 and 4.32) are limited to a low number of removed coefficients and a certain local feature dimension range. The number of removed coefficients is up to six and the local feature dimension is between five and twenty. The region for upper face occlusion (Figures 4.31 and 4.33) is even more limited. A small region of low frequency content provides the best results. On the *Yale2*, *Yale3*, *ExtYale2*, and *ExtYale3* experiments (Figures 4.34, 4.35, 4.38, 4.39), where the illumination variations are not very strong, the correct classification rates are high no matter which frequency band with how many feature dimensions is used. On *Yale2* and *ExtYale2*, lower frequency content is found to be more discriminative than the higher frequency content, but the results are still high with the high frequency content. On the experiments with stronger illumination variations (Figures 4.36, 4.37, 4.40, 4.41), it has been observed that higher dimensional local feature vectors, local feature vectors with ten-dimension or higher, are required in order to reach a high correct classification rate.

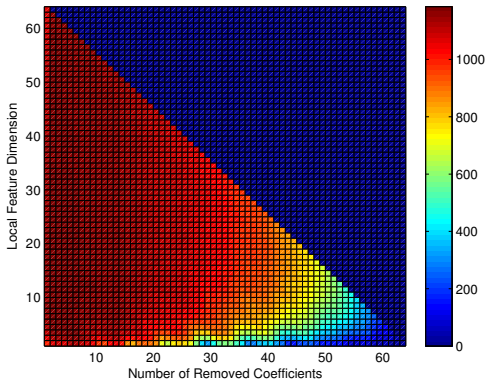



Figure 4.27: Comparison of different feature sets on *FRGC1*.

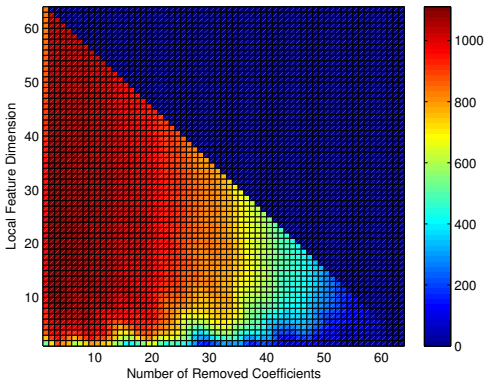



Figure 4.28: Comparison of different feature sets on *FRGC4*.

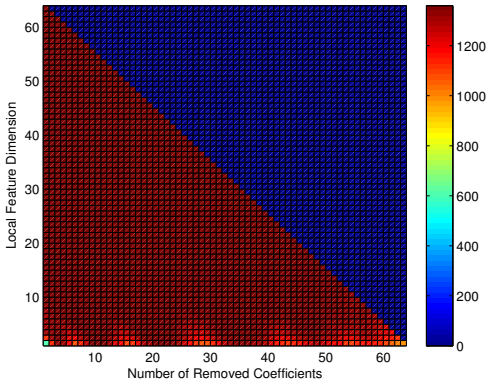



Figure 4.29: Comparison of different feature sets on *CMUPIE*.

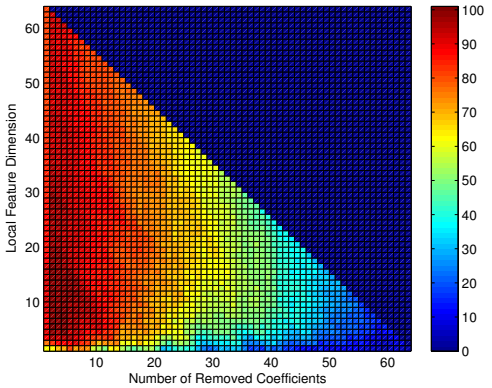



Figure 4.30: Comparison of different feature sets on *AR1scarf*.

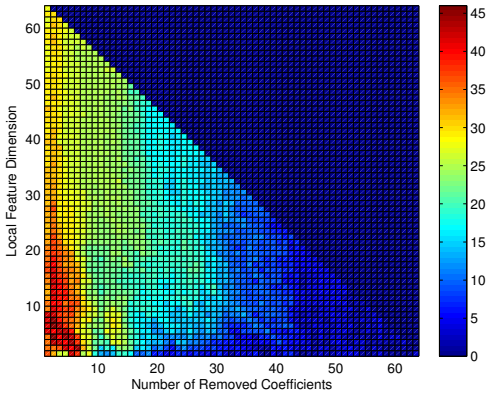



Figure 4.31: Comparison of different feature sets on *AR1sun*.

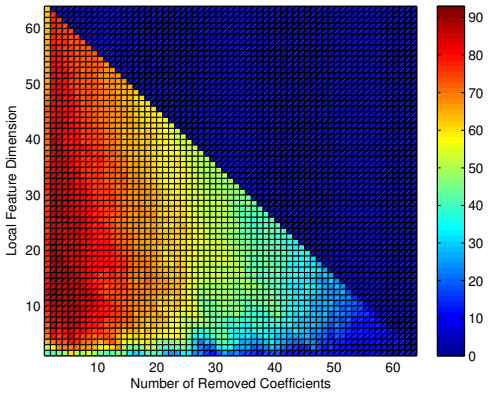



Figure 4.32: Comparison of different feature sets on *ARiscarf*.

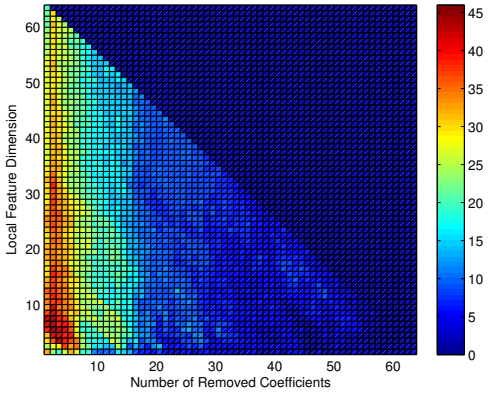



Figure 4.33: Comparison of different feature sets on *ARisun*.

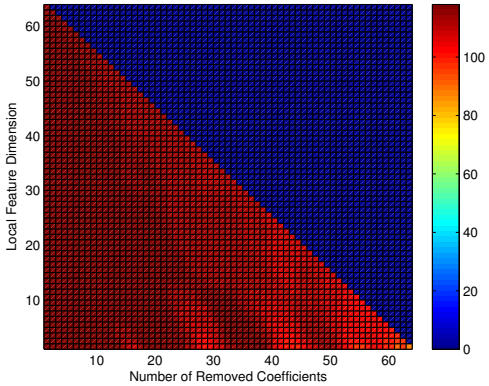



Figure 4.34: Comparison of different feature sets on *Yale2*.

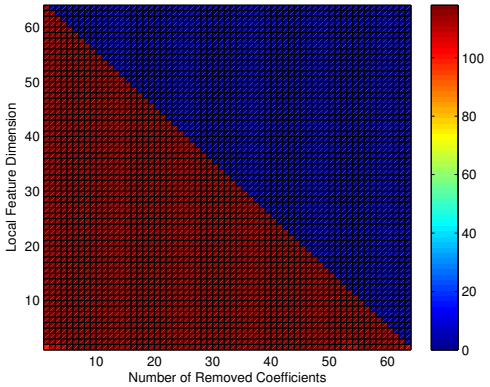



Figure 4.35: Comparison of different feature sets on *Yale3*.

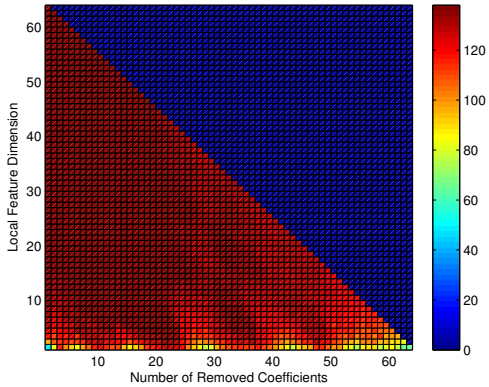



Figure 4.36: Comparison of different feature sets on *Yale4*.

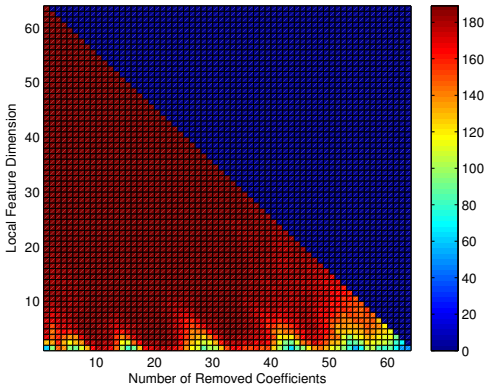



Figure 4.37: Comparison of different feature sets on *Yale5*.

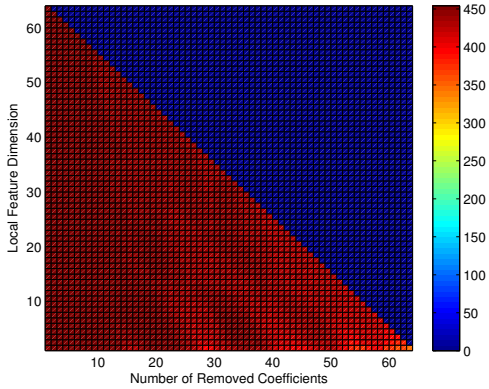



Figure 4.38: Comparison of different feature sets on *ExtYale2*.

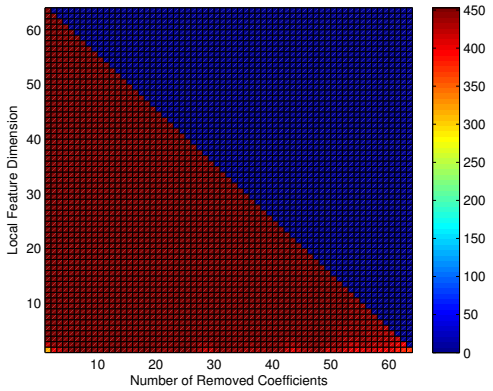



Figure 4.39: Comparison of different feature sets on *ExtYale3*.

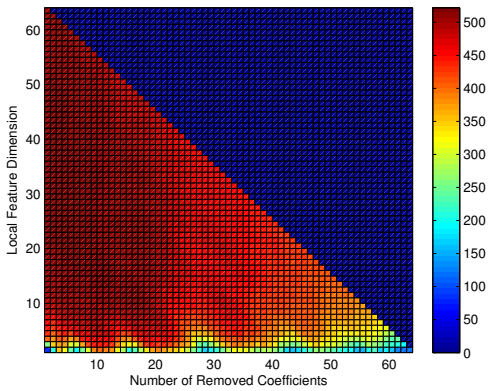



Figure 4.40: Comparison of different feature sets on *ExtYale4*.

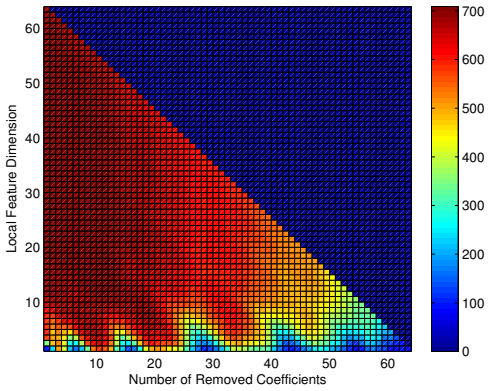



Figure 4.41: Comparison of different feature sets on *ExtYale5*.

In Figure 4.42 reconstruction outputs, that are generated using different frequency bands, are given. Different frequency bands that contain ten DCT coefficients are selected by moving a window from beginning to the end of the ordered DCT coefficients. This way 55 different frequency bands are obtained. In the figure, the top-left face image is the input face image, the one next to it corresponds to the reconstruction output with the first ten DCT coefficients, the third one in the first row corresponds to the reconstruction output with the DCT coefficients between and including the second and eleventh ones. The correspondence between the outputs and the used frequency band continues in this fashion. The number of DCT coefficients, that are removed from the beginning, increases from left to right and from top to bottom. The bottom-right image corresponds to the reconstruction output with the last ten DCT coefficients. As can be observed, the reconstructions with low frequency bands contain coarse information, whereas the ones with high frequency bands contain finer details.

The corresponding classification results for ten-dimensional feature vector are plotted in Figures 4.43, 4.44, 4.45, and 4.46. On the FRGC experiments (Figure 4.43), removing the first DCT coefficient improves the results. The results change only very slightly with additional removal of the coefficients. The correct recognition rate deteriorates if too many coefficients are removed. Similarly, on the AR experiments (Figure 4.44), where occlusion exists, the performance increases with the removal of the first coefficient and drops when too many coefficients are removed. On the experiments with illumination variations (Figures 4.45 and 4.46), more coefficients are required to be removed, since low frequency content is sensitive to the appearance changes due to illumination variations.

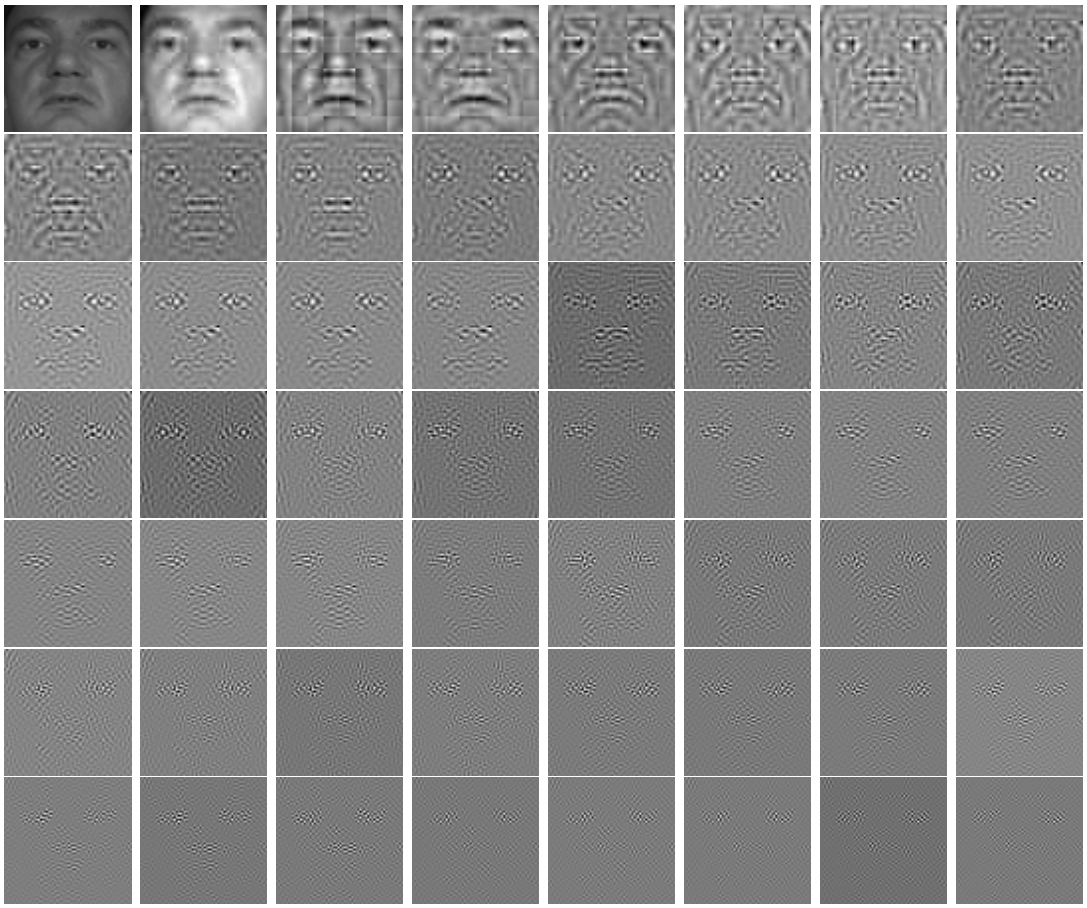


Figure 4.42: Reconstruction outputs generated by using different frequency bands.

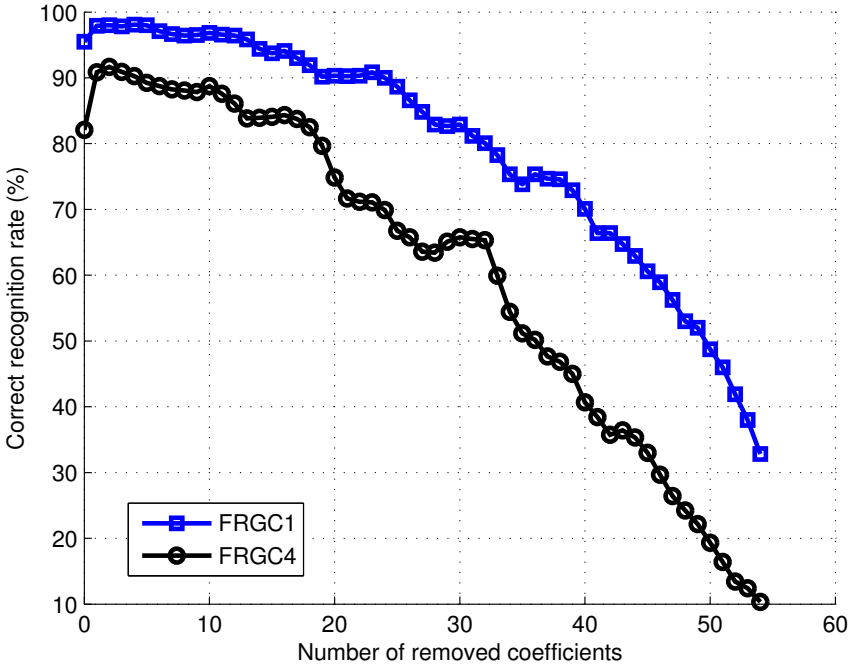



Figure 4.43: Comparison of ten-dimensional local features on the FRGC experiments.

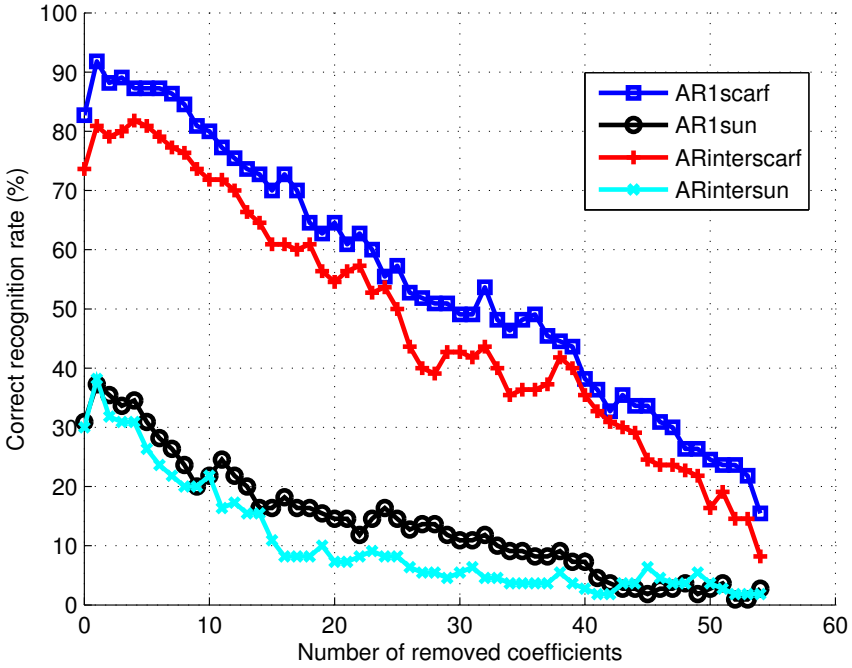



Figure 4.44: Comparison of ten-dimensional local features on the AR experiments.

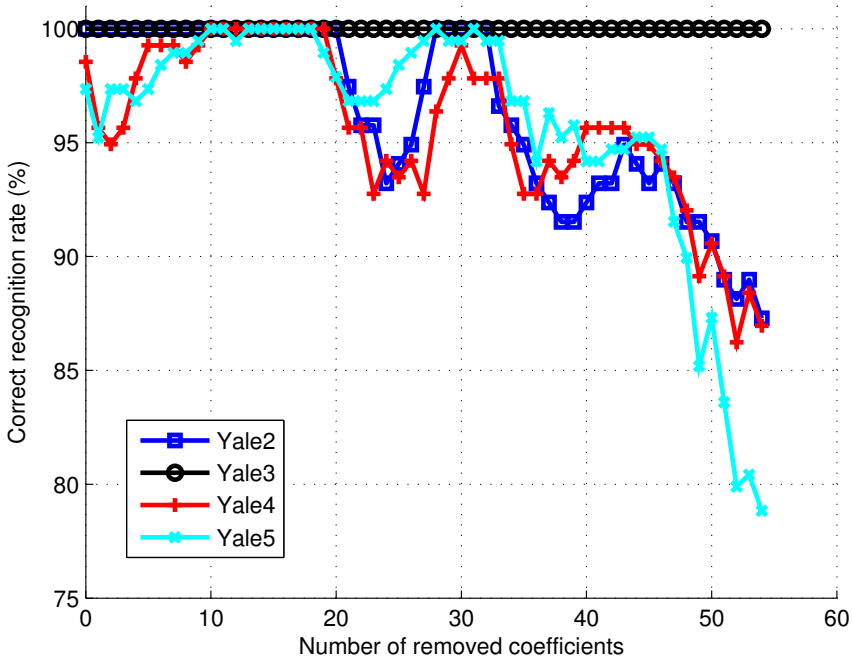



Figure 4.45: Comparison of ten-dimensional local features on the Yale experiments.

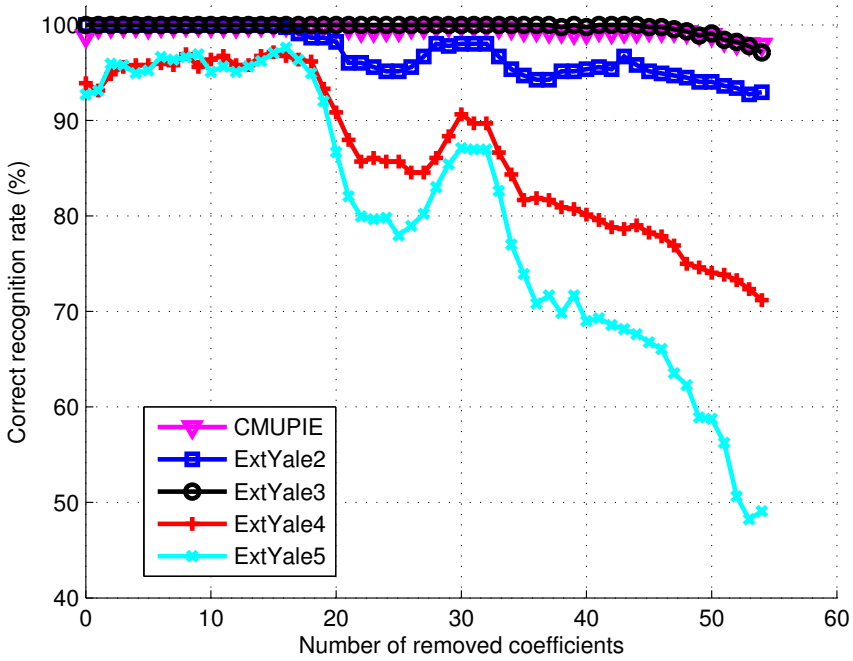



Figure 4.46: Comparison of ten-dimensional local features on the CMU PIE and extended Yale experiments.

## 4.5 Generic vs. Salient Region-based Partitioning

In this part of the experiments, five different salient region-based partitioning schemes, that are derived from previous modular/component/patch based studies [BP93, PMS94, HSP07, KKHK05, LKK05], are compared for the local appearance-based face recognition approach. These salient region-based partitioning schemes are also compared with generic partitioning of the face image. In the implementation, the salient regions are divided into $8 \times 8$ pixels resolution non-overlapping blocks and the DCT is applied on each block. From the DCT coefficients that are ordered according to the zig-zag scan pattern, ten of them are selected by omitting the first DCT coefficient and selecting the following ten of them. The selected coefficients are divided by their standard deviations. Afterwards, the local feature vector is normalized to the unit norm. The overall feature vector for a salient region is constructed by concatenating the local feature vectors that are extracted from the blocks of the corresponding salient region. The feature vector of the combined regions is generated by concatenating the local feature vectors of each region.

The first partitioning scheme (*P1*) is similar to the one in [BP93]. It consists of three regions: eyes, nose, and mouth. A sample image, illustrating this partitioning scheme, is given in Figure 4.47. The obtained results with the individual components and the combined representation on each experimental setup is shown in Figure 4.48. The correspondences between the abbreviations on the x-axis and the experiment labels are the same as the ones in Section 4.2. The best results are obtained with the combined representation except on the *ARintersun* experiment, where the mouth region provides the highest correct classification rate. The eye region is found to be the second best performing region, except in the experiments where upper face occlusion exists. On *Yale2* and *ExtYale2*, 100% correct recognition rate is achieved using only the eye region. Depending on the experimental setup, either the nose region or mouth region comes third. In the experiments with high illumination variations, such as *CMUPIE*, *Yale4*, *Yale5*, *ExtYale4*, and *ExtYale5*, the mouth region is found to be more useful for identification than the nose region. This is expected, since in the case of illumination variation due to cast shadows the appearance of the nose region is affected severely. The other reason for this outcome is the lack of expression variations in the used data sets for these experiments. The mouth region works also better in the experiments that contain upper face occlusion, namely, *AR1sun* and *ARintersun* experiments. Having sunglasses decreases the amount of the discriminative information that the nose region contains. In the *FRGC1*, *FRGC4*, *AR1scarf*, and *ARinterscarf* experiments the nose region reaches higher recogniton rates than the mouth region. The expression variations in *FRGC1* and *FRGC4* experiments deteriorate the performance of the

mouth region. Obviously, in the case of lower face occlusion, as in *AR1scarf* and *ARinterscarf* experiments, the mouth region has no use.

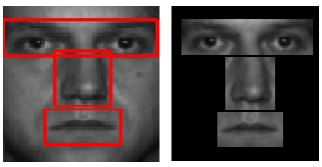


Figure 4.47: Salient regions obtained with the *P1* partitioning scheme.

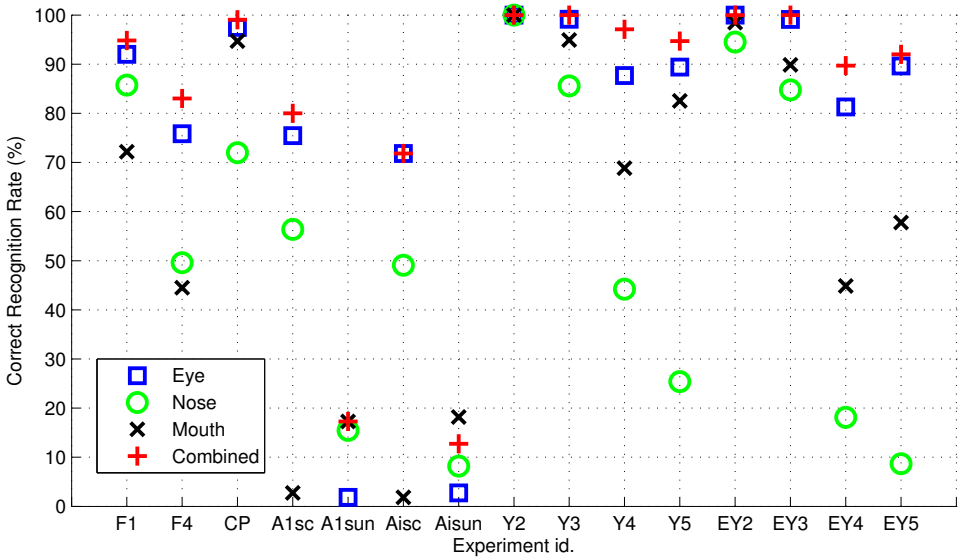



Figure 4.48: Correct identification rates obtained with the *P1* partitioning scheme.

The second partitioning scheme (*P2*) is from [PMS94]. Four salient regions are used for face recognition: left eye, right eye, nose, and mouth. The partitioning on a sample image is shown in Figure 4.49. The results obtained by this partitioning scheme can be seen in Figure 4.50. The outcomes are similar to the ones obtained with the *P1* partitioning scheme. Combined representation achieves the best results. Eye regions have the second place. There is no big difference in left and right eye regions' correct classification rates. The same observations are valid for the nose and mouth regions. The only difference is observed on the *CMUPIE* experiment where the mouth region performs slightly better than the eye regions. It can also be observed that the performance difference between the mouth region and the eye regions is less than the one obtained on the experiments that contain illumination variations with *P1* partitioning scheme. The reason is, the region that contains both of the eyes has more discriminative power than the individual eye regions.

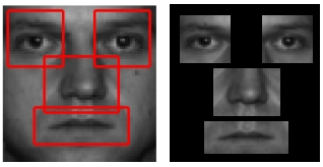


Figure 4.49: Salient regions obtained with the *P2* partitioning scheme.

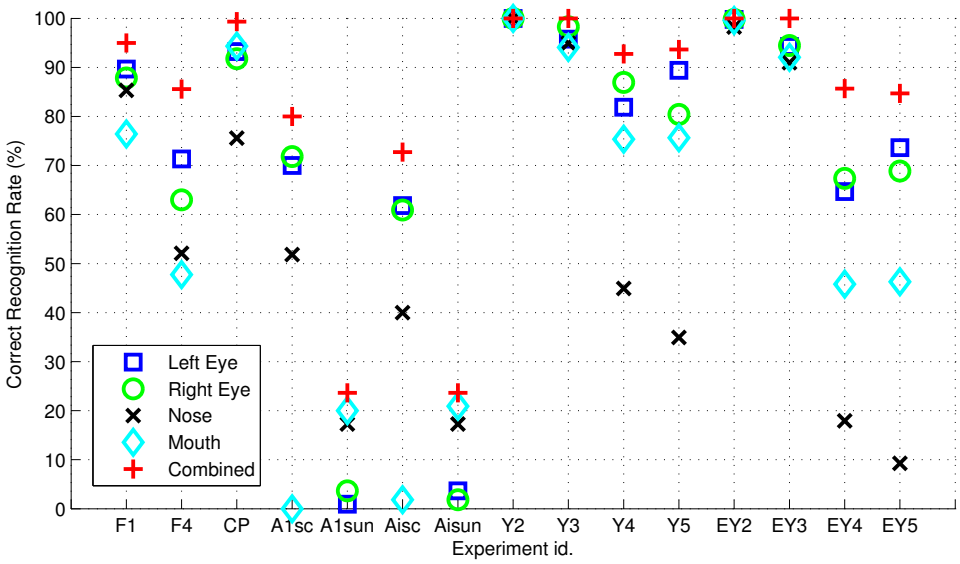



Figure 4.50: Correct identification rates obtained with the *P2* partitioning scheme.

The third partitioning scheme (*P3*) is derived from [LKK05]. Larger left eye and right eye regions that contain partially the nose and some parts below the eyes, and the nose region are the salient regions used in this partitioning scheme. The partitioning on a sample image is shown in Figure 4.51. The correct identification rates that are achieved with the *P3* partitioning scheme is presented in Figure 4.52. Combined representation attains the highest correct recognition rates in most of the experiments. On *AR1sun* and *ARintersun*, nose region achieves the best results, whereas on *AR1scarf*, right eye region outperforms the others. One more time, it has been observed that, except for upper face occlusion, eye regions contain more discriminative power than the nose region.

Figure 4.51: Salient regions obtained with the *P3* partitioning scheme.

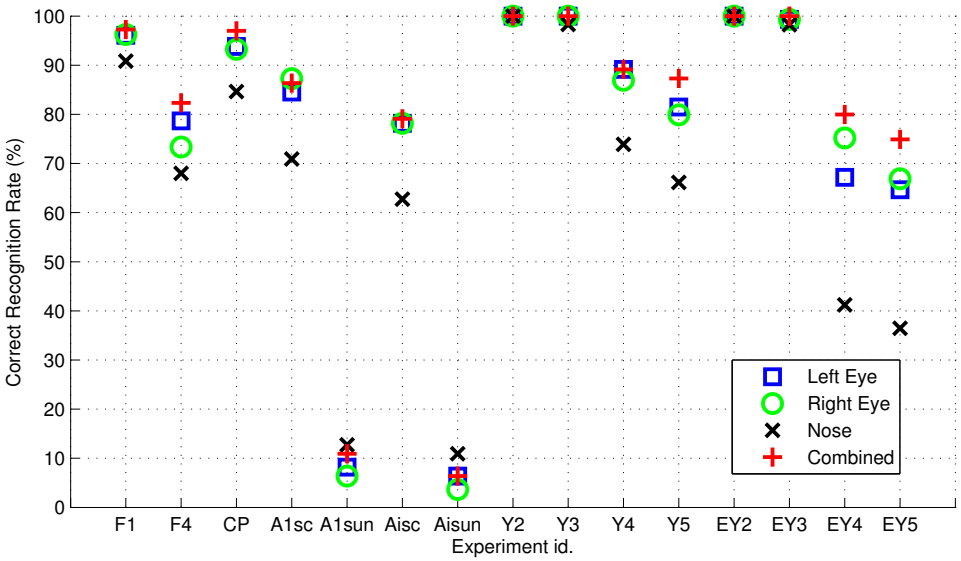



Figure 4.52: Correct identification rates obtained with the *P3* partitioning scheme.

The fourth partitioning scheme (*P4*) is an approximation of the one in [KKHK05]. It has five regions: forehead, left eye, right eye, lower left, and right parts of the face. The partitioning on a sample image is shown in Figure 4.53. Figure 4.54 shows the correct identification rates obtained by the *P4* partitioning scheme. The best performance is always achieved with the combined representation. On the experiments with lower face occlusion, as expected, lower face regions

perform poorly and on the experiments with upper face occlusion, eye regions perform poorly. In most of the cases the forehead region achieves higher correct recognition rates compared to the other salient regions on the experiments that contain large illumination variations, since this region is less affected from the changes in lighting. Both the eye regions and lower facial parts contain partially the nose region which makes them sensitive to the changes in appearance due to cast shadows.

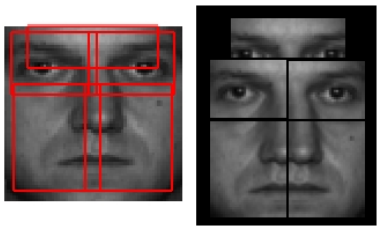


Figure 4.53: Salient regions obtained with the *P4* partitioning scheme.

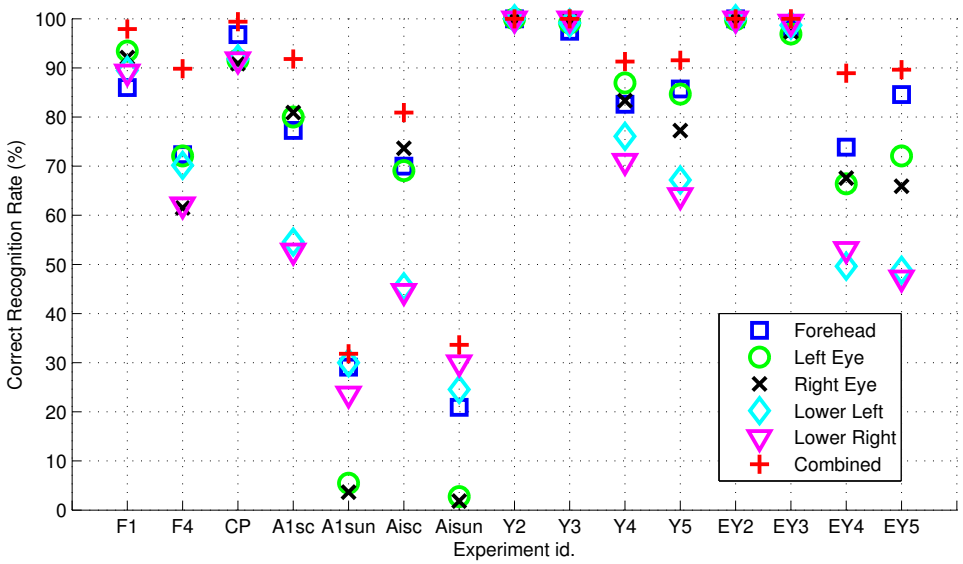



Figure 4.54: Correct identification rates obtained with the *P4* partitioning scheme.

The fifth partitioning scheme (*P5*) is derived from [HSP07]. There are 14 learned components as shown in Figure 4.55. Correct identification rates obtained by the *P5* partitioning scheme is given in Figure 4.56. Most of the time the best performance is achieved with the combined representation. Only on *AR1scarf*, *ARinterscarf* experiments the nose bridge region and on *AR1sun* the right cheek region performs better. Depending on the experimental condition, the performance order of the facial parts changes. Except for the experiments with upper face occlusion, right eye, left eye, right eyebrow, and left eyebrow regions consistently achieve high classification rates.

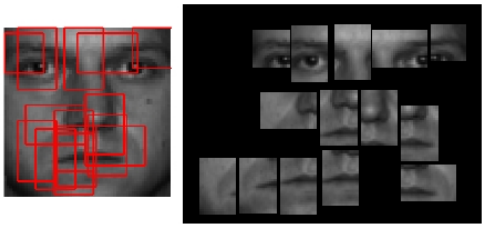


Figure 4.55: Salient regions obtained with the *P5* partitioning scheme.

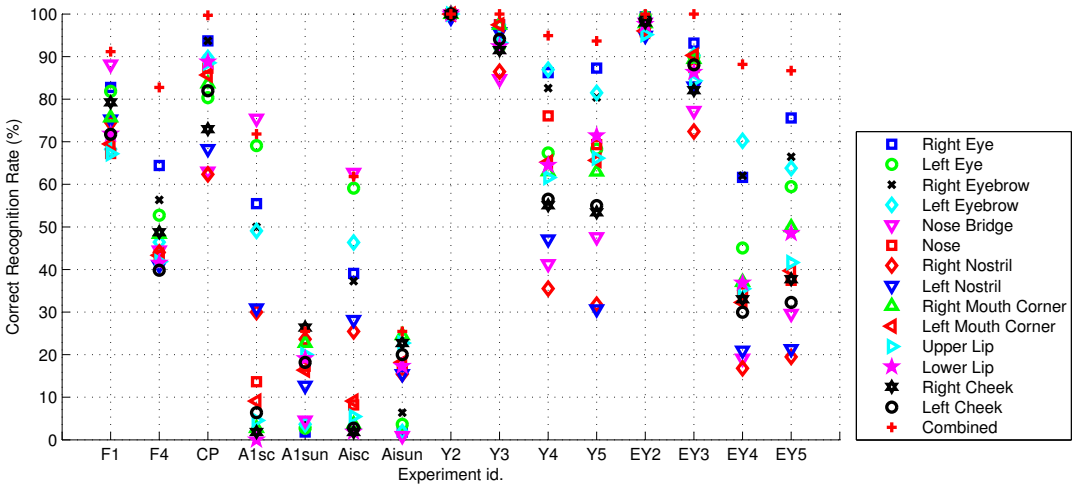



Figure 4.56: Correct identification rates obtained with the *P5* partitioning scheme.

The comparison of the combined representation of different partitioning schemes are shown in Figure 4.57. Generic partitioning is found to be superior to the salient region-based partitioning in most of the cases. For example, on *FRGC4*, the performance is 83.0% with *P1*, 85.6% with *P2*, 82.3% with *P3*, 89.8% with *P4*, 82.8% with *P5*, and 90.8% with generic partitioning. Only on the *Yale4* experiment *P1* partitioning scheme outperforms generic partitioning. However, on *ExtYale4*, which contains *Yale4* as a subset, generic partitioning provides better results. *P4* partitioning scheme also provides consistently high results. On the experiments that contain large illumination variations, *P3* partitioning scheme is found to be the poorest performing one. The reason is that on each part the nose is included to some extend, which makes it sensitive to cast shadows. *P1* performs better than *P2* on these experiments again due to its less sensitivity to the cast shadows. These results indicate that there is no need to detect any salient regions and perform salient region-based partitioning.

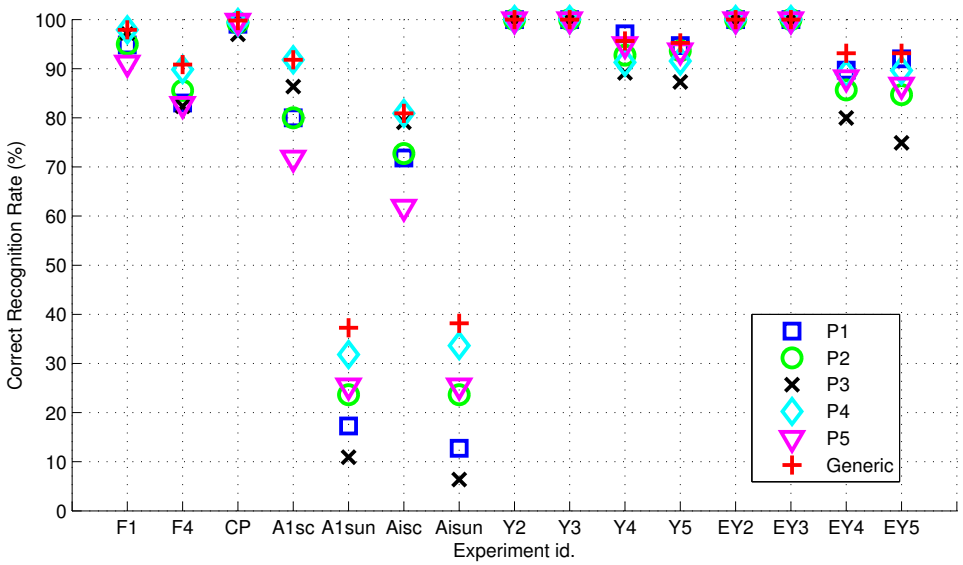



Figure 4.57: Correct identification rates obtained with the combined representation schemes.

## 4.6 Comparison of Local Appearance Representation Methods

In this section, the discrete cosine transform based local appearance representation is compared with different well-known transformation methods that can also be used to represent local regions. These are Karhunen-Loève transform (KLT), Walsh-Hadamard transform (WHT), Fourier transform (FT), and wavelet transform (WT). In the experiments, the input face image is divided into $8 \times 8$ pixel blocks and these basis functions are used to represent these blocks. In KLT, ten-dimensional local feature vectors are used, which are obtained by projecting input face images onto the first ten eigenvectors of the face space that is learned during the training stage. Ten-dimensional feature vectors that are extracted by removing the first DCT coefficient and keeping the following first ten of them are used in DCT. In WHT, the same feature extraction setup as the one used for DCT is utilized. In FT, the magnitudes of the Fourier coefficients are used, which provides 64-dimensional local feature vector. For wavelet transform, the Daubechies 4 wavelet is used, which has been shown to perform better in terms of computation time and recognition performance with respect to the other order Daubechies wavelets, and other well-known wavelets [ES05]. The first order scaling component that provides 16-dimensional representation is used as the feature vector.

Figure 4.58 gives the correct recognition rates obtained with each basis function. The correspondences between the abbreviations on the x-axis and the experiment labels are the same as the ones in Section 4.2. As can be seen in most of the experiments the best results are achieved with the DCT. On the *FRGC1* and *FRGC4*, DCT and WHT are found to be superior to the other representation methods. On *FRGC1*, DCT and WHT reach 97.9% and 98.2%, whereas KLT, FT, and WT obtain 91.2%, 92.6%, and 90.1%, respectively. On *FRGC4*, the correct recognition rate attained by DCT is 90.8%, it is 88.3% by WHT, both being again significantly higher than the ones obtained by KLT —70.8%—, FT —62.9%— and WT —71.0%. This indicates that against expression variations and uncontrolled conditions DCT and WHT provide robust representations. In the case of occlusion, DCT outperforms the other basis functions, except on the *AR1sun* experiment, where WHT, FT, and DCT perform very closely. Except FT, in the case of illumination variations the basis functions are found to work well. DCT-based representation achieves consistently high correct classification rates. For example, on *CMUPIE* experiment, DCT achieves 99.8%, WHT 99.5%, KLT 97.4%, and WT 95.5%. On *Yale4* and *Yale5*, KLT performs slightly better than DCT. The correct recognition rates for KLT are 97.1% and 97.4% and for DCT they are 95.7% and 95.2%, respectively. On the other hand on *ExtYale4*, *ExtYale5* experiments, DCT outperforms KLT. The results are 93.1%, 93.1% for DCT versus 91.0%, 90.6% for KLT. Note that Yale face

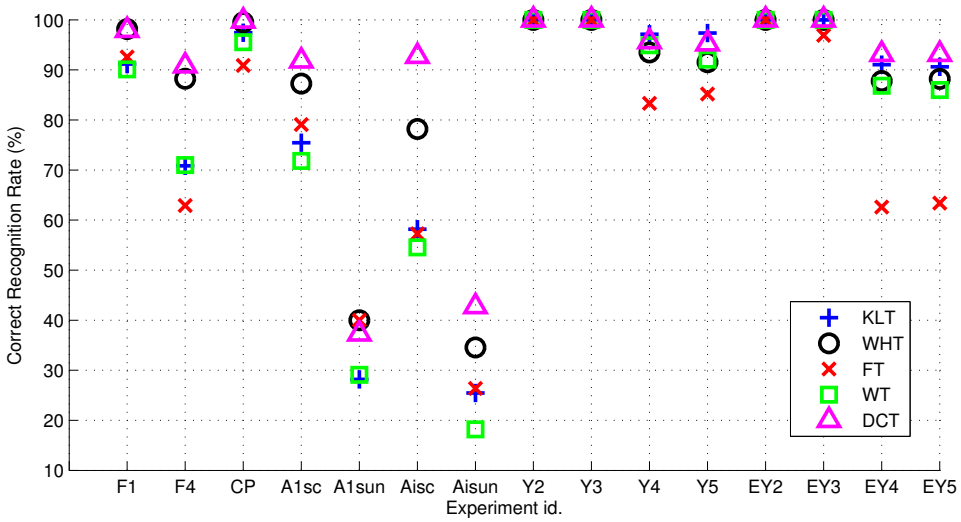



Figure 4.58: Performance comparison of local appearance representation methods.

database B is a subset of the extended Yale face database B and contains less number of subjects. Overall, DCT consistently achieves very high correct classification rates over different conditions, which validates its robust representation capabilities.

## 4.7 Performance Analysis against Compression

The robustness of the proposed local appearance-based approach against JPEG compression is also assessed. Two different experiments are conducted. In one of the experiments, both the training and testing images are compressed with the same quality factor. In the other one, the original, uncompressed face images are used for training, while the testing face images are compressed with varying quality factors. Ten different quality factors are used. Sample compressed images can be seen in Figure 4.59. Especially at low quality factors, compression defects are strongly visible. The quality factors and the corresponding mean compression rates for the face images, which are calculated on the training samples of all the used data sets, are depicted in Figure 4.60. The compression rate is around two when the quality factor is 100. At the quality factor 90, it doubles and becomes four. The compression rate continues to increase with the decreasing quality factor and at the quality factor ten, the compression rate becomes eleven.

Figure 4.59: Sample JPEG compressed face images with different quality factors. Top row, from left to right, with quality factors 10, 20, 30, 40, 50. Bottom row, from left to right, with quality factors 60, 70, 80, 90, 100.

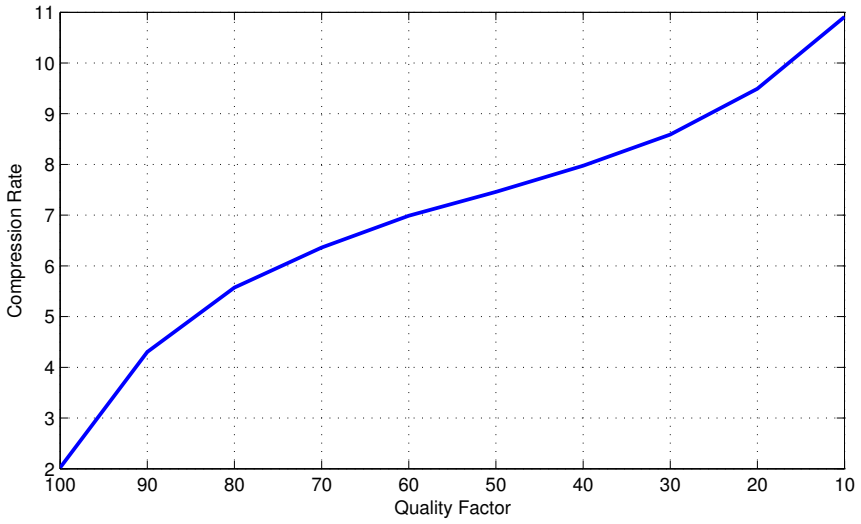



Figure 4.60: Compression rate vs. quality factor.

The results of the face recognition experiments are shown in Figures 4.61-4.66. Ten-dimensional DCT features, that are extracted by omitting the first DCT coefficient and having the following first ten of them, are used in the experiments. The plots are divided into three classes in order to have better visualization: the *FRGC* experiments —*FRGC1*, *FRGC4*—, the *occlusion* experiments —*AR1scarf*, *AR1sun*, *ARinterscarf*, *ARintersun*—, and the *illumination* experiments —*CMUPIE*, *Yale4*, *Yale5*, *ExtYale4*, *ExtYale5*. The results of *Yale2*, *Yale3*, *ExtYale2*, *ExtYale3* experiments are not plotted, since these experiments are relatively easy and 100% correct recognition rate is achieved on these experiments at each quality factor. In the figures, *Org* stands for the original image.

Figures 4.61 and 4.62 plot the correct recognition rates obtained on the *FRGC* experiments. Figure 4.61 corresponds to the matched case where both the training and testing face images are compressed, whereas Figure 4.62 corresponds to the unmatched case where only the testing face images are compressed. As can be observed, on the *FRGC1* experiment, both at the matched and unmatched case, the recognition rates remain stable till the quality factor is 30. It decreases after this point. The results at matched and unmatched conditions are similar, except with the quality factor ten, at which unmatched condition is worse. The recognition rates on the *FRGC4* experiments show slight decrease till the quality factor is 40. A high drop in the performance is observed at the quality factor 10. The unmatched condition's correct recognition rates are better than the matched condition's at lower quality factors.

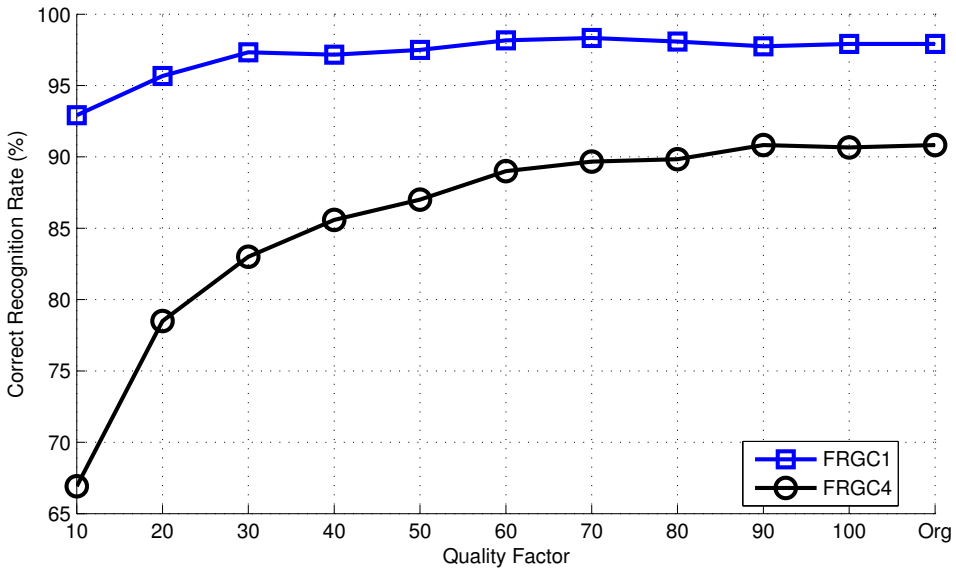



Figure 4.61: Correct recognition rates obtained with respect to different quality factors on the FRGC experiments. Training and testing face images are compressed with the same quality factor.

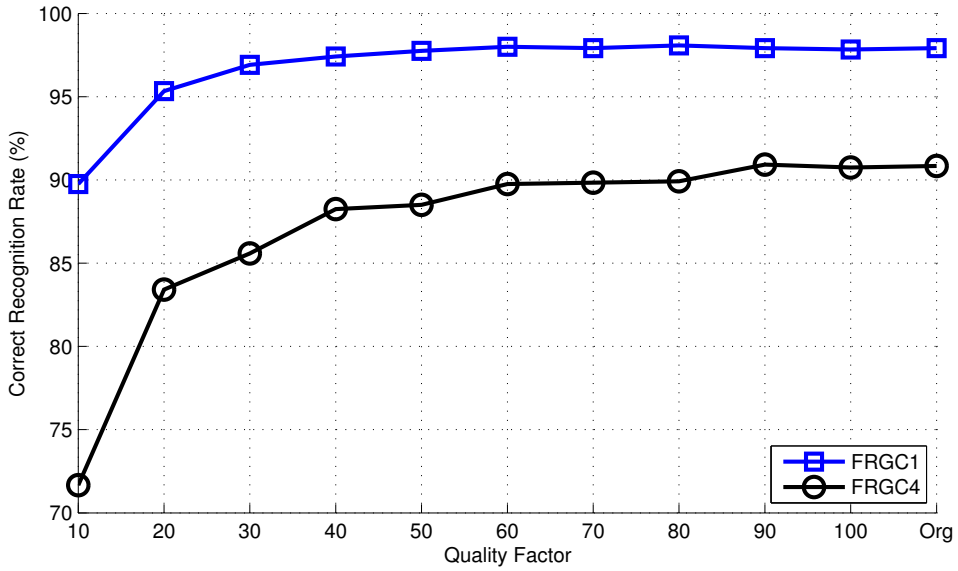



Figure 4.62: Correct recognition rates obtained with respect to different quality factors on the FRGC experiments. Original training images are used. Only the testing face images are compressed.

The correct recognition rates obtained on *occlusion* experiments can be seen in Figures 4.63 and 4.64. The correct recognition rates drop at very low quality factors. The performance is stable at high quality factors. There is no significant performance difference between the matched and unmatched cases at high quality factors. The biggest differences are observed on *AR1scarf* experiment at the quality factor 10 and on *ARinterscarf* experiment at the quality factor 30. Except these points, the results are similar.

Figures 4.65 and 4.66 show the results on the *illumination* experiments. On *CMUPIE*, *Yale4*, and *ExtYale4* experiments, the observations are similar to the ones obtained on the *FRGC* and *occlusion* experiments, that is, the correct recognition rates remain relatively constant at high quality factors and they decrease at low quality factors. However, on the experiments *Yale5* and *ExtYale5*, where strong illumination variations exist, the performance decreases after the quality factor 100. This indicates the importance of high frequency components in doing identification under strong illumination variations, which are eliminated during compression. The unmatched cases are found to perform better than the matched cases on the illumination experiments.

As a summary, it has been shown that the proposed approach is robust against compression both in the case of matched and unmatched training testing combinations. The performance remains stable till very low quality factors. The highest drop in the correct recognition rates occur when the quality factor becomes ten. Nevertheless, in the situation of strong illumination variations, the

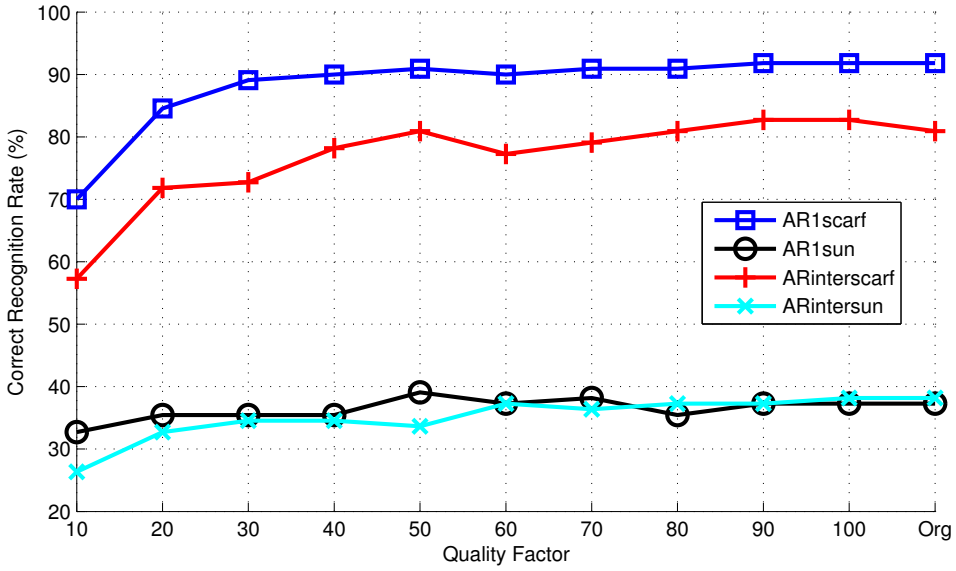



Figure 4.63: Correct recognition rates obtained with respect to different quality factors on the occlusion experiments. Training and testing face images are compressed with the same quality factor.

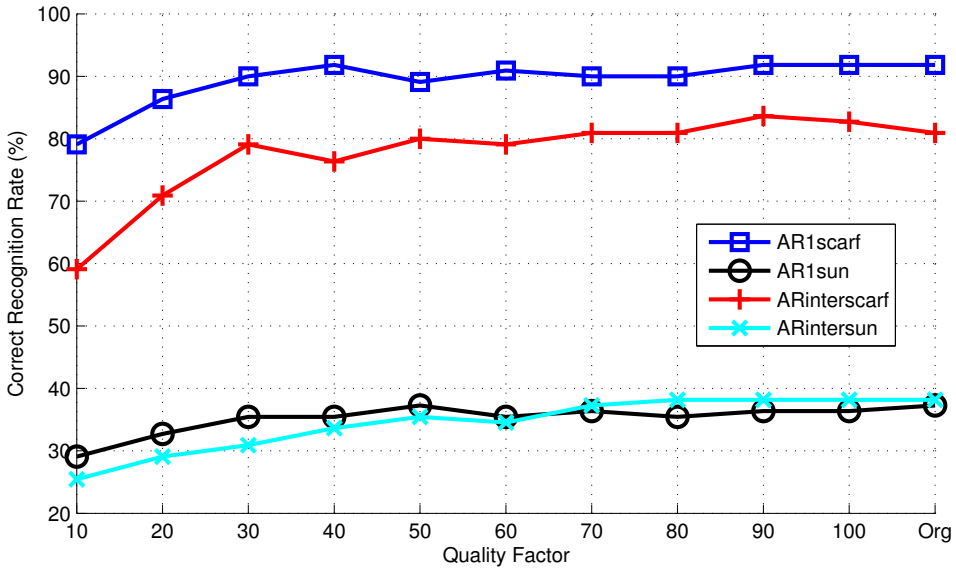



Figure 4.64: Correct recognition rates obtained with respect to different quality factors on the occlusion experiments. Original training images are used. Only the testing face images are compressed.

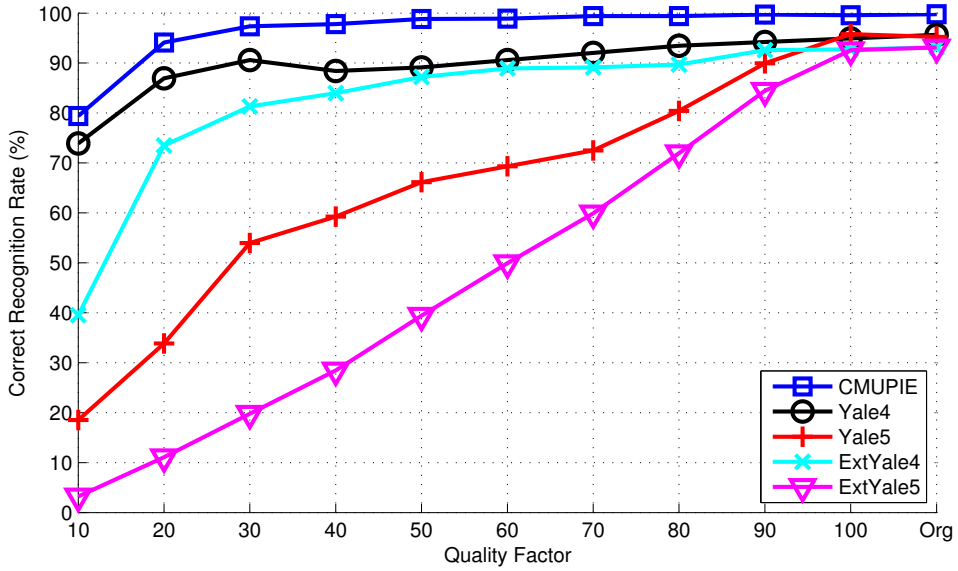



Figure 4.65: Correct recognition rates obtained with respect to different quality factors on the illumination experiments. Training and testing face images are compressed with the same quality factor.

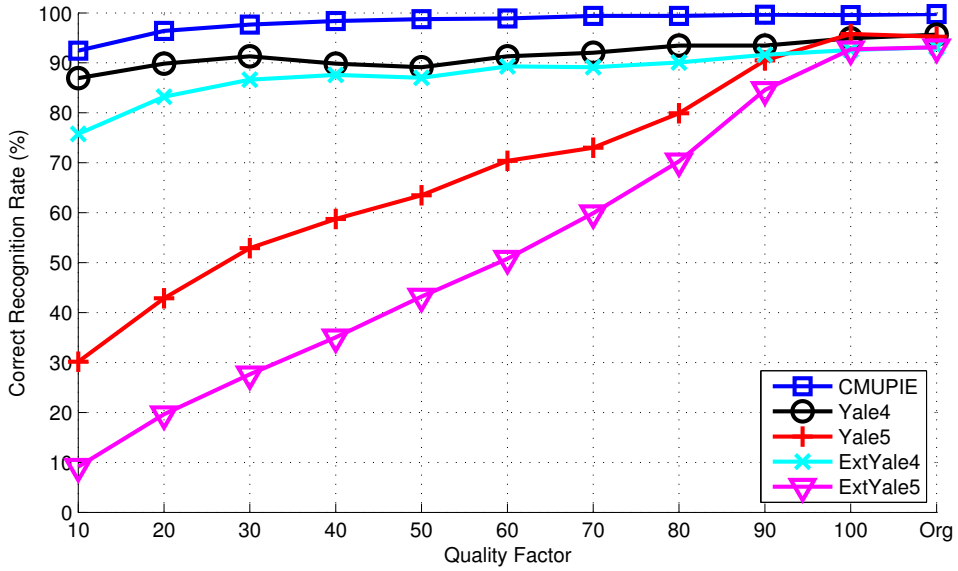



Figure 4.66: Correct recognition rates obtained with respect to different quality factors on the illumination experiments. Original training images are used. Only the testing face images are compressed.

decrease in the performance starts after the quality factor 100, which implies that high frequency content plays a crucial role in the classification of such face images.

# 5 Robust Face Recognition

This chapter presents two extensions to the proposed local appearance-based face recognition algorithm. These extensions consist of an automatic face registration approach that provides robustness against registration errors caused by imprecise facial feature localization and an automatic frequency band selection approach that provides robustness against facial appearance variations caused by changing illumination conditions. The chapter also contains a detailed comparison of the proposed approach with well-known face recognition algorithms.

## 5.1 Performance Analysis against Registration Errors

Registration is one of the important factors that affect pattern recognition systems' performance. For example, in [WHH+89], a time-delay neural network (TDNN) structure is proposed to provide shift invariance. This way superior phoneme recognition performance is achieved. Also, for handwriting recognition a multi-state time-delay neural network (MS-TDNN), which is an extension of the TDNN, is utilized to attain high recognition rates [MB94]. Similarly, registration is an important component of a face recognition system. When an appearance-based approach is to be utilized, precise alignment is crucial in order to achieve high correct recognition rates [LP02, RSPP06]. To analyze the robustness of the local appearance-based face recognition approach against registration errors, in the experiments, training images are registered with respect to the manually annotated eye center positions, whereas equally distributed random noise, ranging from 1% to 15% of the distance between the eyes, is added to the manually annotated eye center positions of the test images. For each noise level a separate classification is done. This way, translation, rotation, and scale variations are provided at different strengths and the local appearance-based approach's performance is assessed at each variation-level. A face image registered with the manually generated labels and sample misaligned face images can be seen from Figure 5.1. The top-left image corresponds to the original registered image. The amount of the added noise increases from left to right and top to bottom. As the modification in the labels increases, the change in registered

images becomes more visible. Especially, when the added noise is too high, it is obvious that there can not be a meaningful comparison between a well registered training image and a misaligned test image when an appearance-based approach is applied, since the appearances are no longer similar.

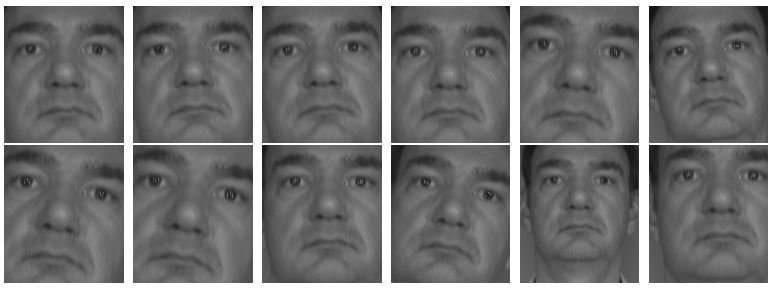


Figure 5.1: Sample misaligned face images with different amount of noise added to the manually labeled eye center positions. First row, from left to right, no modification is induced and induced modifications of 1, 2, 3, 4, 5% of interocular distance. Second row, from left to right, induced modifications of 6, 7, 8, 9, 10, 15% of interocular distance.

The results of the experiments are plotted in Figure 5.2. In the experiments, ten-dimensional feature vectors that are extracted by removing the first DCT coefficient and keeping the following first ten of them are used. Experiments are performed only on the AR and FRGC data sets, where both the original images and corresponding manually annotated eye center labels are available. The CMUPIE, Yale, and ExtYale setups are excluded from the experiment, since only the already aligned images are available from these data sets. However, the outcomes of these experiment are also valid for these setups, since misalignment is a source of variation that is decoupled from the facial appearance changes. The obtained results also justify this claim. No matter what the facial appearance variation in the training-testing combination is, the same performance trends are observed. The performance deteriorates with increasing amount of added noise. For example, in the *FRGC1* experiment, the correct classification rate achieved on the face images that are registered with the original labels is 97.9%, it drops to 96.6% when 1% of the interocular distance is added as a noise to the labels. The performance continues to decrease with increasing noise. It becomes 92.7% and 80.5% with addition of 2% and 3% of the interocular distance as a noise, respectively. At the noise level of 9% of the interocular distance, the correct recognition rate drops below 10%. When 15% of interocular distance is added to the labels as a noise, the attained performance is only 1.1%.

These results validate that the registration step plays a crucial role in appearance-based face recognition. In order to have high correct recognition rates, the registration should be done as precise as possible, otherwise an appearance-based approach is destined to perform poorly. However, a precise registration is not

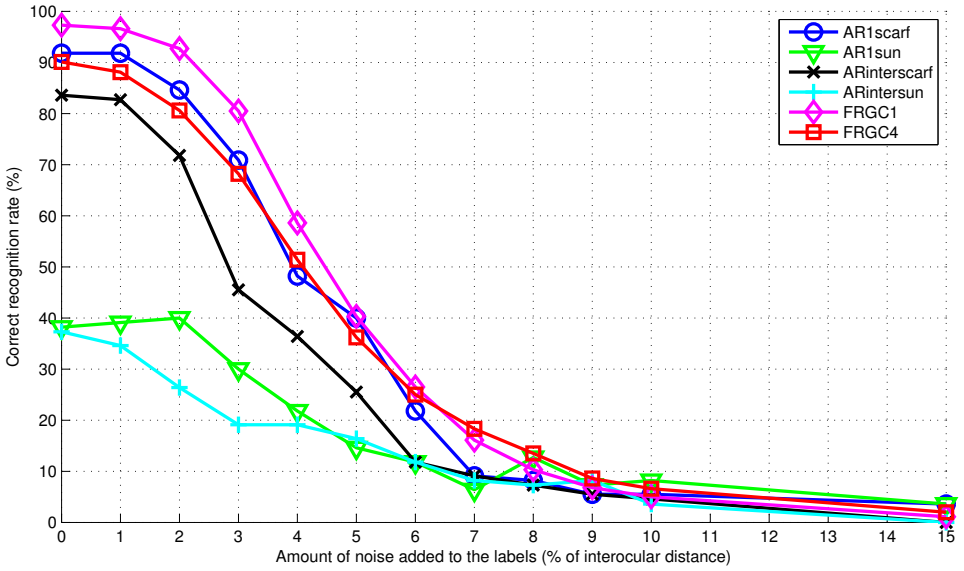



Figure 5.2: Performance of the local appearance-based face recognition approach with respect to registration errors.

always possible due to imperfections in feature localization. On account of this reason, one should take this problem into consideration while designing a face recognition algorithm. The problem of imprecise registration can also be seen from Figures 5.3, 5.4 and 5.5. In order to analyze the effect of having imperfections in eye center localization, 1 to 10 pixels variation is added to the manual eye center annotations of a subset of training samples from the AR face database [MB98]. The face images are aligned according to these newly created labels. This way, positive and negative x-shift and y-shift; smaller and larger scale; clockwise and counterclockwise rotation is provided. From the registered face images, ten-dimensional DCT-based local feature vectors are extracted. The L1 norm distance is calculated between the feature vector of the face image registered with the manual label and the feature vector of the same face image registered with the modified label. The obtained averaged distances are plotted in Figure 5.3. It is observed that the distance value increases with the increasing modification. Another interesting point that can be derived from the figure is that the scale and rotation causes larger distances than shifts. In Figure 5.4, the distribution of the distances between the best two matches for correct and false classifications are given. These values are obtained by classifying the images that are annotated as "neutral expression" in the second recording session of the AR face database. The training face images are from the first recording session and they have the same annotation. As can be seen from the figure, the false classifications occur when the distance between the best two matches is less than twenty. This value is also the one obtained when a picture is shifted by one pixel and the distance value is calculated between the feature vectors of

the original image and the shifted image. So, obviously, it is not reliable to determine the identity if the distance between the best two matches are very close. The distribution of the distances to the closest matches are given in Figure 5.5. As expected, the false classifications occur when the distance is high. Considering the distance values due to mislocalization in Figure 5.3, the sensitivity of the classification results to the registration errors is apparent.

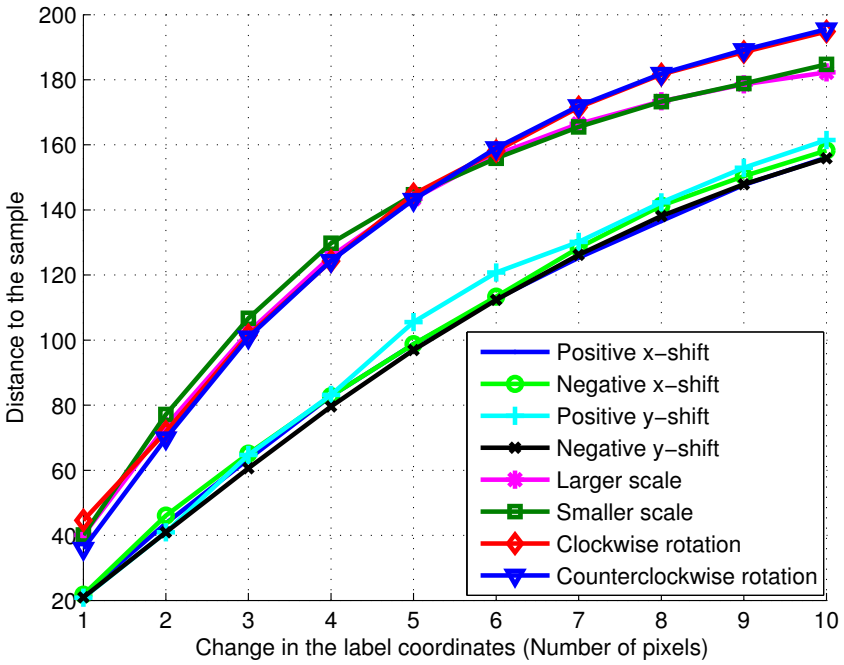



Figure 5.3: Obtained distance values with respect to the change in label coordinates. The distance is calculated between the feature vector of the face image which is registered with the original label and the feature vector of the same face image registered with the modified label.

## 5.2 Face Registration by Minimizing the Closest Classification Distance

In this section, a face registration approach, in which alignment is done by minimizing the closest distance at the classification step, is presented. This method eliminates the need of a feature localization step that exists in traditional face recognition systems and formulates alignment as an optimization process during classification. In other words, instead of performing a separate feature localization step and localizing facial features according to some type of feature matching scores, in the proposed method, alignment is done in such a way that directly the classification score is optimized. Moreover, it is shown

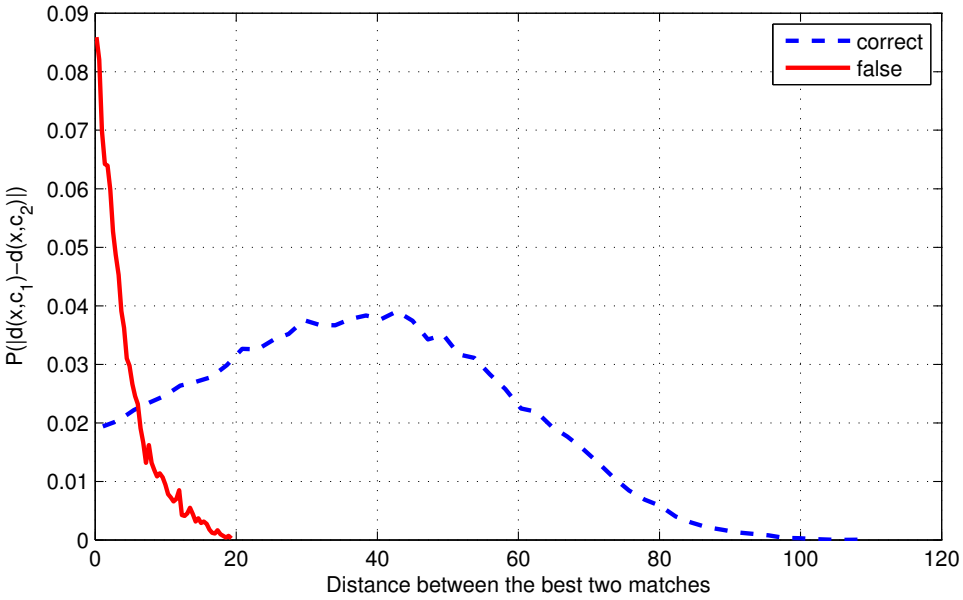



Figure 5.4: Distribution of the distances between the best two matches for correct and false classifications.

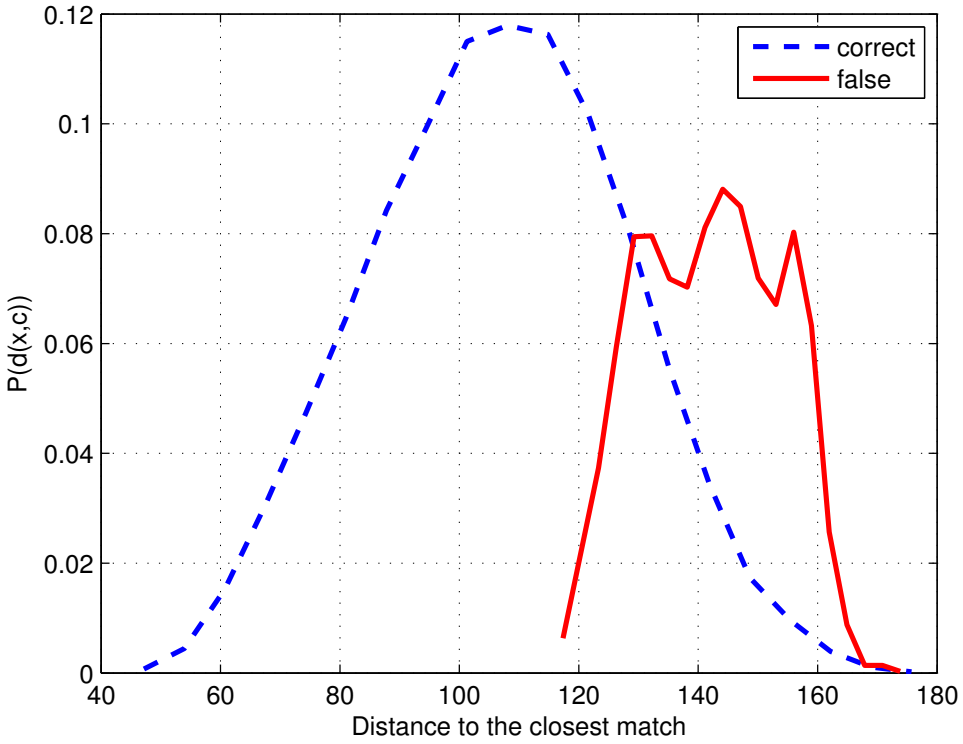



Figure 5.5: Distribution of the distances to the closest matches for correct and false classifications.

that the proposed method can also be used in face recognition systems that do registration via feature localization in order to combat against problems due to erroneous feature localization.

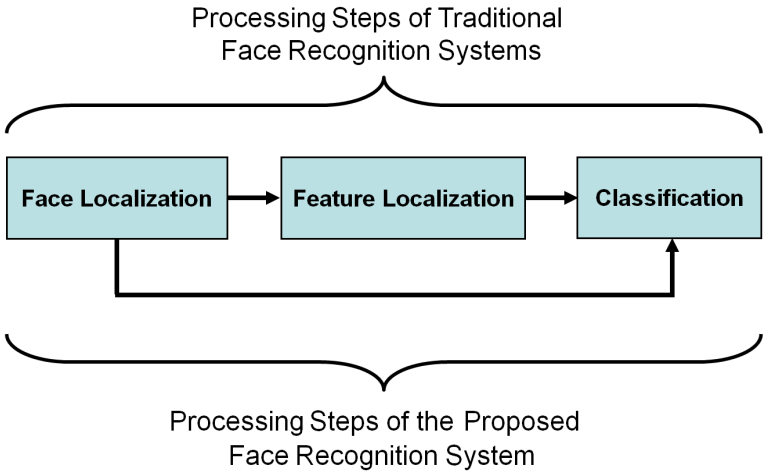



Figure 5.6: Traditional face recognition systems versus the proposed face recognition system.

Since all humans have the same facial feature configuration, once the face is located it is easy to roughly estimate the locations of the facial features. In order to show this, the relative eye center positions with respect to the center of the bounding boxes of faces in the AR [MB98] and the FRGC [PFS+05] databases is plotted in Figure 5.7. For this analysis, the training samples of the AR, FRGC1 and FRGC4 data sets are used. The face bounding boxes are detected with a generic face detector [Int08]. The center of the face rectangle is subtracted from the manually labeled eye center locations and the obtained distances are scaled according to the width of the face bounding box. As can be observed from Figure 5.7, despite using different databases and a generic automatic face detector, the normalized relative eye center positions are densely located. The median values of these eye center positions are calculated both for the left and right eye to produce an eye center hypothesis with respect to the automatically located face rectangle. The obtained eye center position, in pixels, for the left eye is $(-11.5, 6.5)$, and for the right eye it is $(11.5, 6.5)$. In order to contain all the deviations from these calculated values due to variations in feature positions across different identities and variations in the bounding boxes generated by the automatic face detector, a window size of $11 \times 11$ pixels is determined around the eye centers. It can be seen in Figure 5.7 that a window size of $9 \times 9$ pixels would suffice to cover all the points. Furthermore, with a fine tuned face detector and improved face segmentation accuracy, this region can become even smaller. However, for the sake of having a generic approach without relying on the accuracy of the face detector and in order to tolerate in-plane and out-of plane rotations up to some extend, a larger window size is selected. These

regions are used as search regions to determine the best matching eye centers between the test image and training samples. This fine eye center localization is integrated to the classification step, bypassing a separate feature detection process. In the proposed approach, the eye center positions of the test image are determined in such a way that the classification distance of the test image to a training sample is minimized. This way, for each training sample, separate eye center positions are determined for the test face image, which lead to the aligned test face image that has the minimum distance to the training image. In this approach, inconsistencies across manual eye center labels of the training images are also handled, since, as already mentioned, for each training sample a separate eye localization is performed by optimizing the classification distance. This is different from the traditional face recognition approaches where only one eye center estimate is used to match the test image against all the training samples ignoring the possibility of having inconsistencies among the manually labeled eye center positions of the training samples. In order to save processing time during testing, in the implementation, the generation of aligned face images using the eye center coordinates within the determined region is performed offline, on the training side. That is, a window size of $11 \times 11$ pixels around the manually labeled eye center position is used as possible eye center region and additional aligned training face images are generated with respect to the possible eye center combinations from this region. The algorithm can be summarized as follows:

- Training:
  - (i) Have all the eye center location combinations between the left and right eyes within the $11 \times 11$ pixels window around the manually labeled eye center positions,
  - (ii) Generate aligned face images according to these eye center positions,
  - (iii) Extract a feature vector from each aligned face image.
- Testing:
  - (i) Do face localization and estimate the eye center positions by adding $(-11.5, 6.5)$ for the left eye center and $(11.5, 6.5)$ for the right eye center to the center of the scaled face bounding box,
  - (ii) Align test face image with respect to the estimated eye center position,
  - (iii) Compare the aligned test face image with all the aligned face images generated from a training sample,
  - (iv) Find the aligned face image from the training sample that provides the minimum classification distance,
  - (v) Perform steps (iii) and (iv) for each training sample,

(vi) Find the training sample that provides minimum classification distance,

(vii) Assign the identity of the best matching training sample to the test image.

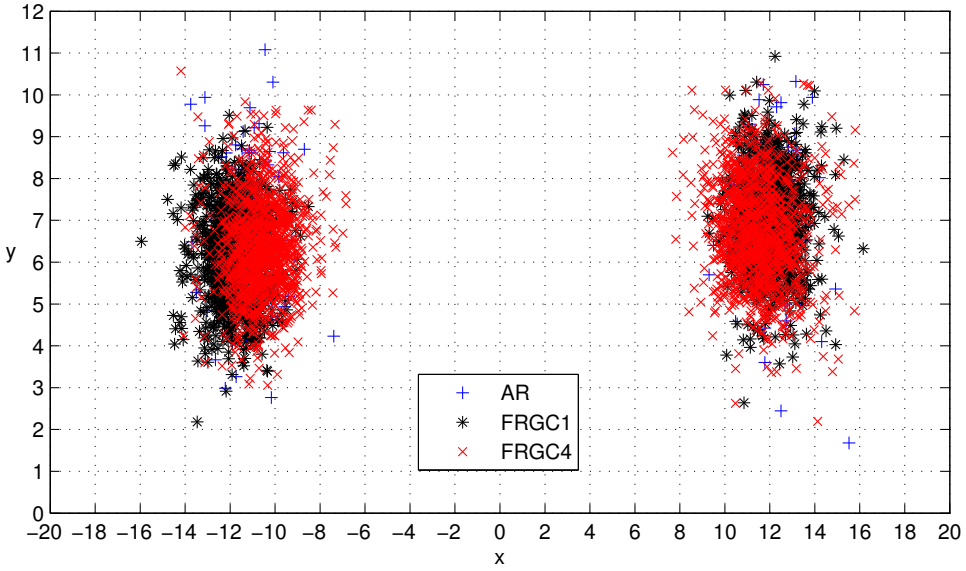
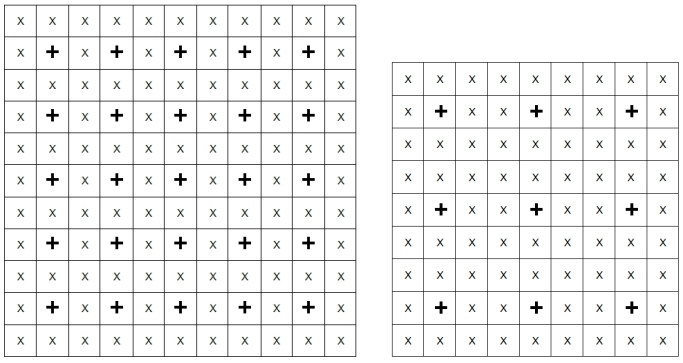



Figure 5.7: Distribution of the eye centers with respect to the center of the face bounding box.

One can notice that having a window size of $W \times W$ pixels around the eye centers causes $W^4$ eye center coordinate combinations. In the case of a window size of $W = 11$, the number of combinations is 14641, which means 14641 times as many feature vector comparisons have to be done. However, the amount of comparisons can be significantly decreased by utilizing a hierarchical search scheme. Instead of having all the position combinations within $\pm 5$ pixels, first the combinations at $\pm 2$ pixels locations can be searched. This way, the number of eye center position combinations at the first search step becomes 625. After determining the combination that provides the minimum classification distance at the first step, at the second step, the search is done $\pm 1$ pixel around the determined eye center positions from the first search step. Thus, at the second step $W$ becomes 3, providing 81 combinations. Since the classification is already done with the face image aligned using the determined eye center positions from the first search step, only 80 additional comparisons are needed to be done at the second step, making overall a total of 705 comparisons per training sample. This search pattern is depicted in Figure 5.8(a). The '+' shows the search locations at the first step, whereas 'x' shows the search locations at the second step. The computational load can be further decreased by having a window size of $W = 9$, and performing the search first at the combinations of $\pm 3$ pixels and then at

the combinations of ±1 pixel around the determined eye center locations from the first step, which makes in total 161 comparisons per training sample. The search pattern for $W = 9$ is shown in Figure 5.8(b).

It should be noted that, although on one hand the amount of computation increases due to higher number of feature comparisons, on the other hand, due to omitting a separate feature detection step, some amount of computation is saved. Moreover, feature comparison consists of only a subtraction operation which can be done very fast.



(a) (b)

Figure 5.8: Search pattern for (a) $W = 11$, (b) $W = 9$. The '+' shows the search locations at the first step, whereas 'x' shows the search locations at the second step.

The experimental results of the proposed automatic face recognition system which does not need facial feature localization are presented in Table 5.1. In the experiments ten-dimensional local feature vectors that are extracted from each $8 \times 8$ pixels block by removing the first DCT coefficient and keeping the following first ten of them are used. The selected coefficients are divided by their standard deviations and normalized to unit norm. As in the robustness analysis experiments against registration errors in Section 5.1, experiments are performed only on the AR and FRGC data sets, where both the original images and corresponding manual eye center labels are available. The achieved correct recognition rates on the *AR* experiments are very low when the training images are aligned using only the manually annotated eye center labels and the test face images are detected with a generic automatic face detector [Int08] and aligned using the estimated eye center positions. The results improve significantly when all the eye center combinations within the determined eye center regions are used to align the training face images. For example, the correct recognition rate increases from 32.7% to 97.3% in the *AR1scarf* experiment and from 20% to 95.5% in the *AR1sun* experiment. As can be seen from the table, there is no significant performance difference between brute-force search and hierarchical search when the window size is $W = 11$. However, the results

are lower, when the window size is $W = 9$, especially on the expeirments that contain occlusion. The main reason is, due to occlusion, face detection quality is very low in these cases, which in turn causes poor eye center position estimates that cannot be covered with a smaller window size. Another interesting observation that can be derived from the table is that, very high correct classification rates are obtained against occlusion problem. Especially eye region occlusion is known to be one of the biggest challenges in face recognition. The obtained results imply that mainly the erroneous feature localization, thus imprecise face alignment causes the poor performance in the case of eye region occlusion. It is also intriguing to observe that the correct recognition rate obtained in the *ARinterscarf* experiment is lower than the one obtained in the *ARintersun* experiments, although lower face occlusion is known to be an easier problem than the upper face occlusion. The reason can be the textured surface of the scarfs which might affect the classification decision more than the black sunglasses. It should be also considered that, as already shown, the main issue with the upper face occlusion is the misalignment and once it is handled, very high performance can be reached. Note that the achieved correct identification rates are significantly higher than the ones presented in the literature [GSC01, Mar02, TCZZ05, PLL05, FSL06, WGYM07, JM08].

The results obtained on the FRGC experiments are similar to the ones attained on the AR experiments. There is a big improvement in the correct recognition rates when the proposed approach is used. Again, hierarchical search performs as good as brute force search.

| | Est. labels only | Brute-force search | Hierar. search $W = 11$ | Hierar. search $W = 9$ |
|---|---|---|---|---|
| AR1scarf | 32.7% | 97.3% | 97.3% | 94.6% |
| AR1sun | 20.0% | 95.5% | 95.5% | 90.9% |
| ARinterscarf | 40.0% | 90.0% | 89.1% | 84.6% |
| ARintersun | 13.6% | 93.6% | 93.6% | 79.1% |
| FRGC1 | 66.4% | 98.2% | 98.2% | 98.3% |
| FRGC4 | 43.8% | 93.8% | 93.4% | 91.8% |

Table 5.1: Correct recognition rates obtained on the AR and FRGC face data sets. The results in the first column are obtained using just the manually labeled eye center positions to align a training sample. The ones in the other columns are achieved using all the eye center position combinations to align the training sample. Either brute force or hierarchical search is conducted to find the best matching aligned training sample.

## 5.3 Robust Face Recognition against Registration Errors

Although the proposed system is free from feature localization, it is still possible to integrate a feature detector to it. In this case, instead of using the estimated eye center positions, the output of the facial feature detector is used as the initial point of the optimization process. In the scaled faces, the distance between the eyes is 27 pixels. Conducting search within $\pm 5$ pixels in the case of window size $W = 11$ provides insensitivity to the localization errors of up to 18% of the interocular distance, whereas conducting search within $\pm 4$ pixels in the case of window size $W = 9$ provides insensitivity to the localization errors of up to 14% of the interocular distance.

In this section, to analyze the contribution of the proposed method to the performance of the face recognition system that uses a separate feature detection step, several experiments are conducted by using the manually labeled eye center positions of the test face image as the localization output of an automatic feature detector and adding different levels of noise to them to imitate the registration errors. In the experiments, the same experimental setup, as the one in experiments against registration errors in Section 5.1, is used. That is, training images are registered with respect to the manually annotated eye center labels, whereas random noise, ranging from 1% to 15% of the distance between the eyes, are added to the manually annotated eye center labels of the test images. For each noise level a separate classification is done. A face image registered with the manually generated labels and sample misaligned face images can be seen from Figure 5.1. Ten-dimensional local feature vectors that are extracted from each $8 \times 8$ pixels block by removing the first DCT coefficient and keeping the following first ten of them are used. The selected coefficients are divided by their standard deviations and normalized to unit norm. Correct recognition rates obtained by brute force and hierarchical search with window size $W = 11$ are plotted. In these experiments, hierarchical search with window size $W = 9$ has been also used, which has been found to perform similarly to the hierarchical search with window size $W = 11$, except at the noise level of 15% of the distance between the eyes, where a slight decrease in performance has been observed.

In Figures 5.9, 5.10 and 5.11 the correct recognition rates are plotted with respect to varying localization errors added to the manually annotated eye center positions. The results of the experiments in Section 5.1 are also plotted in the figures to visualize the improvement provided by the proposed method. In Figure 5.9, the correct recognition rates obtained on the *ARscarf* experiments are depicted. The *AR1scarf* and the *ARinterscarf* plots correspond to the results attained using just the provided eye center positions. The plots with the suffix "bf" correspond to using multiple eye center position combinations and doing

brute force search, and the ones with the suffix "hs" correspond to using multiple eye center position combinations and doing hierarchical search. As can be observed, even without adding any errors and using the provided manually labeled eye center positions (the point "0" in the x-axis), the proposed method improves the correct recognition rates significantly. On *AR1scarf*, the performance increases from 91.8% to 97.3% and on *ARinterscarf*, it increases from 83.6% to 90.0%. Both brute force and hierarchical search provide the same results at this level. As can be seen, as the error increases, the performance deteriorates, however, with the proposed method the achieved correct recognition rates stay consistent with respect to different error levels. There is no significant performance difference observed between doing brute force search and hierarchical search.

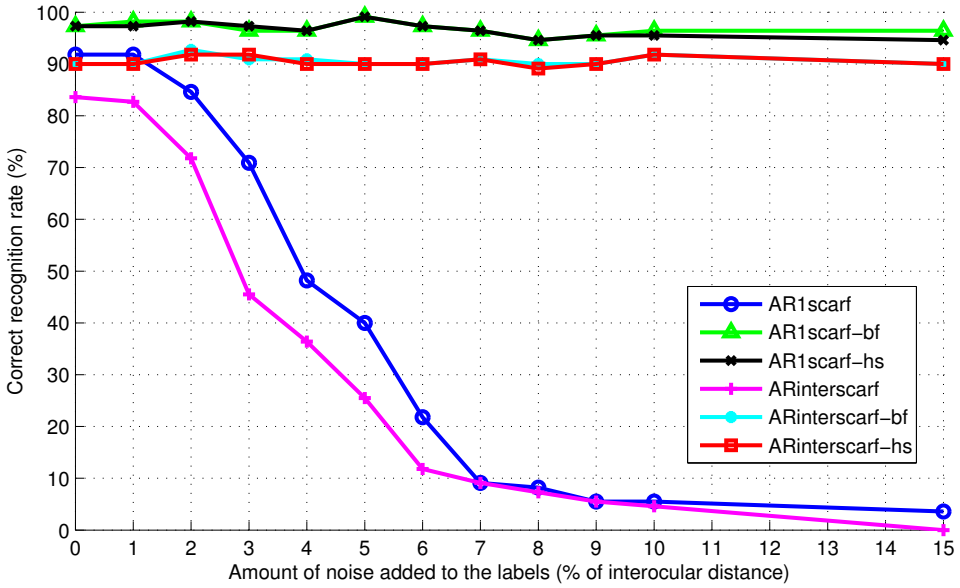



Figure 5.9: Correct recognition rates obtained on the AR scarf experiments with respect to localization errors. The plots without suffix correspond to using only provided labels, the ones with the suffix "bf" correspond to using multiple eye center position combinations and doing brute force search, and the ones with the suffix "hs" correspond to using multiple eye center position combinations and doing hierarchical search. The proposed approach provides stable results over localization errors.

The correct recognition rates obtained on the AR sun experiments are plotted in Figure 5.10. The improvement in correct classification rates provided by the proposed approach is even more remarkable in these experiments. The correct recognition rate is 38.2% on the *AR1sun* and 37.3% on the *ARintersun* experiments when only the manually labeled eye center positions are used. They become 97.3% and 95.5%, respectively, when all the eye center position com-

binations are utilized within the determined eye region. This outcome is not surprising, since it is not possible to precisely label the actual eye centers even manually due to occlusion caused by sunglasses which leads to misalignment. Again, correct recognition rates remain stable with respect to varying error levels. However, this time, it decreases when 15% of the interocular distance is added to the manually annotated eye center labels as noise. As stated before, in these experiments, these labels are assumed to be precise and the errors are induced to these labels in order to imitate the localization errors. Nevertheless, these labels are not precise in the case of wearing sunglasses. Because of this reason, when the eye center positions are modified by 15% of the interocular distance, depending on how precise the manually generated label is, on some test images the modification could be higher than 18% of the interocular distance with respect to the actual eye center positions, which is the upper limit of localization error that the proposed system can tolerate.

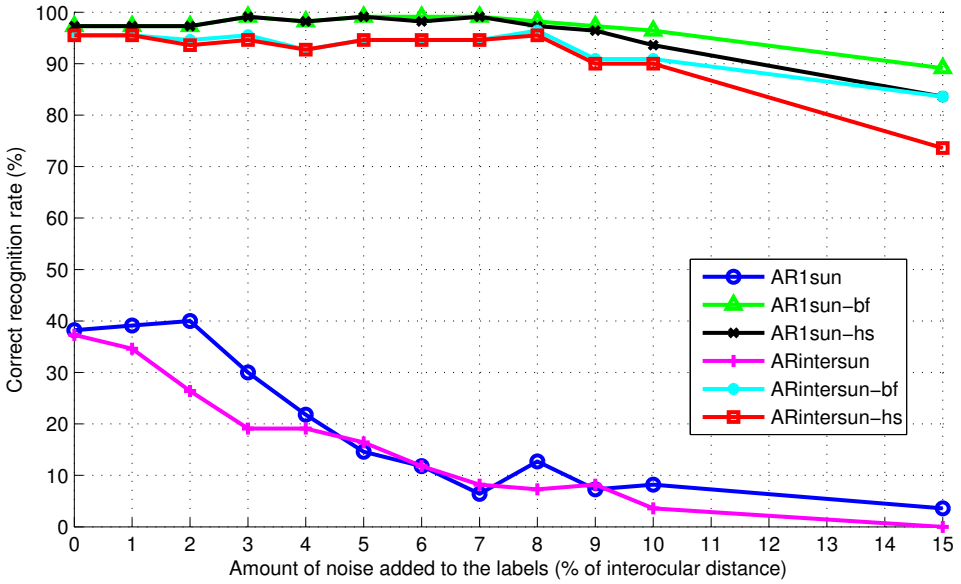



Figure 5.10: Correct recognition rates obtained on the AR sun experiments with respect to localization errors. The plots without suffix correspond to using only provided labels, the ones with the suffix "bf" correspond to using multiple eye center position combinations and doing brute force search, and the ones with the suffix "hs" correspond to using multiple eye center position combinations and doing hierarchical search. The proposed approach provides stable results over localization errors.

In Figure 5.11, the results of the FRGC experiments can be seen. One more time, it is shown that the proposed approach provides robustness against localization errors, having consistent correct recognition rates over varying localization errors. The high resolution face images in the FRGC data sets lead better manual

labeling of the eye center positions. Therefore, on the FRGC experiments, the increase in the performance attained with the proposed method, at the condition where no errors are added to the manually annotated eye center positions, is less than the ones obtained on the AR experiments. It is 1.5% on the *FRGC1* where the face resolution is the highest, and 3.7% on the *FRGC4* experiment, while 5.5%, 59.1%, 6.4% and 58.2% increases are achieved on the *AR1scarf*, *AR1sun*, *ARinterscarf* and *ARintersun* experiments, respectively.

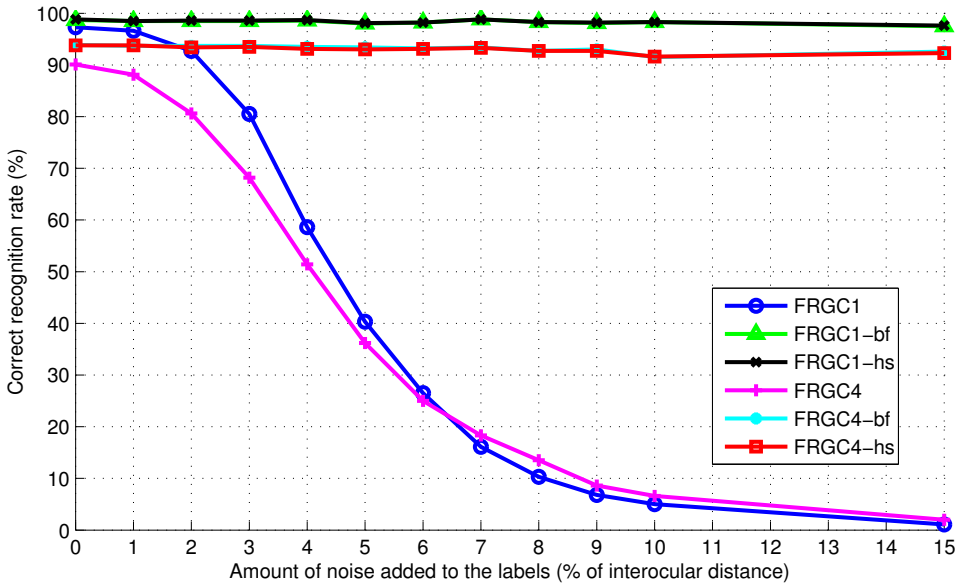



Figure 5.11: Correct recognition rates obtained on the FRGC experiments with respect to localization errors. The plots without suffix correspond to using only provided labels, the ones with the suffix "bf" correspond to using multiple eye center position combinations and doing brute force search, and the ones with the suffix "hs" correspond to using multiple eye center position combinations and doing hierarchical search. The proposed approach provides stable results over localization errors.

Again note that the achieved correct recognition rates are very high compared to the ones presented in the literature that are attained on the same database [GSC01, Mar02, TCZZ05, PLL05, FSL06, WGYM07, JM08]. For example, in [GSC01], the experiments are conducted on 116 subjects and the best result obtained against lower face occlusion due to scarf is 81% in the within session experiment and around 65% in the between session experiment. The best results obtained against upper face occlusion due to sunglasses are 48% and 35%, in within and between session experiments, respectively. In [Mar02], where 50 subjects are used for the experiments, the correct recogntion rates are 82% and 50% against scarf, and 80% and 55% against sunglasses, in within session and between session experiments, respectively. Another study [PLL05] that only

performs within session experiments on 135 subjects, reports correct recognition rates of 85.2% against scarf and 80.7% against sunglasses.

The results of using only manually generated labels and the proposed alignment approach are given in Table 5.3. As mentioned before, the correct recognition rates obtained with the proposed alignment approach is superior to the ones obtained using the manually generated labels. No significant performance difference is observed between doing hierarchical search with $W = 11$ and $W = 9$.

| | Manual labels only | Brute-force search | Hierar. search $W = 11$ | Hierar. search $W = 9$ |
|---|---|---|---|---|
| AR1scarf | 91.8% | 97.3% | 97.3% | 98.2% |
| AR1sun | 38.2% | 97.3% | 97.3% | 99.1% |
| ARneutral | 92.7% | 100% | 100% | 99.1% |
| ARinterscarf | 83.6% | 90% | 90% | 88.2% |
| ARintersun | 37.3% | 95.5% | 95.5% | 94.6% |
| FRGC1 | 97.3% | 98.7% | 98.8% | 98.1% |
| FRGC4 | 90.1% | 93.9% | 93.8% | 93.8% |

Table 5.2: Correct recognition rates obtained on the AR and FRGC face data sets. The results in the first column are obtained using just the manually labeled eye center positions to align a training sample. The ones in the other columns are achieved using all the eye center position combinations to align the training sample. Either brute force or hierarchical search is conducted to find the best matching aligned training sample.

Besides the proposed approach, only a few studies have focused on building face recognition systems which are robust to misalignment [Mar02, SCG+04, WYH+08]. In [Mar02], from each training sample, 6615 additional images are generated by perturbing the facial feature locations. These images are projected onto the eigenspace and for each subject the resulting feature vectors are modeled with a Gaussian or a mixture of Gaussians. The identification is done by finding the closest Gaussian or mixture of Gaussians model. Similarly, in [SCG+04], 81 additional samples are derived from each training sample by modifying the eye center locations. These samples are then used as input to the Fisherfaces algorithm [BHK97]. Different from these two studies, in which additional samples are generated, in [WYH+08], misalignment parameters are learned by solving an optimization problem. Although these approaches have improved performance over the baseline, they still cannot fully handle misalignment. For example, in [Mar02], on the AR face database [MB98], around 80% and 50% correct recognition rates are obtained against occlusion over randomly selected 50 subjects, for the within session and between session experiments, respectively, whereas in this study around 95% and 90% correct recognition rates

are achieved for the same conditions over 110 subjects. In [SCG+04], even with two pixels translation, the performance drops by 30%, whereas in this study, the performance stays constant with respect to translations up to 18% of the interocular distance. Similarly, in [WYH+08], there is still a significant performance drop even in the case of small modifications, such as rotation of $\pm 5^{\circ}$, scaling with [0.95, 1.05], or $\pm 1$ pixel shift, whereas in this study rotations up to $\pm 30^{\circ}$, scaling factors up to [0.64, 1.36], and shifts up to 18% of the interocular distance, are tolerated.

## 5.4 What affects more: Occlusion or Registration?

As can be observed from the Figure 5.10, registration plays a very important role in face classification. It also shows that the poor results reported in the literature under the condition of upper face occlusion is not mainly due to missing discriminative information that exists around eyes, but due to imprecise labeling of the eye center coordinates. To justify this finding, additional experiments were performed on the *FRGC1* and *FRGC4* data sets. In the experiments face images are aligned using the manually labeled eye center positions. After the alignment, the blocks at the second and third row of the testing face images are painted to black as shown in Figure 5.12. Again, ten-dimensional local feature vectors that are extracted from each $8 \times 8$ pixels block by removing the first DCT coefficient and keeping the following first ten of them are used. The selected coefficients are divided by their standard deviations and normalized to unit norm.

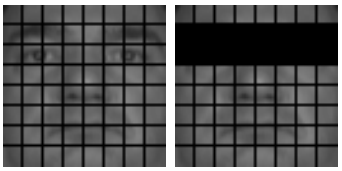


Figure 5.12: Sample aligned face image and corresponding occluded face image.

The classification results with this experimental setup is given in Table 5.3. The results obtained on the test face images without occlusion are also presented in the table for comparison purposes. It is apparent that missing eye region information causes a decrease in the correct classification rate. Especially, if the experiment is more difficult, as in the *FRGC4* experiment, where the training and testing data is collected under uncontrolled conditions, the decrease is more prominent. However, compared to the results in Figure 5.11, the decrease is much less than the one caused by erroneous feature localization. This

validates that registration has more influence on the classification results than occlusion.

| | FRGC1 | FRGC4 |
|---|---|---|
| Without occlusion | 97.9% | 90.8% |
| With occlusion | 95.9% | 83.8% |

Table 5.3: Correct recognition rates obtained on the FRGC experiments. The results in the first row are obtained using test face images that have no occlusion. The ones in the second row are attained using test face images that contain upper face occlusion as depicted in Figure 5.12.

## 5.5 Automatic Feature Selection

Different types of facial appearance variations can be handled by different frequency bands. Therefore, in the proposed face recognition algorithm an automatic frequency selection scheme is employed. In the utilized method the classification is done using multiple frequency bands, that is, by selecting different DCT coefficients with a sliding window of size $M$ and performing classification with each frequency band. The band that provides the maximum separation between the closest two candidates is chosen to be the most reliable band, and its decision is used as the classification result. This way, by using the appropriate frequency band, the algorithm can adapt itself automatically to the changing illumination conditions. The correct classification rates obtained this way can be seen in Table 5.4. As can be observed, automatic feature selection contributes to the performance significantly, especially in the case of uncontrolled conditions and illumination variations.

## 5.6 Comparison with Well-Known Face Recognition Algorithms

In this section, the proposed local appearance-based face recognition algorithm is compared with the well-known generic face recognition algorithms. These algorithms are eigenfaces [TP91], Fisherfaces [BHK97], embedded hidden Markov model [Nef99], and Bayesian face recognition [MJP00].

The obtained correct recognition rates on the face images that are aligned with respect to the manually annotated eye center positions are presented in Table 5.5. In the experiments, for eigenfaces the MAHCOS metric is used in nearest-neighbor classification as suggested in [YDB02]. Fisherfaces and Bayesian face

| Experiment | Without Auto. Feature Sel. | With Auto. Feature Sel. |
|---|---|---|
| FRGC1 | 98.8% | 98.5% |
| FRGC4 | 93.8% | 96.2% |
| CMUPIE | 99.8% | 100% |
| AR1scarf | 97.3% | 97.3% |
| AR1sun | 97.3% | 98.2% |
| ARinterscarf | 90% | 93.6% |
| ARintersun | 95.5% | 95.5% |
| Yale2 | 100% | 100% |
| Yale3 | 100% | 100% |
| Yale4 | 95.6% | 100% |
| Yale5 | 96.8% | 100% |
| ExtYale2 | 100% | 100% |
| ExtYale3 | 100% | 100% |
| ExtYale4 | 93.1% | 98.7% |
| ExtYale5 | 93.1% | 99.0% |

Table 5.4: Results of automatic feature selection experiments.

recognition algorithms are not used when there exists only one training sample per subject. As can be observed, the proposed algorithm achieves significantly higher correct identification rates than well-known face recognition approaches, especially when the experiment is more difficult, for example as in the experiments *FRGC4*, *(Ext)Yale4*, *(Ext)Yale5*. As can be seen, the algorithm can handle illumination variations very well. Compared to the published results in the recent studies [CWX+06, LHK05, ZACJ07], the proposed algorithm performs as well as these approaches. Note that in [CWX+06, LHK05, ZACJ07] prior illumination-related information is utilized and the approaches are tested on databases that contain only illumination variations. The proposed algorithm also reaches very high correct recognition rates even when the face images are recorded under uncontrolled conditions, as in *FRGC4*. The main challenge results from occlusion. Especially sunglasses cause a significant performance drop, since they cause misaligned face images.

The obtained correct recognition rates on the face images that are aligned with respect to the eye center labels provided by the proposed registration approach are given in Table 5.6. Experiments are performed only on the FRGC and AR data sets, where both the original images and corresponding manually annotated eye center labels are available. As can be seen, there is a significant improvement in the performance of the proposed face recognition algorithm, whereas the other algorithms are not able to benefit from the proposed face registration technique.

| | LAFR | Eigenfaces | Fisherfaces | EHMM | Bayesian |
|---|---|---|---|---|---|
| FRGC1 | 97.8% | 92.8% | 97.8% | 90.2% | 95.6% |
| FRGC4 | 91.6% | 57.3% | 65.6% | 42.3% | 63.4% |
| AR1scarf | 89.1% | 28.2% | - | 16.4% | - |
| AR1sun | 32.7% | 17.3% | - | 7.3% | - |
| ARintercarf | 78.2% | 17.3% | - | 10.9% | - |
| ARintersun | 32.7% | 17.3% | - | 3.6% | - |
| CMUPIE | 100 | 64.9% | - | 52.6% | - |
| Yale2 | 100% | 100% | 100% | 100% | 100% |
| Yale3 | 100% | 97.5% | 100% | 55.9% | 100% |
| Yale4 | 100% | 60.1% | 58.7% | 21.0% | 68.1% |
| Yale5 | 100% | 41.3% | 19.6% | 11.6% | 24.3% |
| ExtYale2 | 100% | 100% | 100% | 98.5% | 100% |
| ExtYale3 | 100% | 97.8% | 99.1% | 32.0% | 100% |
| ExtYale4 | 98.7% | 49.2% | 34.6% | 4.4% | 37.6% |
| ExtYale5 | 99.0% | 15.8% | 7.7% | 3.7% | 8.1% |

Table 5.5: Comparison of local appearance-based face recognition algorithm (LAFR) with well-known face recognition algorithms. Face images are aligned with respect to manually labeled eye center positions.

| | LAFR | Eigenfaces | Fisherfaces | EHMM | Bayesian |
|---|---|---|---|---|---|
| FRGC1 | 98.5% | 89.7% | 98.1% | 88.8% | 47.3% |
| FRGC4 | 96.2% | 49.6% | 56.1% | 36.6% | 32.7% |
| AR1scarf | 97.3% | 23.6% | - | 17.3% | - |
| AR1sun | 98.2% | 26.4% | - | 16.4% | - |
| ARinterscarf | 93.6% | 15.5% | - | 14.6% | - |
| ARintersun | 95.5% | 23.6% | - | 8.2% | - |

Table 5.6: Comparison of local appearance-based face recognition algorithm (LAFR) with well-known face recognition algorithms. Face images are aligned with respect to the eye center labels provided by the proposed registration approach.

# 6 Real-World Applications

Several real-world systems, which are based on the proposed face recognition algorithm, have been developed. These systems are classified into three groups:

1. *Face recognition for smart environments:* This application group comprises the identification tasks at a constant location [SES07, EJFS07, EFS07, SBE+06, EP06]. For example, in a smart home, family members can be identified while they are entering the rooms of the house and their location can be determined in order to automatically route incoming phone calls. This application group requires identification of people without any cooperation and under uncontrolled conditions, without any constraints on head pose, illumination, use of accessories, etc.

2. *Face recognition for smart machines:* In this application group, a machine identifies the subject that it interacts with. For instance a car that identifies its driver [SEE+07], a laptop that recognizes its user, or a robot that recognizes the person it serves [SEF+07]. In this application group an implicit cooperation exists between the person and the machine due to the standard actions the person performs, e.g. the driver looking at the road, or the computer user looking at the screen. Therefore, the head pose variations are limited in such systems. The difficulty in this group arises due to changing environmental conditions.

3. *Face recognition for smart image/video retrieval:* In this application group, face recognition is used as a search tool to retrieve relevant images or videos. It is the most difficult application case, since all the conditions are completely unconstrained.

In this chapter, some of the sample systems from these application groups are presented briefly.

## 6.1 Person Identification in Smart Rooms

Person identification in smart rooms is one of the sample applications of *face recognition for smart environments*. The system presented in this section is developed for the CLEAR evaluations [SBB+07], in which the person identification

system needs to identify participants of the lecture-like seminars and interactive small working group seminars.

Doing person identification in a smart room poses many challenges. In terms of face recognition, there is no cooperation of the subjects being identified, there are no constraints on head pose, illumination conditions, use of accessories, etc. Moreover, depending on the distance between the camera and the subject the face resolution varies and generally the face resolution is low. In terms of speaker identification, again, there is no cooperation and the system should handle a large variety of speech signals, corrupted by adverse environmental conditions. The only factors that can help to improve the person identification performance in smart rooms are the video data of the individuals from multiple views provided by several cameras and the multi-channel speech signal provided by microphone arrays that are mounted in the smart room. With the fusion of these modalities, the correct identification rates can be improved further. A sample smart room layout and sample images from different cameras are shown in Figure 6.1.

### 6.1.1 Video-based Face Recognition

The face recognition system is based on the local appearance-based face recognition approach and it processes multi-view video data provided by four fixed cameras. In the training stage all the images from all the cameras are put together. Although the manual annotations of the images are available in the database, due to the low resolution faces these manual labels might be imprecise. In order to prevent the registration errors that can be caused by these imprecise labels, 24 additional samples are also generated by modifying the manually annotated face bounding boxes by moving the center of the bounding box by one pixel and changing the width or height by two pixels. In the testing stage, at an instant all four camera views are compared to the representatives in the database. Their distances are converted to confidence scores using min-max normalization [SUM+05],

$$ns = 1 - \frac{s - min(S)}{max(S) - min(S)}, \tag{6.1}$$

where, $s$ corresponds to a distance value of the test image to one of the training images in the database and $S$ corresponds to a vector that contains the distance values of the test image to the ten closest matches among the training images. The division is subtracted from one, since the lower the distance is, the higher the probability that the test image belongs to that identity class. This way, the score is normalized to the value range of [0,1], closest match having the score '1' and the furthest match having the score '0'. To have equal contribution of each

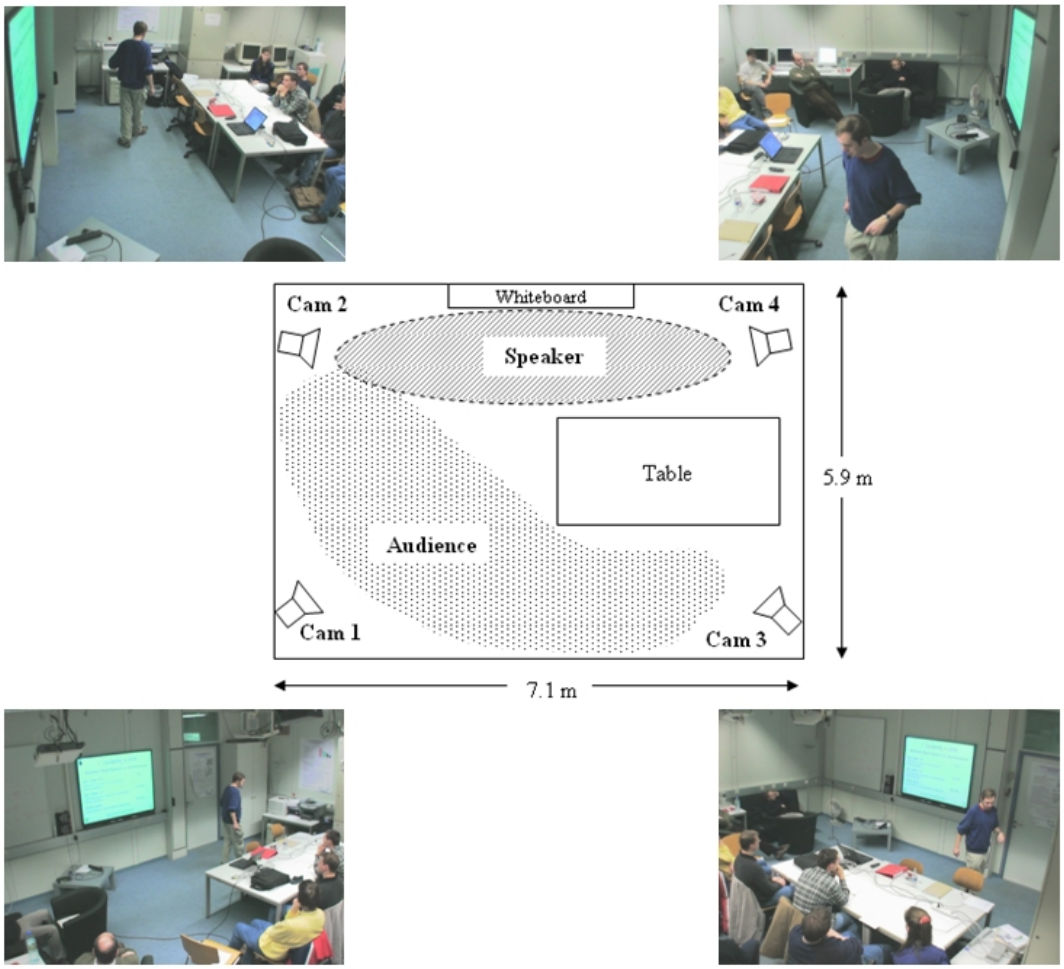



Figure 6.1: A sample smart room layout and sample images captured by cameras mounted at the corners.

frame, these scores are re-normalized by dividing them to the sum of their values. Each frame is weighted using the distance-to-second-closest (DT2ND) metric [SES07]. In [SES07], it has been observed that the difference of the distances, $x$, between the closest and the second closest training samples is generally smaller in the case of a false classification than in the case of a correct classification. It has been found that the distribution of these distances resembles an exponential distribution:

$$\epsilon(x;\lambda) = 0.1\lambda e^{-2x} \text{ with } \lambda = 0.05. \tag{6.2}$$

The weights are then computed as the cumulative distribution function:

$$\omega_{DT2ND}(x;\lambda) = 1 - e^{-2x}. \tag{6.3}$$

The obtained confidence scores are summed over camera views and over image sequence. The identity of the face image is assigned as the person who has the highest accumulated score.

### 6.1.2 Speaker Identification

The speaker identification system is based on mel-frequency cepstral coefficients (MFCC) that are modeled with Gaussian mixture models (GMM) [Fur97, Rey95]. Feature warping and reverberation compensation are applied on MFCC in order to improve robustness against channel mismatch. The reverberation compensation approach uses a different noise estimation compared to the standard spectrum subtraction approach [JPS06]. The feature warping method warps the distribution of a cepstral feature stream to a standardized distribution over a specified time interval [JPS06, PS01, XCN+02]. The identification decision is made as follows:

$$s = i_{\text{argmax}} \left\{ L\left(Y|\Theta_i\right) \right\} \; Y = (y_1, y_2, \ldots, y_N) \tag{6.4}$$

where $s$ is the identified speaker and $L\left(Y|\Theta_i\right)$ is the likelihood that the test feature set $Y$ was generated by the GMM $\Theta_i$ of speaker $i$, which contains $M$ weighted mixtures of Gaussian distributions

$$\Theta_i = \sum_{m=1}^{M} \lambda_m N\left(X, U_m, \Sigma_m\right) \; i = 1, 2, \ldots, S \tag{6.5}$$

where $X$ is the set of training feature vectors to be modeled, $S$ is the total number of speakers, $M$ is the number of Gaussian mixtures, $\lambda_m$, $U_m$, and $\Sigma_m$

are the weight, mean, and diagonal covariance matrix of the $m^{th}$ Gaussian distribution.

As there are 64 channels for each speech recording, GMMs are trained for each speaker on each of the 64 channels. Channel 7 is randomly selected as the test channel. The "frame-based score competition (FSC)" approach is applied when computing the likelihood scores of test features given a speaker with $M$ GMMs. The idea of the FSC approach is to use the set of multiple GMM models rather than a single GMM model. A multiple microphone setup emits speech samples from multiple channels. As a consequence, multiple GMM models can be built for each speaker $k$, one for each channel $i$ and they are referred as $\Theta_k, Chi$. For a total number of 64 channels one gets $\Theta_k = \{\Theta_k, Ch1, \ldots, \Theta_k, Ch64\}$ models for speaker $k$. In each frame the incoming feature vector of channel $Ch7$ is compared to all GMMs $\{\Theta_k, Ch1, \ldots, \Theta_k, Ch64\}$ of speaker $k$. The highest log likelihood score of all GMM models is chosen to be the frame score. Finally, the log likelihood score of the entire test feature vector set $X$ from channel $h$ is estimated as:

$$LL\left(X|\Theta_k\right) = \sum_{n=1}^{N} LL\left(x_n|\Theta_k\right) = \sum_{n=1}^{N} max\left\{LL\left(x_n|\Theta_{k,Chj}\right)\right\}_{j=1,j\neq h}^{64} \tag{6.6}$$

This competition process based on multiple channels differs from the standard scoring process based on one channel in that the per-frame log likelihood scores are not necessarily derived from the same microphone. This speaker identification system is developed by Qin Jin [EJFS07].

### 6.1.3 Fusion

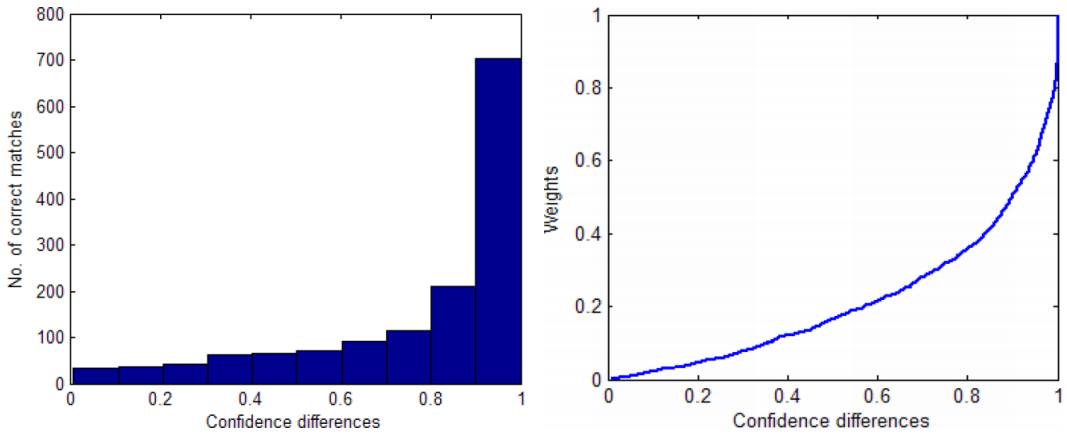
The effects of three main steps of the fusion process are investigated. They are: score normalization, modality weighting, and modality combination.

Score normalization is the first step in the process of modality fusion. Due to the different ways of feature extraction and classification, the distribution of the resulting scores may differ between the modalities. For example, in this study, the face recognition system generates accumulated min-max normalized scores, whereas the speaker identification system provides likelihood scores. In order to combine these scores, two well-known score normalization methods, namely the min-max and hyperbolic tangent normalization are utilized and compared. The min-max normalization can be calculated as in Equation 6.1 without the need of subtracting the obtained division value from one, since the modality scores are directly proportional to the modality confidences. Hyperbolic tangent normalization nonlinearly maps the confidence scores to the (0,1) range and calculated as,

$$nc = \frac{1}{2}\left[\tanh\left(0.01\frac{(c - mean(C))}{std(C)}\right) + 1\right] \tag{6.7}$$

where, $nc$ denotes the normalized confidence score, $c$ denotes a confidence score of an identity candidate in the database whose image or speech signal has been compared with, and $C$ denotes the vector that contains the confidence scores of all the identity candidates in the database.

Modality weighting is the second step in the fusion process. A new adaptive modality weighting scheme is introduced which is based on the separation of the best two matches. It is named as cumulative ratio of correct matches (CRCM) and utilizes non-parametric modeling of the distribution of the correct matches with respect to the confidence differences between the best two matches. It relies on the observation that the difference of the confidences between the closest and the second closest training samples is generally smaller in the case of a false classification than in the case of a correct classification. The greater the confidence difference between the best two matches is, the higher the weight the individual modality receives. Figure 6.2 shows the obtained correct match distribution over the confidence differences and the corresponding weighting model for the face recognition system. This weighting model has been computed on a validation set by taking the cumulative sum of the number of correct matches achieved at a confidence difference between the best two matches.




(a) (b)

Figure 6.2: (a) Distribution of the correct matches, (b) The weighting model.

In addition to the adaptive modality weighting scheme, individual correct identification rates of each modality that are obtained on the validation set are also taken into account. Modalities are assigned with fixed weights according to their performance.

Finally, the modalities are combined using the well-known classifier combination methods: sum rule, product rule, and max rule [KHDM98].

### 6.1.4 Experimental Results

The experiments have been conducted on a database that has been collected by the CHIL consortium [WSS04] for the CLEAR 2007 evaluations [SBB+07]. The recordings are from lecture-like seminars and interactive small working group seminars that have been held at different CHIL sites: AIT, Athens, Greece, IBM, New York, USA, ITC-IRST, Trento, Italy, UKA, Karlsruhe, Germany, and UPC, Barcelona, Spain. Sample images from the recordings can be seen in Figure 6.3. The data used for the identification task consists of short video sequences of 28 subjects, where the subject is both speaking and visible to the cameras at the same time. The recording conditions are uncontrolled and depending on the camera view and the position of the presenter/participant low resolution faces ranging between 10 to 50 pixels resolution are acquired. Two different training —15 and 30 seconds— and four different testing durations —1, 5, 10, 20 seconds— are used in the experiments. Identity estimates are provided at the end of each test sequence duration using the available audio-visual data.

Figure 6.3: Sample images from different smart rooms.

In the database, face bounding box labels are available for every 200ms. Only these labeled frames are used for the experiments. The face images are cropped and scaled to $40 \times 32$ pixels resolution. They are then divided into $8 \times 8$ pixels resolution non-overlapping blocks making 20 local image blocks. From each image block ten-dimensional DCT-based feature vectors are extracted and they are concatenated to construct the final 200-dimensional feature vector.

13-dimensional MFCC is extracted from the speech signal as the speaker feature. A GMM with 32 mixtures is trained for each speaker using the expectation-maximization (EM) algorithm under the 30 seconds training condition and 16 mixtures for each speaker under the 15 seconds training condition. The classification is performed as described in Section 6.1.2.

### Experiments on the Validation Set

The correct identification rates of the face recognition and speaker identification systems obtained on the validation set are presented in Table 6.1. In the table, the first row contains training durations and the second row contains testing durations. The third and fourth rows show the results for different training-testing duration combinations. As expected, as the duration of training or testing increases the correct identification rate increases. Both of the systems achieve 100% correct identification when the systems are trained with 30 seconds of data and tested with the sequences of 20 seconds duration. Face recognition is found to be significantly superior to speaker identification at the other training-testing duration combinations.

These results are used to determine the fixed weights that each modality receives. It is done in two different ways. The first way is by determining the weights directly proportional to the correct identification rates. For example, if the face recognition system has 100% and the speaker identification system has 85% correct identification rates, then they are weighted by 1 and 0.85 respectively for that training-testing duration combination. The second way is by determining the weights inversely proportional to the false identification rates. For instance, if the face recognition system has 5% and the speaker identification system has 10% false identification rates, then the face recognition system receives twice as much weight than the speaker identification system.

| Training duration (sec) | 15 | | | | 30 | | | |
|---|---|---|---|---|---|---|---|---|
| Testing duration (sec) | 1 | 5 | 10 | 20 | 1 | 5 | 10 | 20 |
| Face Reco. (%) | 91.4 | 99.1 | 100 | 100 | 94.3 | 100 | 100 | 100 |
| Speaker Id. (%) | 56.4 | 67.9 | 89.3 | 92.9 | 61.1 | 84.8 | 98.2 | 100 |

Table 6.1: Correct identification rates of the individual modalities on the validation set.

### Experiments on the Test Set

The correct identification rates of the face recognition and speaker identification systems obtained on the test set are given in Table 6.2. Similar to the obtained

results on the validation set, as the duration of training or testing increases the correct identification rate increases. Only at one second testing, it does not change too much for the speaker identification system. As can be noticed, on the test set the speaker identification performs as well as or even better than the face recognition at longer duration test segments. In the case of fixed modality weighting, this implies that the validation set is misleading, since on the validation set face recognition has been found to be more successful at these test segments. The other observation that can be derived by comparing Tables 6.1 and 6.2 is the lower correct identification rates obtained on the testing set. The main reason is that the time gap between training set and test set recordings is larger than the one between the recordings of training and validation sets.

| Training duration (sec) | 15 | | | | 30 | | | |
|---|---|---|---|---|---|---|---|---|
| Testing duration (sec) | 1 | 5 | 10 | 20 | 1 | 5 | 10 | 20 |
| Face Reco. (%) | 84.6 | 90.8 | 93.3 | 94.6 | 89.3 | 94.4 | 94.6 | 96.4 |
| Speaker Id. (%) | 41.9 | 69.6 | 92.0 | 96.4 | 41.2 | 78.3 | 96.4 | 99.1 |

Table 6.2: Correct identification rates of the individual modalities on the test set.

### Fusion Experiments

In the following subsections, steps of the fusion process are examined.

***Comparison of Score Normalization Methods***: In the first fusion experiment, the min-max and hyperbolic tangent score normalization methods are compared. For modality weighting, fixed weights are used which are directly proportional to the correct identification rates obtained on the validation set. Sum rule is utilized for classifier combination. The resulting correct recognition rates are shown in Table 6.3. As can be observed, there is no significant performance difference between using min-max or hyperbolic tangent methods for score normalization. Although, for this experiment a very simple fixed weighting scheme is used, in most of the training-testing duration combinations the correct identification rates are higher than the ones obtained by the individual modalities.

| Training duration (sec) | 15 | | | | 30 | | | |
|---|---|---|---|---|---|---|---|---|
| Testing duration (sec) | 1 | 5 | 10 | 20 | 1 | 5 | 10 | 20 |
| min-max (%) | 84.8 | 91.1 | 94.2 | 94.6 | 89.8 | 94.9 | 95.5 | 97.3 |
| tanh (%) | 85.8 | 91.7 | 95.1 | 94.6 | 90.4 | 96.0 | 96.9 | 98.2 |

Table 6.3: Comparative results of min-max and hyperbolic tangent score normalization methods.

***Comparison of Modality Weighting Methods***: In the second fusion experiment, the modality weighting schemes are analyzed. First, the fixed weighting schemes, at which the weights are either directly proportional to the correct classification rate or inversely proportional to the false identification rate obtained on the validation set, are compared. These fixed modality weighting schemes are named as *DPC* and *IPF*, respectively. Sum rule is used for classifier combination. The results are presented in Table 6.4. As already mentioned in the previous subsection, even with these simple weighting schemes, in most of the training-testing duration combinations the correct recognition rates are higher than the ones obtained by the individual modalities.

| Training duration (sec) | 15 | | | | 30 | | | |
|---|---|---|---|---|---|---|---|---|
| Testing duration (sec) | 1 | 5 | 10 | 20 | 1 | 5 | 10 | 20 |
| DPC, min-max (%) | 84.8 | 91.1 | 94.2 | 94.6 | 89.8 | 94.9 | 95.5 | 97.3 |
| DPC, tanh (%) | 85.8 | 91.7 | 95.1 | 94.6 | 90.4 | 96.0 | 96.9 | 98.2 |
| IPF, min-max (%) | 84.6 | 90.8 | 93.3 | 94.6 | 89.4 | 94.4 | 94.6 | 97.3 |
| IPF, tanh (%) | 84.8 | 90.8 | 93.3 | 94.6 | 89.9 | 94.4 | 94.6 | 98.2 |

Table 6.4: Comparative results of fixed weighting schemes.

The results obtained with the more sophisticated adaptive modality weighting scheme are given in Table 6.5. Again, sum rule is used for classifier combination. Compared to the Table 6.4, CRCM provides a significant increase in correct identification rates. Note that, in terms of performance of each modality, the validation set was not quite representative. On the validation set, at some training-testing duration combinations face recognition was found to be superior, but on the test set speaker identification performed better. Therefore, performance based fixed weighting can be misleading. The results obtained by CRCM indicate that confidence differences are more robust cues for modality weighting.

| Training duration (sec) | 15 | | | | 30 | | | |
|---|---|---|---|---|---|---|---|---|
| Testing duration (sec) | 1 | 5 | 10 | 20 | 1 | 5 | 10 | 20 |
| CRCM, min-max (%) | 86.3 | 93.5 | 98.2 | 99.1 | 89.6 | 97.3 | 98.7 | 99.1 |
| CRCM, tanh (%) | 81.0 | 91.7 | 97.3 | 98.2 | 84.7 | 96.9 | 99.1 | 100 |

Table 6.5: Results of adaptive weighting scheme.

The adaptive weight and the fixed weights are also combined. Score normalization is done with min-max normalization and the classifiers are combined with the sum rule. As can be observed from Table 6.6, there is no significant performance difference between the CRCM and DPC+CRCM results. The performance degrades with IPF+CRCM. The reason is the hard modality weighting in IPF. Since, on the validation set at some training-testing combinations, face

recognition had 0% false identification rate, at these combinations only the face recognition system's decision is trusted.

| Training duration (sec) | 15 | | | | 30 | | | |
|---|---|---|---|---|---|---|---|---|
| Testing duration (sec) | 1 | 5 | 10 | 20 | 1 | 5 | 10 | 20 |
| DPC + CRCM (%) | 86.7 | 93.5 | 98.2 | 99.1 | 89.9 | 97.3 | 98.7 | 99.1 |
| IPF + CRCM (%) | 86.7 | 91.7 | 93.3 | 94.6 | 89.9 | 94.4 | 94.6 | 99.1 |

Table 6.6: Comparative results of combined adaptive and fixed weighting schemes.

***Comparison of Classifier Combination Methods***: Finally, the weighted scores are combined using the sum rule, product rule, and max. rule. Combined DPC and CRCM is used for modality weighting. From Table 6.7, it can be seen that the max. rule operates better on the min-max normalized confidence scores. Sum rule and max. rule are found to perform slightly better than the product rule. However, no big difference is observed in the correct identification rates.

| Training duration (sec) | 15 | | | | 30 | | | |
|---|---|---|---|---|---|---|---|---|
| Testing duration (sec) | 1 | 5 | 10 | 20 | 1 | 5 | 10 | 20 |
| Sum, min-max (%) | 86.7 | 93.5 | 98.2 | 99.1 | 89.9 | 97.3 | 98.7 | 99.1 |
| Sum, tanh (%) | 83.2 | 92.9 | 97.8 | 98.2 | 87.2 | 96.9 | 99.1 | 100 |
| Product, min-max (%) | 86.7 | 92.6 | 95.5 | 95.5 | 90.4 | 96.2 | 97.3 | 98.2 |
| Product, tanh (%) | 86.8 | 92.6 | 95.5 | 95.5 | 90.7 | 96.2 | 96.9 | 98.2 |
| Max., min-max (%) | 84.8 | 92.2 | 97.8 | 99.1 | 88.2 | 96.2 | 90.1 | 100 |
| Max., tanh (%) | 68.8 | 83.5 | 93.8 | 97.3 | 72.7 | 89.1 | 97.8 | 100 |

Table 6.7: Comparative results of classifier combination methods.

### 6.1.5 Summary of the Experiments

In the experiments, no significant performance difference between well-known score normalization or classifier combination approaches is found. It is observed that the modality weighting has a major impact on the correct identification rate. An adaptive modality weighting model is proposed, which is derived from the confidence differences between the best two matches. It is named as cumulative ratio of correct matches (CRCM) and the weighting model is computed by taking the cumulative sum of the number of correct matches achieved at a confidence difference between the best two matches. In Table 6.8, the correct identification rates of the individual modalities and the multimodal system are listed. The multimodal system included in the table uses min-max normalized confidence scores, CRCM modality weighting, and the sum rule. From the table,

it is clear that multimodal fusion contributes to the performance significantly. This also indicates that the face and voice modalities are complementary biometric traits.

| Training duration (sec) | 15 | | | | 30 | | | |
|---|---|---|---|---|---|---|---|---|
| Testing duration (sec) | 1 | 5 | 10 | 20 | 1 | 5 | 10 | 20 |
| Face Reco. (%) | 84.6 | 90.8 | 93.3 | 94.6 | 89.3 | 94.4 | 94.6 | 96.4 |
| Speaker Id. (%) | 41.9 | 69.6 | 92.0 | 96.4 | 41.2 | 78.3 | 96.4 | 99.1 |
| Combined (%) | 86.3 | 93.5 | 98.2 | 99.1 | 89.6 | 97.3 | 98.7 | 99.1 |

Table 6.8: Comparative results of individual modalities and the multimodal system.

## 6.2 Door Monitoring System

The door monitoring system is one of the sample applications of *face recognition for smart environments*. The real-time face recognition system presented in this section monitors the entrance door of a seminar room. Individuals are recognized when they enter the room. They behave naturally, since they are not required to interact with the recording system in any special way, e.g., looking at the camera. As a consequence, the system is confronted with real-world facial appearance variations that are caused by partial face occlusion, changing illumination, and head pose (Figure 6.4).

Figure 6.4: Sample images from the door monitoring system.

Faces are detected in a two-stage process. First, regions of interest are determined by skin color segmentation and then the eyes are detected with a classifier cascade of Haar-like features [VJ04]. The eye positions are used to register the faces to a fixed orientation and scale (Figure 6.5). Please note the variations

in expression, illumination, pose, and resolution as well as blurring effects from motion.

Figure 6.5: Sample aligned images from the door monitoring system.

To evaluate the recognition performance, both a k-nearest-neighbors (k-NN) and a Gaussian mixture model (GMM) approach are used. In the k-NN case, video-based classification is achieved by accumulating the normalized individual frame scores. In the GMM approach, this is done with Bayesian inference. As not all frames are of the same quality, a weighting scheme consisting of two sub-schemes is introduced into the k-NN approach to weight each frame's influence on the final decision. The scheme distance-to-model identifies frames that are inconsistent with the training data, therefore modeled inappropriately, and assigns them a lower weight. Distance-to-second-closest compares the top-2 matches and reduces a frame's weight if the classification is ambiguous, that is, if the top-2 matches are very close. A smoothed version of the GMM approach is also developed with the underlying idea that the identity of a person does not change over time. Consequently, frames which are inconsistent with the current hypothesis get a small weight. This approach still allows a change of identity if there is strong enough evidence, but it avoids rough sudden jumps between different classifications [SES07].

In order to show the robustness of the local appearance-based face recognition approach under real-world conditions, it is first compared with several well-known face recognition algorithms, such as eigenfaces [TP91], Fisherfaces [BHK97, ZCP99], and Bayesian face recognition [MJP00] on the collected door database. This experiment is conducted frame-based, that is, the classification is performed based on single frames. A database of 2294 video sequences of 41 individuals, which have been automatically recorded during seven months with the developed system [SES07], is used for the experiments. The data is divided into training and testing sets according to the recording date. The sequences recorded earlier are used for training and the ones recorded later are used for testing. Five-dimensional local DCT-based features are used for the local appearance-based face recognition algorithm, making a 320-dimensional combined feature vector. The feature vectors are classified using a nearest neighbor classifier. The L1 norm is used as a distance metric. The same feature dimensionality is used for the other face recognition approaches as well. For the eigenfaces, Mahalanobis cosine (MAHCOS) [YDB02] is also used as a distance

metric in nearest neighbor classification. The correct identification rates are given in Table 6.9. The local appearance-based face recognition approach outperforms well-known face recognition algorithms. The most interesting result that can be observed from this table is the very low correct identification rate obtained by Bayesian face recognition [MJP00] which has been known to be the one of the best performing face recognition algorithms. Varying pose, illumination changes, registration errors make the intra-personal and extra-personal variations almost identical, which causes this low performance.

| Method | Performance |
|---|---|
| Local DCT | 80.6% |
| LDA | 75.9% |
| PCA, L1 | 68.7% |
| PCA, MAHCOS | 66.1% |
| Bayesian | 28.0% |

Table 6.9: Frame-based experiment results.

In video-based evaluations, the same database is used, but this time the classification is performed using the entire sequence. As can be seen from Table 6.10, the system successfully extends the frame-based approach to video-based data. In the table, *Video-based* results correspond to equally weighting each frame, whereas *Weighted* corresponds to frame weighting approach in k-NN and *Smooth* corresponds to smoothing process in the GMM approach. Correct recognition rates are significantly higher when the entire sequence of images are used as the increased amount of input data compensates for low-quality frames. Results improve further if bad frames can be identified and their influence is reduced. Note that, the frame-based result with k-NN in Table 6.10 is lower than the one in Table 6.9. The reason is that the training samples are clustered in video-based face recognition to make the system run in real-time.

| Classifier | Frame-based | Video-based | Weighted | Smooth |
|---|---|---|---|---|
| KNN | 68.4% | 90.9% | 92.5% | N/A |
| GMM | 62.7% | 86.7% | N/A | 87.8% |

Table 6.10: Correct recognition rates achieved by the door monitoring system. Weighted and Smooth are only available for k-NN and GMM, respectively.

## 6.3 Visitor Interface

The visitor interface system is one of the sample applications of *face recognition for smart machines*. The system performs open-set face recognition and

Figure 6.6: A snapshot of the visitor interface system in operation.

it has been designed as a visitor interface, where a visitor looks at the monitor before knocking on the door. A welcome message is displayed on the screen. While the visitor is reading the welcome message, the system identifies the visitor unobtrusively without needing the person's cooperation. According to the identity of the person, the system customizes the information that it conveys about the host. For example, if the visitor is unknown, the system displays only availability information about the host. If the visitor is known, depending on the identity of the person, more detailed information about the host's status is displayed. A snapshot of the system in operation can be seen in Figure 6.6.

Open-set identification can be seen as the most generic form of face recognition problem. Several approaches can be considered to solve it. One of them is to perform verification and classification sequentially, that is, to perform first verification to determine whether the encountered person is known or unknown and then, if the person is known, finding out who he/she is by doing classification. An alternative approach can be training an unknown identity class and running just a classifier. A third option is running just verifiers. A test image is compared against each known subject to see whether it belongs to that subject or not. If all the verifiers reject, then the image is classified as belonging to an unknown person. If one or more verifiers accept, then the image is classified as belonging to a known person. Among these approaches, the last one is opted for, which is named as the multi-verification approach. The main reason for this choice is the better discrimination provided via multiple verifications. The first method requires a verifier to determine known/unknown persons. This requires training the system with face images of known and unknown persons. Since human faces are very similar, generating a single known/unknown verifier can not be highly discriminative. In the second method, training a separate unknown class would not be feasible. Because the unknown class covers unlimited number of subjects that cannot be completely modeled using a limited number

of subjects. On the other hand, with the multi-verification approach, only the available subjects are modeled.

In the system, an identity verification component is trained for each known subject in the database. When a test image is captured, it is verified against each known subject in the gallery, either using a support vector machine or nearest neighbor classifier. If all of the verifiers reject, the person is reported as unknown; if one accepts, the person is accepted as known and the verified identity is assigned to him/her; if more than a single verifier accepts, the person is accepted as known and the identity of the verifier with the highest confidence is assigned to him/her. Verifier confidences are inversely proportional to the distance values obtained by the nearest-neighbor or SVM classifier.

The system is evaluated on a data set that consists of short video recordings of 50 subjects captured in front of an office over four months. There is no control on the recording conditions. The sequences consist of 150 consecutive frames where both face and eyes are detected. Figure 6.7 shows some sample captured frames. As can be seen, the recording conditions can change significantly due to lighting, motion blur, distance to camera, and change of the view angle. The subjects are assigned to two separate groups as known and unknown subjects. In the experiments, five subjects, who are the members of a research group, are classified as known people. 45 subjects who are mainly university students and some external guests, are classified as unknown people. The set of recording sessions is then further divided into training and testing data. Known subjects' recordings are splitted into non-overlapping training and testing sessions. From the 45 unknown subjects, 25 of them are used for training and twenty of them are used for testing. There is no overlap between the unknown subjects who are used for training and testing. The organization of the data can be seen in Table 6.11. As can be noticed, for each verifier training, there exists around 600 frames (4 sessions, 150 frames per session) from the known subject. On the other hand the number of available frames from the unknown subjects is around 3750 frames (25 sessions, 150 frames per session). In order to limit the influence of data imbalance during verifier training, unknown recordings are undersampled to 30 images per used training session, making a total of 750 frames.

| Training data | | |
|---|---|---|
| Known | 5 subjects | 4 sessions |
| Unknown | 25 subjects | 1 session |
| Testing data | | |
| Known | 5 subjects | 3 – 7 sessions per person |
| Unknown | 20 subjects | 1 session per person |

Table 6.11: Data organization for open-set face recognition experiments.

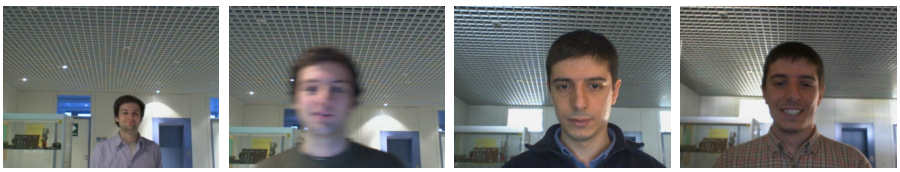


Figure 6.7: Sample images from the data set.

Open-set face recognition systems can make three different types of errors. False classification rate (FCR) indicates the percentage of correctly accepted but misclassified known subjects, whereas false rejection rate (FRR) shows the percentage of falsely rejected known subjects and false acceptance rate (FAR) corresponds to the percentage of falsely accepted unknown subjects. The equal error rate (EER) is defined as the point on the receiver operating characteristic (ROC) curve where $FAR = FRR + FCR$.

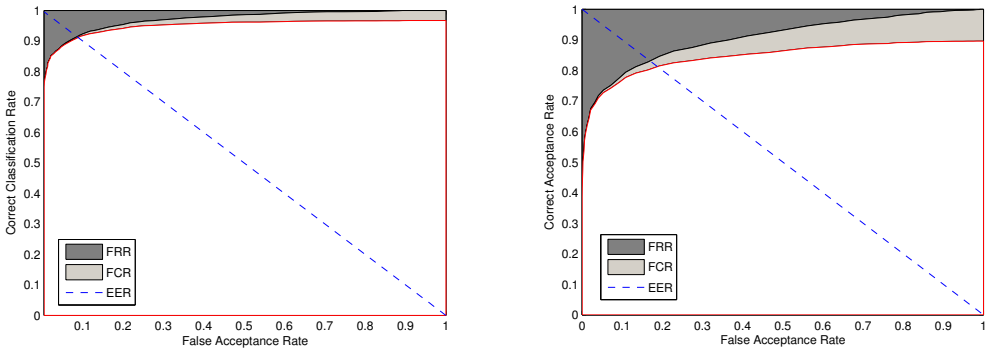
First, the frame-based verification is performed. Frame-based verification implies doing verification using a single frame instead of an image sequence. Each frame in the recordings is verified separately, that is, the decision is taken only using a single frame at a time. The results of this experiment at the point of equal error rate are reported in Table 6.12. In the table CCR denotes the correct recognition rate and CRR denotes the correct rejection rate. The SVM-based classification outperforms the nearest-neighbor approach in correct classification performance by almost 10%, false acceptance rate and false rejection rate are also lower.

| Classifier | CCR | FRR | FAR | CRR | FCR |
|---|---|---|---|---|---|
| NN | 81.6 % | 15.5 % | 17.8 % | 82.2 % | 2.9 % |
| SVM | 90.9 % | 8.6 % | 8.5 % | 91.5 % | 0.5 % |

Table 6.12: Frame-based nearest-neighbor and SVM classification results.

Figure 6.8 shows the ROC curves for both SVM and NN-based verification. To analyze the effect of FRR and FCR on the performance, they are plotted separately in these figures. The dark gray colored region corresponds to the errors due to false known/unknown separation and the light gray colored region corresponds to the errors due to misclassification. Similar to the finding in [SES07], it is observed that determining whether a person is known or unknown is a more difficult problem than finding out who the person is.

As the data set consists of short video sequences, the additional information can be used to further improve classification results. In the video-based verification, the decision is taken after using all the frames of the entire video. Table 6.13 shows the improved results with the help of accumulated scores. As




(a) ROC curve of SVM-based classification. (b) ROC curve of NN-based classification.

Figure 6.8: ROC curves of the frame-based verification.

can be seen SVM-based classification outperforms the nearest neighbor-based classification.

| | CCR | FRR | FAR | CRR | FCR |
|---|---|---|---|---|---|
| NN | 95 | 5 | 15 | 85 | 0 |
| SVM | 100 | 0 | 0 | 100 | 0 |

Table 6.13: Video-based nearest-neighbor and SVM classification results.

## 6.4 Face Recognition for Humanoid Robots

Similar to the *visitor interface* system, face recognition for humanoid robots is a sample application of *face recognition for smart machines*. In this scenario, the face recognition system identifies the person, who interacts with the robot. This system is developed for the Karlsruhe humanoid robot [SEF+07]. It performs open-set identification and has the same working principles as the *visitor interface* system. The only difference is, it is integrated to a more sophisticated tracker [NS07] than the one used for the visitor interface. A sample snapshot of the system can be seen in Figure 6.9.

## 6.5 Person Identification in Movies

The proposed face recognition algorithm is also used for person retrieval in videos, which is a sample application of *face recognition for smart image/video retrieval*. The developed system first segments the input video into its shots and

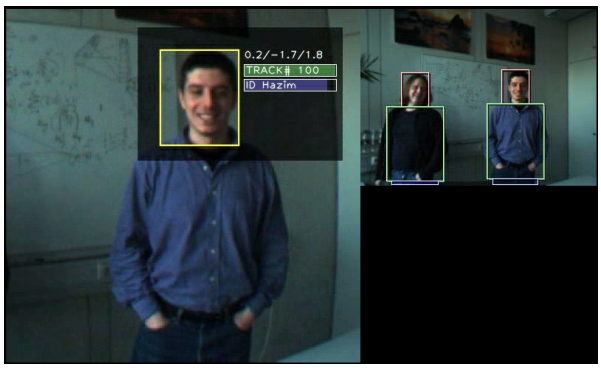



Figure 6.9: A snapshot illustrating face recognition with a humanoid robot.

in each shot automatically detects and tracks persons. It extracts features from these tracks which can be used to reliably identify the persons in the video. This application group poses many challenges. Unlike the first application group, the location is not fixed. There exists no implicit cooperation between the subjects and the system as in the second application group. Therefore, this application group is the most difficult application case, since all the conditions are completely unconstrained. Sample images from this application scenario can be seen in Figure 6.10.

Figure 6.10: Sample frames from a TV series.

The shot boundary detection system contains four separate detectors [EFG+07], one for each type of shot boundary transition: cuts, fast dissolves, fade in/fade outs, and dissolves. Within the shot, the faces are tracked with a particle filter approach. The tracking algorithm uses skin color, which is modeled in HSV color space and a face detector, which is based on a cascade of classifiers that use Haar-like features [VJ04], as observation cues. The face detector has been modified, so that in addition to providing binary decisions —face exists or not—, it also generates confidence values.

For face classification, three types of application scenarios are considered. The first one is closed-set identification. In this application scenario, given a set of main characters in a TV series, the system is required to determine who is who in each shot. The second application scenario is automatic retrieval. In this scenario, given a set of query faces of a person, it is required to find the faces of the same person in different shots. The last one is interactive retrieval. This

| Algorithm | Correct classification rate |
|---|---|
| DCT (10) | 70.5% |
| DCT (5) | 69.4% |
| EHMM [Nef99] | 67.9% |
| Fisherfaces [BHK97] | 63.2% |
| Eigenfaces [TP91] | 50.4% |
| Bayesian [MJP00] | 27.7% |

Table 6.14: Results for the closed-set identification scenario.

scenario is similar to the automatic retrieval, the only difference is that, the user can provide feedback to the system to refine the search.

The developed system is evaluated on an episode of British TV Series *Coupling*. The face tracks and the corresponding identities are labeled to provide the ground truth. The obtained closed-set correct identification results are given in Table 6.14. The experiment is conducted frame-based. That is, the classification is performed and evaluated on single frames. Two different local feature dimensions, five and ten, are used for the proposed local appearance-based face recognition approach. Some of the well-known face recognition algorithms are also tested on the same data. It can be observed that the correct recognition rates are lower than the ones obtained on the standard benchmark databases. This was also the case in frame-based experiments of the *door monitoring* system. The main reason is that under uncontrolled conditions, using single frames is not sufficient to produce a reliable result. Moreover, it has also been observed that one of the main characters has significantly less amount of training data compared to the other main characters, leading to false classifications of the test face images belonging to that character.

Person retrieval results are plotted in Figure 6.11. In the interactive retrieval, first, the top matching faces are returned to the user. The similarity of these matches to the query face image set exceeds a certain threshold. The user then selects the correctly retrieved faces and with this additional training data the system refines and enlarges its retrieval results. The performance of the system improves significantly with the interactive setup. For example, with ten feedbacks from the user, the recall rate reaches around 90% at a precision rate of 95%.

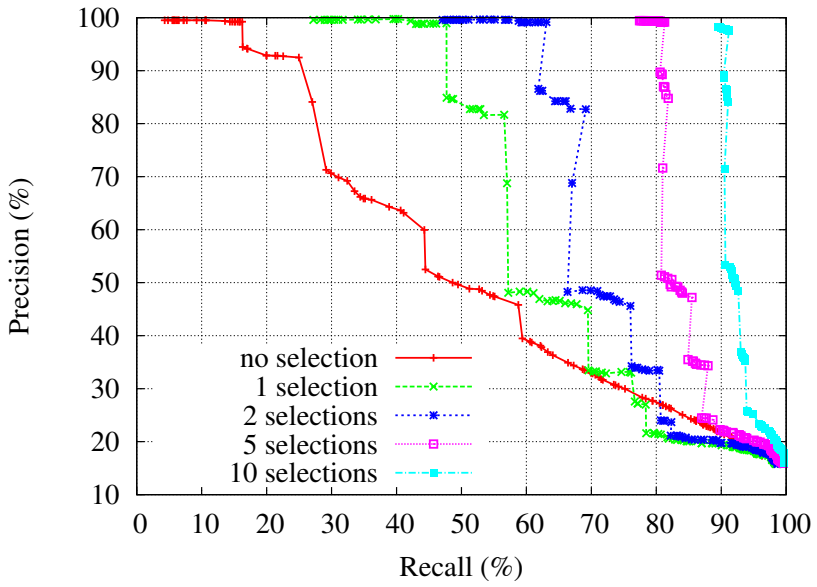



Figure 6.11: Person retrieval results.

## 6.6 Local Depth-based 3D Face Recognition

In addition to the real-world systems, the proposed algorithm is also extended for 3D face recognition. In this section, this 3D face recognition system is presented.

### 6.6.1 Discrete Cosine Transform-based Local Depth Models

Feature extraction from depth images using local appearance-based face representation can be summarized as follows: The input depth image is divided into blocks of $8 \times 8$ pixels size. Each block is then represented by its DCT coefficients. These DCT coefficients are ordered using the zig-zag scanning pattern [GW01]. From the ordered coefficients, $M$ of them are selected according to the feature selection strategy resulting in an $M$-dimensional local feature vector. Finally, the DCT coefficients extracted from each block are concatenated to construct the overall feature vector of the corresponding depth image.

### 6.6.2 Experimental Results

Extensive experiments have been conducted on the Face Recognition Grand Challenge (FRGC) version 2.0 data set [PFS+05] to analyze the performance

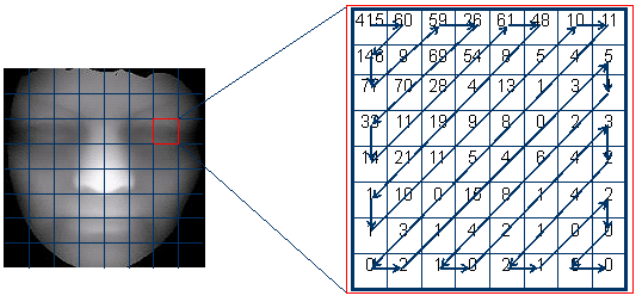


Figure 6.12: Local blocks in depth image, DCT features are extracted using zig-zag scan.

of the proposed local depth-based 3D face recognition approach. The 3D data corpus of the FRGC database was collected by imaging subjects with a range scanner. For the experiments, the subjects who have at least two range images in the spring 2003 recordings of the database are selected and their images from these recordings are used for training. For testing, the range images of these subjects from the spring 2004 recordings are used. The training data contains neutral expressions, whereas the testing data contains different expressions, such as frowning, smiling, etc. In total 218 range images of 109 subjects are used for training, where each individual has two samples, and 758 range images are used for testing, where each individual has a different number of samples, ranging from one to twelve. Sample pre-processed range images and the corresponding registered depth images from the training and testing data sets are shown in Figure 6.13. The depth images are scaled to $64 \times 64$ pixels resolution.

In the experiments, a nearest neighbor classifier is used with the L1 norm as distance metric, since it has been shown that the L1 norm provides better results than the L2 norm and normalized correlation [ES06]. A support vector machine (SVM) classifier is also tested as a more sophisticated classifier.

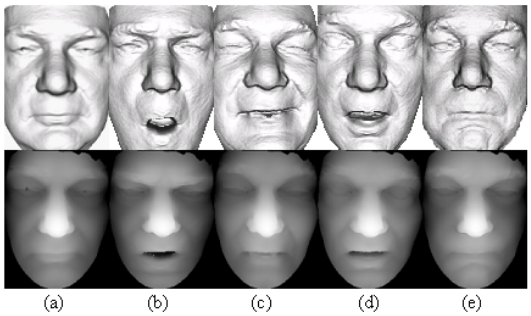



Figure 6.13: First row: Pre-processed range images rendered with shade model in training and test set. Second row: registered depth images. (a) neutral (b) frowning (c) smiling (d) surprised (e) puffy

**Analysis of Local Depth-based 3D Face Recognition**

In the first part of the experiments, the effects of local feature dimension, feature selection, and face registration on face recognition performance are analyzed. Figure 6.14 shows the face recognition performance with respect to increasing local feature dimensionality. In this experiment, the 3D faces were registered using all of the eleven manually labeled landmark points, and depth images were generated using ray-casting. At each local block, the first coefficient was removed from the ordered DCT coefficients, since it only represents the average depth of a local image block. From the remaining coefficients, the first $M$ of them were selected. The selected local feature vector was normalized to have unit norm as suggested in [ES06] which has been shown to improve the face recognition performance. As can be observed from the figure, high correct recognition rates can be attained by using only five-dimensional local feature vectors. The performance continues to increase slightly till to the feature dimension of ten. The correct recognition rate remains the same or decreases slightly, when the dimensionality increases further. Therefore, ten-dimensional local feature vectors are used for the rest of the experiments.

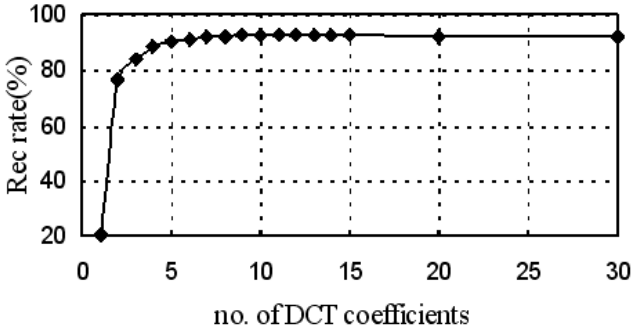



Figure 6.14: Correct recognition rate vs. local feature dimensionality.

The second experiment assesses the effect of frequency content on face recognition performance. In order to consider the correct recognition rate with different sets of features having different frequency contents, the first $N$ ($N = 0, 1, \ldots, 6$) low frequency DCT coefficients are discarded and the following first ten coefficients are used. The same experiments are conducted with three different registration setups to observe whether the selected features produce consistent results over each registration framework. The registration configurations were named $ManualLM\_11$, $AutoLM\_11$, and $ICP$. $ManualLM\_11$ and $AutoLM\_11$ correspond to face registration with eleven manually and automatically labeled landmarks, respectively, whereas $ICP$ corresponds to registration with $ICP$. Ray-casting is used to generate depth images from the registered 3D faces. Correct recognition rates obtained from these three different experimental setups

are plotted in Figure 6.15. In all of the experiments, the best results are obtained using the DCT-4 feature set, which implies that removing the coefficients that represent horizontal and vertical changes as well as the one that represents the average depth, improves the face recognition performance.

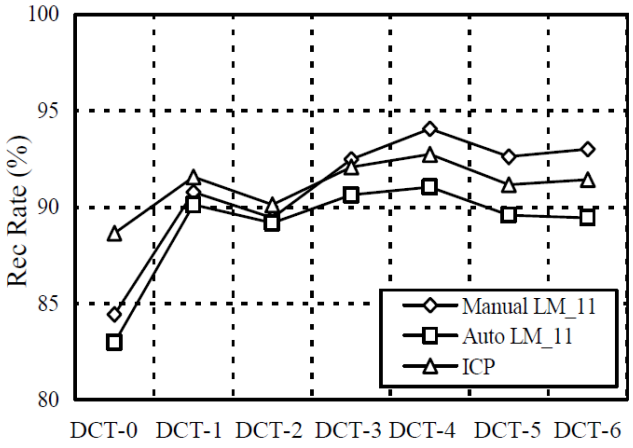



Figure 6.15: Recognition rate of DCT-based local depth approach using different feature sets. (DCT-N: Discard first $N$ coefficients and select the first ten coefficients from the remaining ones.)

The effects of landmark points used for registration and the depth image generation techniques are analyzed in the third experiment. Usually using more landmarks for registration improves correspondence, but if the landmark points are poorly placed, correspondence may get worse. If more than necessary landmark points are used while performing the TPS warping, the cumulative noise of the landmarks may result in degenerate deformations. Therefore, some of the landmarks are discarded to analyze their effectiveness. In the experiment, five possible landmark combinations, illustrated in Figure 6.16a, are tested for registration. Both ray-casting and closest-point methods are used for depth image generation. The DCT-4 feature set is used for classification. The corresponding results can be seen in Figure 6.16b. The highest scores are achieved by selecting ten landmarks, excluding the landmark located in the middle of the mouth. This is expected, since this point is not easy to label on faces that have different facial expressions. As can be observed, ray-casting mapping always outperforms closest-point mapping. With optimal landmark combination and ray-casting, 95.5% correct recognition rate is achieved.

The effect of automatic landmarking is investigated in the last experiment. Table 6.15 compares the performance of the proposed face recognition algorithm on the 3D face images that are registered using manually labeled landmark points, automatically via ICP and using automatically detected landmark points. The depth images were generated by ray-casting and DCT-4 feature set was used for classification. The results obtained using the automatic registration methods

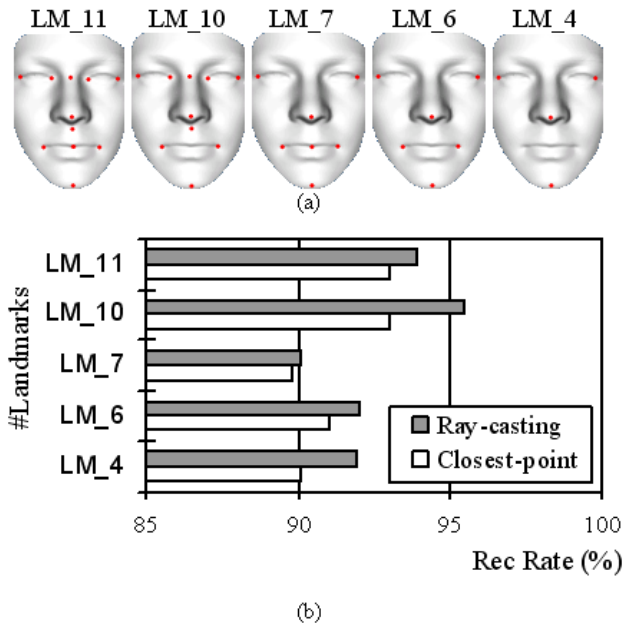



Figure 6.16: (a) Five landmark combinations. (b) Recognition rate of DCT-based local depth approach with different landmark combinations.

-with ICP and with automatically detected landmark points-, are slightly lower than the results obtained on the images that are registered using the manual labels. This decrease in the performance is mainly caused by the errors introduced by ICP registration and automatic landmark detection. Better results were attained on the images that are registered using automatically detected landmark points than on the ones registered via ICP. This indicates that deformation onto a common frame is able to mitigate the effects of expression variations.

| Registration Method | Recognition Rate |
|---|---|
| Manual LM_10 | 95.5% |
| ICP | 92.7% |
| Automatic LM_10 | 93.1% |

Table 6.15: Manual registration vs. automatic registration.

### Analysis of Different Bases

In this part of the experiments, the discrete cosine transform-based local depth representation is compared with different well-known basis functions that can also be used for representing the local regions. In addition, the proposed local depth-based approach is compared with the discrete cosine transform-based

holistic approach. In these experiments, the 3D faces were registered using ten manually labeled landmark points and depth images were generated using ray-casting.

The discrete cosine transform is compared with Karhunen-Love transform (KLT), Walsh-Hadamard transform (WHT), Fourier transform (FT), and wavelet transform (WT). In KLT, 20-dimensional local feature vectors are used; in WHT the same feature setup as the one used for DCT is used; in FT, the magnitudes of the Fourier coefficients are utilized. For wavelet transform, the Daubechies 4 wavelet is used, which has been shown to perform better in terms of computation time and recognition performance with respect to the other order Daubechies wavelets and other well-known wavelets [ES05]. The first-order scaling component is utilized as the feature vector [ES05]. Table 6.16 gives the correct recognition rates obtained with each basis function. As can be seen, DCT achieves the best result. After DCT, WHT also reaches high correct recognition rate compared to the other basis functions. The other basis functions attain lower performance although they use higher dimensional feature vectors. The feature dimension is 20 in KLT, 64 in FT, and 16 in WT, whereas it is 10 in DCT and WHT.

The comparison of DCT-based local and holistic 3D face recognition is given in Table 6.17. In the holistic approach, the same dimensional feature vector is selected for the entire image using the same feature selection strategy as the one used for local depth-based approach, that is, removing the first four DCT coefficients and selecting the remaining first 640 DCT coefficients that are ordered according to the zig-zag pattern. The results show the importance of applying DCT locally and then combining the local analysis results in order to construct the overall feature vector.

The results from Table 6.16 and 6.17 indicate that the obtained performance improvement with the proposed algorithm is not solely based on doing classification in the frequency domain or using the discrete cosine transform. For instance, Fourier transform-based local depth and the discrete cosine transfom-based holistic 3D face recognition approaches have been found to perform poorly. This shows that the performance improvement is provided by performing local analysis and using the discrete cosine transform to represent the local regions.

| Method | Performance |
|---|---|
| DCT | 95.5% |
| KLT | 84.0% |
| WHT | 92.4% |
| FT | 80.6% |
| WT | 74.5% |

Table 6.16: Performance comparison of local depth representation methods.

| Method | Performance |
|---|---|
| Local DCT | 95.5% |
| Holistic DCT | 84.0% |

Table 6.17: Local DCT vs. Holistic DCT.

**Performance Comparison**

In this part of the experiments, the proposed local depth-based 3D face recognition approach is compared with several well-known face recognition algorithms: eigenfaces [TP91], linear discriminant analysis (LDA) [BHK97], Bayesian face recognition [MJP00], embedded hidden Markov model (EHMM) [Nef99], point set difference (PSD) [İGA04], and point distribution model (PDM) [CTCG95]. For the local depth-based approach, an SVM classifier is also used instead of a nearest neighbor classifier to assess the performance of a more sophisticated classification scheme.

Table 6.18 shows the experimental results of each algorithm. Both of the correct recognition rates attained on manually and automatically registered images are given. In eigenfaces, Bayesian face recognition, and PDM algorithms, 100 principal components are used. This is the number of principal components with which the best results are achieved. For LDA, the LDA+PCA algorithm provided in the CSU face identification evaluation system [CSU09] is used. This version of LDA uses a soft distance measure proposed by Zhao et al. [ZCP99]. In EHMM, a $4 \times 4$ size DCT coefficients matrix is used as HMM observation, which is extracted from a $12 \times 12$ image block by using DCT. In local depth-based 3D face recognition, both for nearest neighbor and for SVM classification, the ten-dimensional DCT-4 feature set is used. Radial basis function was the kernel function in the SVM classifier.

From the results given in Table 6.18 it can be observed that the proposed local depth-based approach outperforms the other well-known face recognition algorithms as well as the local DCT features classified with SVM, which may suffer from a small training set problem. The performance of all algorithms decreases slightly when they use the images that are registered using automatically detected landmarks. These results indicate that the proposed local DCT features provide a powerful and robust representation of depth images for classification purposes.

| Method | Manual LM_10 | Automatic LM_10 |
|---|---|---|
| Local DCT | 95.5% | 93.1% |
| Local DCT + SVM | 90.0% | 89.0% |
| EHMM | 87.9% | 85.5% |
| Eigenfaces | 88.6% | 86.5% |
| LDA | 92.4% | 88.5% |
| Bayesian | 94.9% | 89.7% |
| PSD | 81.4% | 80.6% |
| PDM | 87.6% | 84.7% |

Table 6.18: Performance comparison of 3D face recognition methods with manual and automatic landmark-based registration.

# 7 Conclusions

In this thesis a novel face recognition algorithm that is able to work under different conditions is presented. It is the first time in the literature that a generic algorithm can handle facial appearance variations without having specific modifications for each variation. The algorithm is based on appearances of local facial regions that are represented with discrete cosine transform coefficients. The local representation provides robustness against appearance variations in local regions caused by factors such as facial occlusion or expression, whereas automatic frequency band selection provides robustness against changes in illumination. Moreover, different from the traditional face recognition systems, the algorithm does not need a facial feature detection step for face registration.

The algorithm has been tested extensively on 15 different training-testing conditions using well-known face recognition benchmark databases. First, the parameters of the LAFR approach —feature normalization, block size, region partitioning, and local appearance representation— are analyzed. It has been found that feature normalization has paramount importance in local appearance-based face recognition. Balancing the impact of the coefficients on classification by dividing them by their standard deviations and then balancing the impact of different blocks on classification by normalizing the local feature vector to unit norm has improved the performance significantly. For example, in the *FRGC4* experiment, with feature normalization the obtained correct classification rate is 90.8%, whereas without doing any feature normalization it is 63.2%. Experimental results have shown that $8 \times 8$ pixels local block size for an input image size of $64 \times 64$ provides the best recognition performance. It is observed that generic partitioning provides higher correct recognition rates than salient region-based partitioning approaches, which indicates that there is no need to focus on salient regions and perform salient region-based partitioning. Discrete cosine transform is found to be superior in representing the local facial regions compared to principal component analysis, wavelet transform, Fourier transform, and Walsh-Hadamard transform, in terms of face recognition performance.

The robustness of the proposed algorithm is assessed against compression and registration errors. Afterwards, a novel face registration approach is introduced, which does not require a facial feature localization step. The proposed approach formulates the registration problem as an optimization process, in which the closest classification distance is minimized. It has been shown once more that

registration plays a very crucial role in appearance-based face recognition. The following points have been observed in the experiments:

- The facial feature localization step can be eliminated from the face recognition systems and the alignment can be performed by directly aiming at minimizing the closest classification distance.
- The proposed registration approach performs even better than doing registration with manual labels. For instance, on the AR face database [MB98], against lower facial occlusion, the obtained result when the test images are aligned using the manual labels was 91.8%, while with the proposed registration approach it has become 97.3%.
- The main problem with the upper face occlusion is due to registration errors and not the occlusion itself. Due to the sunglasses, the eye center points that are widely used for face alignment can not be reliably labeled even manually. When only the manual labels are used to align the test images, the achieved correct recognition rate against upper facial occlusion with sunglasses is 38.2% on the AR face database [MB98]. When the proposed registration approach is applied, the performance jumps to 97.3%.
- The optimization procedure integrated to the classification step makes the face recognition system insensitive to facial feature localization errors. The algorithm can tolerate up to 18% of the interocular distance as localization error and up to this point it provides stable performance.

The effect of frequency bands is also analyzed and an automatic frequency band selection method is proposed. Significant improvements have been achieved with the use of the proposed automatic frequency selection scheme. For example in the *Yale4* experiment, the correct recognition rate increased from 95.6% to 100%, similarly in the *Yale5* experiment, it increased from 96.8% to 100%.

The LAFR approach has been compared with well-known face recognition algorithms and found to be significantly superior to them. Furthermore, the attained results are as good as or even better than the state-of-the-art algorithms specifically developed for illumination or occlusion variations. The algorithm is extended for 3D face recognition and promising results are obtained. The proposed face recognition system has also been combined with a speaker identification system for multimodal person identification. An adaptive modality weighting scheme, named as cumulative ratio of correct matches (CRCM), where modalities are weighted according to their confidence in their decisions, is developed. Several real world systems —a door monitoring system, a visitor interface, face recognition for humanoid robots, person identification in movies— have been developed, which are shown to work robustly under challenging real-world conditions.

# Bibliography


[AHP06] T. Ahonen, A. Hadid, and M. Pietikäinen, "Face description with local binary patterns: application to face recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, no. 12, pp. 2037–2041, 2006.

[AMU97] Y. Adini, Y. Moses, and S. Ullman, "Face recognition: The problem of compensating for changes in illumination direction," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, no. 7, pp. 721–732, 1997.

[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*, vol. 19, no. 7, pp. 711–720, 1997.

[BP93] R. Brunelli and T. Poggio, "Face recognition: Features versus templates," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 13, no. 10, pp. 1042–1052, 1993.

[CEW05] W. Chen, M. J. Er, and S. Wu, "PCA and LDA in DCT domain," *Pattern Recognition Letters*, vol. 26, no. 15, pp. 2474–2482, 2005.

[CEW06] W. Chen, M. J. Er, and S. Wu, "Illumination compensation and normalization for robust face recognition using discrete cosine transform in logarithm domain," *IEEE Transactions on Systems, Man, and Cybernetics–Part B: Cybernetics*, vol. 36, no. 2, pp. 458–466, 2006.

[CSU09] CSU, "The CSU face identification evaluation system:," 2009, 2009. [Online]. Available: http://www.cs.colostate.edu/evalfacerec/

[CTCG95] T. F. Cootes, C. J. Taylor, D. H. Cooper, and J. Graham, "Active shape models-their training and application," *Computer Vision and Image Understanding*, vol. 61, no. 1, pp. 38–59, 1995.

[CWX+06] T. Chen, Y. Wotao, S. Z. Xiang, D. Comaniciu, and T. Huang, "Total variation models for variable lighting face recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, no. 9, pp. 1519–1524, 2006.

[ECW05] M. J. Er, W. Chen, and S. Wu, "High-speed face recognition based on discrete cosine transform and RBF neural networks," *IEEE Transactions on Neural Networks*, vol. 16, no. 3, pp. 679–691, 2005.

[EFG+07] H. K. Ekenel, M. Fischer, H. Gao, K. Kilgour, J. S. Marcos, and R. Stiefelhagan, "Universität Karlsruhe (TH) at TRECVID 2007," in *Proc. of TRECVID Evaluation Workshop*, 2007.

[EFS07] H. K. Ekenel, M. Fischer, and R. Stiefelhagen, "Face recognition in smart rooms," in *Proc. of Machine Learning for Multimodal Interaction (MLMI)*, 2007, pp. 120–131.

[EJFS07] H. K. Ekenel, Q. Jin, M. Fischer, and R. Stiefelhagen, "ISL person identification systems in CLEAR 2007," in *Proc. of CLEAR Evaluation Workshop*, 2007, pp. 256–265.

[EMR00] S. Eickeler, S. Muller, and G. Rigoll, "Recognition of JPEG compressed face images based on statistical methods," *Image and Vision Computing*, vol. 18, no. 4, pp. 279–287, 2000.

[EP06] H. K. Ekenel and A. Pnevmatikakis, "Video-based face recognition evaluation in the CHIL project -Run 1," in *Proc. of IEEE Intl. Conf. on Automatic Face and Gesture Recognition*, 2006, pp. 85–90.

[ES04] H. K. Ekenel and B. Sankur, "Feature selection in the independent component subspace for face recognition," *Pattern Recognition Letters*, vol. 25, no. 12, pp. 1377–1388, 2004.

[ES05] H. K. Ekenel and B. Sankur, "Multiresolution face recognition," *Image and Vision Computing*, vol. 23, no. 5, pp. 469–477, 2005.

[ES06] H. K. Ekenel and R. Stiefelhagen, "Analysis of local appearance-based face recognition: Effects of feature selection and feature normalization," in *Conf. on Computer Vision and Pattern Recognition Workshop 2006*, 2006, p. 34.

[FCH98] B. J. Frey, A. Colmenarez, and T. S. Huang, "Mixtures of local linear subspaces for face recognition," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 1998.

[FSL06] S. Fidler, D. Skočaj, and A. Leonardis, "Combining reconstructive and discriminative subspace methods for robust classification and regression by subsampling," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, no. 3, pp. 337–350, 2006.

[Fur97] S. Furui, "Recent advances in speaker recognition," *Pattern Recognition Letters*, vol. 18, no. 9, pp. 859–872, 1997.

[GA04] R. Gottumukkal and V. K. Asari, "An improved face recognition technique based on modular PCA approach," *Pattern Recognition Letters*, vol. 25, no. 4, pp. 429–436, 2004.

[GBK01] A. Georghiades, P. Belhumeur, and D. Kriegman, "From few to many: Illumination cone models for face recognition under variable lighting and pose," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 6, pp. 643–660, 2001.

[GMB02] R. Gross, I. Matthews, and S. Baker, "Fisher light-fields for face recognition across pose and illumination," in *Proceedings of the* $24^{th}$ *DAGM Symposium on Pattern Recognition*, 2002, pp. 481–489.

[GMB04] R. Gross, I. Matthews, and S. Baker, "Appearance-based face recognition and light-fields," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 26, no. 4, pp. 449–465, 2004.

[GSC01] R. Gross, J. Shi, and J. F. Cohn, "Que vadis face recognition?" in *Proceedings of Workshop on Empirical Evaluation Methods in Computer Vision*, 2001.

[GW01] R. Gonzales and R. E. Woods, *Digital Image Processing.* Prentice Hall, 2001.

[HHWP03] B. Heisele, P. Ho, J. Wu, and T. Poggio, "Face recognition: component-based versus global approaches," *Computer Vision and Image Understanding*, vol. 91, no. 1-2, pp. 6–21, 2003.

[HL01] Z. M. Hafed and M. D. Levine, "Face recognition using the discrete cosine transform," *International Journal of Computer Vision*, vol. 43, no. 3, pp. 167–188, 2001.

[HSP07] B. Heisele, T. Serre, and T. Poggio, "A component-based framework for face detection and identification," *International Journal of Computer Vision*, vol. 74, no. 2, pp. 167–181, 2007.

[HYH+05] X. He, S. Yan, Y. Hu, P. Niyogi, and H. J. Zhang, "Face recognition using Laplacianfaces," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 3, pp. 328–340, 2005.

[İGA04] M. O. İrfanoğlu, B. Gökberk, and L. Akarun, "Dense surface point distribution models of the human face," in *Proceedings of International Conference on Pattern Recognition*, 2004, pp. 183–186.

[Int08] Intel Corporation, "Open source computer vision library (OpenCV)," 2008, last visit: Nov. 2008. [Online]. Available: http://www.intel.com/technology/computing/opencv/index.htm

[JM08] H. Jia and A. M. Martinez, "Face recognition with occlusions in the training and testing sets," in *Proceedings of IEEE Int'l. Conf. on Automatic Face and Gesture Recognition*, 2008.

[JPS06] Q. Jin, Y. Pan, and T. Schultz, "Far-field speaker recognition," in *Proceedings of International Conference on Acoustic, Speech, and Signal Processing, ICASSP 2006*, 2006, pp. 937–940.

[KD98] V. V. Kohir and U. B. Desai, "Face recognition using a DCT-HMM approach," in *Proceedings of the Fourth IEEE Workshop on Applications of Computer Vision*, 1998, pp. 226–231.

[KHDM98] J. Kittler, M. Hatef, R. P. W. Duin, and J. Matas, "On combining classifiers," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 3, pp. 226–239, 1998.

[KJK02] K. I. Kim, K. Jung, and H. J. Kim, "Face recognition using kernel principal component analysis," *IEEE Signal Processing Letters*, vol. 9, no. 2, pp. 40–42, 2002.

[KKB02] H. C. Kim, D. Kim, and S. Y. Bang, "Face recognition using the mixture-of-eigenfaces method," *Pattern Recognition Letters*, vol. 23, no. 13, pp. 1549–1558, 2002.

[KKB03] H. C. Kim, D. Kim, and S. Y. Bang, "Face recognition using LDA mixture model," *Pattern Recognition Letters*, vol. 24, no. 15, pp. 2815–2821, 2003.

[KKHK05] T. K. Kim, H. Kim, W. Hwang, and J. Kittler, "Component-based LDA face description for image retrieval and MPEG-7 standardisation," *Image and Vision Computing*, vol. 23, no. 7, pp. 631–642, 2005.

[LC04] S. Lucey and T. Chen, "A GMM parts based face representation for improved verification through relevance adaptation," in *Proc. of IEEE Int'l. Conf. on Computer Vision and Pattern Recognition*, 2004, pp. 856–861.

[LCLZ07] S. Z. Li, R. Chu, S. Liao, and L. Zhang, "Illumination invariant face recognition under near-infrared images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 4, pp. 627–639, 2007.

[LHK05] K. Lee, J. Ho, and D. Kriegman, "Acquiring linear subspaces for face recognition under variable lighting," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 5, pp. 684–698, 2005.

[LKK05] H. J. Lee, H. J. Kim, and W. Y. Kim, "Face recognition using component-based DCT/LDA," in *Proc. of IWAIT*, 2005.

[LP02] A. Lemieux and M. Parizeau, "Experiments on eigenfaces robustness," in *Proc. Int'l. Conf. on Pattern Recognition*, vol. 1, 2002, pp. 421–424.

[Mar02] A. M. Martinez, "Recognizing imprecisely localized, partially occluded and expression variant faces from a single sample per class," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 6, pp. 748–763, 2002.

[MB94] S. Manke and U. Bodenhausen, "A connectionist recognizer for online cursive handwriting recognition," in *Proc. of Int'l. Conference on Acoustics, Speech, and Signal Processing (ICASSP'94)*, Adelaide, Australia, 1994.

[MB98] A. M. Martinez and R. Benavente, "The AR face database," CVC, Tech. Rep. 24, 1998.

[MJP00] B. Moghaddam, T. Jebara, and A. Pentland, "Bayesian face recognition," *Pattern Recognition*, vol. 33, no. 11, pp. 1771–1782, 2000.

[Nef99] A. Nefian, "A hidden Markov model-based approach for face detection and recognition," Ph.D. dissertation, Georgia Institute of Technology, USA, 1999.

[NS07] K. Nickel and R. Stiefelhagen, "Fast audio-visual multi-person tracking for a humanoid stereo camera head," in *Proc. of IEEE-RAS Intl. Conference on Humanoid Robots*, 2007.

[PB99] Z. Pan and H. Bolouri, "High speed face recognition based on discrete cosine transforms and neural networks," University of Hertfordshire, UK, Tech. Rep., 1999.

[PFS+05] P. J. Phillips, P. J. Flynn, T. Scruggs, K. W. Bowyer, J. Chang, K. Hoffman, J. Marques, J. Min, and W. Worek, "Overview of the Face Recognition Grand Challenge," in *Proc. 2005 IEEE Computer Society Conf. on Computer Vision and Pattern Recognition*, vol. 1, Los Alamitos, CA, USA, 2005, pp. 947–954.

[PLL05] B. G. Park, K. M. Lee, and S. U. Lee, "Face recognition using face-ARG matching," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 12, pp. 1982–1988, 2005.

[PMRR00] P. J. Phillips, H. Moon, S. A. Rizvi, and P. J. Rauss, "The FERET evaluation methodology for face recognition algorithms," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 10, pp. 1090–1104, 2000.

[PMS94] A. Pentland, B. Moghaddam, and T. Starner, "View-based and modular eigenspaces for face recognition," in *Proc. of IEEE CVPR*, 1994, pp. 84–91.

[PS01] J. Pelecanos and S. Sridharan, "Feature warping for robust speaker verification," in *Proceedings of Speaker Odyssey Conference*, 2001.

[PZ96] C. Podilchuk and X. Zhang, "Face recognition using DCT-based feature vectors," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 4, 1996, pp. 2144–2147.

[Rey95] D. A. Reynolds, "Speaker identification and verification using gaussian mixture speaker models," *Speech Communication*, vol. 17, no. 1–2, pp. 91–108, 1995.

[RSPP06] E. Rentzeperis, A. Stergiou, A. Pnevmatikakis, and L. Polymenakos, "Impact of face registration errors on recognition," in *Artificial Intelligence Applications and Innovations (AIAI06)*. Springer, 2006, pp. 187–194.

[Sam94] F. Samaria, "Face recognition using hidden markov models," Ph.D. dissertation, University of Cambridge, UK, 1994.

[SBB03] T. Sim, S. Baker, and M. Bsat, "The CMU pose, illumination, and expression database," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 25, no. 12, pp. 1615–1618, 2003.

[SBB+07] R. Stiefelhagen, K. Bernardin, R. Bowers, R. T. Rose, M. Michel, and J. Garofolo, "The CLEAR 2007 evaluation," in *Proc. of the International Evaluation Workshops CLEAR 2007 and RT 2007*, 2007, pp. 3–34.

[SBE+06] R. Stiefelhagen, K. Bernardin, H. K. Ekenel, J. McDonough, K. Nickel, M. Voit, and M. Wölfel, "Audio-visual perception of a lecturer in a smart seminar room," *Signal Processing*, vol. 86, no. 12, pp. 3518–3533, 2006.

[SCG+04] S. Shan, Y. Chang, W. Gao, B. Cao, and P. Yang, "Curse of misalignment in face recognition: Problem and a novel mis-alignment learning solution," in *Proc. of the* $6^{th}$ *IEEE International Conference on Automatic Face and Gesture Recognition (FGR'04)*, 2004, pp. 314–320.

[Sco03] W. L. Scott, "Block-level discrete cosine transform coefficients for autonomic face recognition," Ph.D. dissertation, Louisiana State University, USA, May 2003.

[SEE+07] J. Stallkamp, H. K. Ekenel, H. Erdoğan, R. Stiefelhagen, and A. Erçil, "Video-based driver identification using local appearance face recognition," in *Proc. of Workshop on DSP in Mobile and Vehicular Systems*, 2007.

[SEF+07] R. Stiefelhagen, H. K. Ekenel, C. Fügen, P. Gieselmann, H. Holzapfel, F. Kraft, K. Nickel, M. Voit, and A. Waibel, "Enabling multimodal human-robot interaction for the Karlsruhe humanoid robot," *IEEE Transactions on Robotics*, vol. 23, no. 5, pp. 840–851, 2007.

[SES07] J. Stallkamp, H. K. Ekenel, and R. Stiefelhagen, "Video-based face recognition on real-world data," in *Proc. of IEEE International Conference on Computer Vision, ICCV 2007*, 2007, pp. 1–8.

[SH94] F. S. Samaria and A. C. Harter, "Parameterisation of a stochastic model for human face identification," in *Proceedings of the Second IEEE Workshop on Applications of Computer Vision*, 1994, pp. 138–142.

[SK01] T. Sim and T. Kanade, "Combining models and exemplars for face recognition: An illumination example," in *Proceedings of Workshop on Models versus Exemplars in Computer Vision*, 2001.

[SP03a] C. Sanderson and K. K. Paliwal, "Features for robust face-based identity verification," *Signal Processing*, vol. 83, no. 5, pp. 931–940, 2003.

[SP03b] C. Sanderson and K. K. Paliwal, "Noise compensation in a person verification system using face and multiple speech features," *Pattern Recognition*, vol. 36, no. 2, pp. 293–302, 2003.

[SRR01] A. Shashua and T. Riklin-Raviv, "The quotient image: Class-based re-rendering and recognition with varying illuminations," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 2, pp. 129–139, 2001.

[SUM+05] R. Snelick, U. Uludag, A. Mink, M. Indovina, and A. K. Jain, "Large-scale evaluation of multimodal biometric authentication using state-of-the-art systems," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 3, pp. 450–455, 2005.

[TC02] D. S. Turaga and T. Chen, "Face recognition using mixtures of principal components," in *Proceedings of IEEE International Conference on Image Processing*, 2002.

[TCZZ05] X. Tan, S. Chen, Z. H. Zhou, and F. Zhang, "Recognizing partially occluded, expression variant faces from single training image per person with SOM and soft $k$-NN ensemble," *IEEE Transactions on Neural Networks*, vol. 16, no. 4, pp. 875–886, 2005.

[TLV04] R. Tjahyadi, W. Liu, and S. Venkatesh, "Application of the DCT energy histogram for face recognition," in *Proceedings of the* $2^{nd}$ *International Conference on Information Technology and Applications*, 2004, pp. 70–75.

[TP91] M. Turk and A. Pentland, "Eigenfaces for recognition," *Journal of Cognitive Neuroscience*, vol. 3, no. 1, pp. 71–86, 1991.

[VJ04] P. Viola and M. J. Jones, "Robust real-time face detection," *International Journal of Computer Vision*, vol. 57, no. 2, pp. 137–154, 2004.

[VT02] M. A. O. Vasilescu and D. Terzopoulos, "Multilinear image analysis for face recognition," in *Proceedings of the* $16^{th}$ *International Conference on Pattern Recognition*, 2002, pp. 511–514.

[WFKM97] L. Wiskott, J. M. Fellous, N. Kruger, and C. Malsburg, "Face recognition by elastic bunch graph matching," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, no. 7, pp. 775–779, 1997.

[WGYM07] J. Wright, A. Ganesh, A. Yang, and Y. Ma, "Robust face recognition via sparse representation," University of Illinois, USA, Tech. Rep., 2007.

[WHH+89] A. Waibel, T. Hanazawa, G. Hinton, K. Shikano, and K. J. Lang, "Phoneme recognition using time-delay neural networks," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 37, no. 3, pp. 328–339, 1989.

[WLWZ03] H. Wang, S. Z. Li, Y. Wang, and W. Zhang, "Illumination modeling and normalization for face recognition," in *Proc. of the IEEE International Workshop on Analysis and Modeling of Faces and Gestures*, 2003.

[WSS04] A. Waibel, H. Steusloff, and R. Stiefelhagen, "CHIL - computers in the human interaction loop," in *Proc. of Int'l. Workshop on Image Analysis for Multimedia Interactive Services*, Lisbon, Portugal, 2004.

[WYH+08] H. Wang, S. Yan, T. Huang, J. Liu, and X. Tang, "Misalignment-robust face recognition," in *Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2008)*, 2008, pp. 1–6.

[XCN+02] B. Xiang, U. V. Chaudhari, J. Navratil, G. N. Ramaswamy, and R. A. Gopinath, "Short-time gaussianization for robust speaker verification," in *Proceedings of International Conference on Acoustic, Speech, and Signal Processing, ICASSP 2002*, vol. 1, 2002, pp. 681–684.

[YAK00] M. H. Yang, N. Ahuja, and D. Kriegman, "Face recognition using kernel eigenfaces," in *Proceedings of IEEE International Conference on Image Processing*, 2000, pp. 37–40.

[Yan02] M. H. Yang, "Kernel eigenfaces vs. kernel Fisherfaces: Face recognition using kernel methods," in *Proceedings of the fifth IEEE International Conference on Automatic Face and Gesture Recognition*, 2002, pp. 215–221.

[YDB02] W. S. Yambor, B. A. Draper, and J. R. Beveridge, "Analyzing PCA-based face recognition algorithms: Eigenvector selection and distance measures," in *Proceedings of Workshop on Emprical Evaluation Methods in Computer Vision*, 2002.

[ZACJ07] S. K. Zhou, G. Aggarwal, R. Chellappa, and D. W. Jacobs, "Appearance characterization of linear lambertian objects, generalized photometric stereo, and illumination invariant face recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 2, pp. 230–245, 2007.

[ZCP99] W. Zhao, R. Chellappa, and P. J. Phillips, "Subspace linear discriminant analysis for face recognition," UMD, Tech. Rep. TR4009, 1999.

[ZCPR03] W. Zhao, R. Chellappa, P. J. Phillips, and A. Rosenfeld, "Face recognition: A literature survey," *ACM Computing Surveys*, vol. 35, no. 4, pp. 399–458, 2003.

[ZR04] S. Zhao and R. R.Grigat, "Multiblock-fusion scheme for face recognition," in *Proceedings of ICPR 2004*, 2004.

# Publications


[BES06] K. Bernardin, H. K. Ekenel, and R. Stiefelhagen, "Multimodal identity tracking in a smartroom," in *Proceedings of the* $3^{rd}$ *IFIP Conference on Artificial Intelligence Applications & Innovations (AIAI)*, 2006.

[DFB+05] M. Danninger, G. Flaherty, K. Bernardin, H. K. Ekenel, T. Köhler, R. Malkin, R. Stiefelhagen, and A. Waibel, "The Connector — facilitating context-aware communication," in *Proceedings of the* $7^{th}$ *International Conference on Multimodal Interfaces*, 2005.

[EEE+05] H. Erdoğan, A. Erçil, H. K. Ekenel, S. Y. Bilgin, İ. Eden, and M. Kirişçi, "Mutlimodel person recognition for vehicular applications," in *Proceedings of the* $6^{th}$ *International Workshop on Multiple Classifier Systems*, 2005.

[EFG+07] H. K. Ekenel, M. Fischer, H. Gao, K. Kilgour, J. S. Marcos, and R. Stiefelhagan, "Universität Karlsruhe (TH) at TRECVID 2007," in *Proc. of TRECVID Evaluation Workshop*, 2007.

[EFJS07] H. K. Ekenel, M. Fischer, Q. Jin, and R. Stiefelhagen, "Multi-modal person identification in a smart environment," in *Proceedings of Intl. Conf. on Computer Vision and Pattern Recognition Workshop 2007*, 2007.

[EFS07] H. K. Ekenel, M. Fischer, and R. Stiefelhagen, "Face recognition in smart rooms," in *Proc. of Machine Learning for Multimodal Interaction (MLMI)*, 2007, pp. 120–131.

[EGS07] H. K. Ekenel, H. Gao, and R. Stiefelhagan, "3-D face recognition using local appearance-based models," *IEEE Transactions on Information Forensics and Security*, vol. 2, no. 3, pp. 630–635, 2007.

[EJ06] H. K. Ekenel and Q. Jin, "ISL person identification systems in the CLEAR evaluations," in *Proceedings of the CLEAR Evaluation Workshop*, 2006.

[EJFS07] H. K. Ekenel, Q. Jin, M. Fischer, and R. Stiefelhagen, "ISL person identification systems in CLEAR 2007," in *Proc. of CLEAR Evaluation Workshop*, 2007, pp. 256–265.

[EP06] H. K. Ekenel and A. Pnevmatikakis, "Video-based face recognition evaluation in the CHIL project —Run 1," in *Proc. of IEEE Intl. Conf. on Automatic Face and Gesture Recognition*, 2006, pp. 85–90.

[ES04] H. K. Ekenel and B. Sankur, "Feature selection in the independent component subspace for face recognition," *Pattern Recognition Letters*, vol. 25, no. 12, pp. 1377–1388, 2004.

[ES05a] H. K. Ekenel and B. Sankur, "Multiresolution face recognition," *Image and Vision Computing*, vol. 23, no. 5, pp. 469–477, 2005.

[ES05b] H. K. Ekenel and R. Stiefelhagen, "A generic face representation approach for local appearance based face verification," in *Proceedings of the IEEE CVPR Workshop on Face Recognition Grand Challenge Experiments*, 2005.

[ES05c] H. K. Ekenel and R. Stiefelhagen, "Local appearance-based face recognition using discrete cosine transform," in *Proceedings of the* $13^{th}$ *European Signal Processing Conf. (EUSIPCO 2005)*, 2005.

[ES06a] H. K. Ekenel and R. Stiefelhagen, "Analysis of local appearance-based face recognition: Effects of feature selection and feature normalization," in *Proceedings of Intl. Conf. on Computer Vision and Pattern Recognition Workshop 2006*, 2006, p. 34.

[ES06b] H. K. Ekenel and R. Stiefelhagen, "Block selection in the local appearance-based face recognition scheme," in *Conf. on Computer Vision and Pattern Recognition Workshop 2006*, 2006, p. 43.

[ES07] H. K. Ekenel and R. Stiefelhagen, "An un-awarely collected real world face database: The ISL-Door face database," in *Proceedings of the International Conference on Computer Vision Systems*, 2007.

[ESG+07] H. K. Ekenel, J. Stallkamp, H. Gao, M. Fischer, and R. Stiefelhagen, "Face recognition for smart interactions," in *Proc. of International Conference on Multimedia and Expo*, 2007, pp. 1017–1010.

[GNE+05] T. Gehrig, K. Nickel, H. K. Ekenel, U. Klee, and J. McDonough, "Kalman filters for audio-video source localization," in *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 2005.

[HSE+06] H. Holzapfel, T. Schaaf, H. K. Ekenel, C. Shaa, and A. Waibel, "A robot learns to know people —first contacts of a robot," in *Proceedings of the* $29^{th}$ *German Conference on Artificial Intelligence (KI2006)*, 2006.

[KHEW07] S. Könn, H. Holzapfel, H. K. Ekenel, and A. Waibel, "Integrating face-ID into an interactive person-ID learning system," in *Proceedings of the International Conference on Computer Vision Systems*, 2007.

[NEVS06] K. Nickel, H. K. Ekenel, M. Voit, and R. Stiefelhagen, "Audio-visual perception of humans for a humanoid robot," in *Proceedings of the* $2^{nd}$ *International Workshop on Human-Centered Robotics Systems*, 2006.

[NGE+06] K. Nickel, T. Gehrig, H. K. Ekenel, J. McDonough, and R. Stiefelhagen, "An audio-visual particle filter for speaker tracking on the CLEAR06 evaluation dataset," in *Proceedings of the CLEAR Evaluation Workshop*, 2006.

[SBE+06] R. Stiefelhagen, K. Bernardin, H. K. Ekenel, J. McDonough, K. Nickel, M. Voit, and M. Wölfel, "Audio-visual perception of a lecturer in a smart seminar room," *Signal Processing*, vol. 86, no. 12, pp. 3518–3533, 2006.

[SEE+07] J. Stallkamp, H. K. Ekenel, H. Erdoğan, R. Stiefelhagen, and A. Erçil, "Video-based driver identification using local appearance face recognition," in *Proc. of Workshop on DSP in Mobile and Vehicular Systems*, 2007.

[SEF+07] R. Stiefelhagen, H. K. Ekenel, C. Fügen, P. Gieselmann, H. Holzapfel, F. Kraft, K. Nickel, M. Voit, and A. Waibel, "Enabling multimodal human-robot interaction for the Karlsruhe humanoid robot," *IEEE Transactions on Robotics*, vol. 23, no. 5, pp. 840–851, 2007.

[SES07] J. Stallkamp, H. K. Ekenel, and R. Stiefelhagen, "Video-based face recognition on real-world data," in *Proc. of IEEE International Conference on Computer Vision, ICCV 2007*, 2007, pp. 1–8.

[WE05] M. Wölfel and H. K. Ekenel, "Feature weighted mahalanobis distance: Improved robustness for Gaussian classifiers," in *Proceedings of the* $13^{th}$ *European Signal Processing Conf. (EUSIPCO 2005)*, 2005.