N° d'ordre : ..../..../....

الجمهـــورية الجـــزَائرية الديمـــقراطية الشّعـــــبية

RÉPUBLIQUE ALGÉRIENNE DÉMOCRATIQUE ET POPULAIRE

وزَارة التـعليم العـالي وَ البـحث العـلمي

MINISTÈRE DE L'ENSEIGNEMENT SUPÉRIEUR ET DE LA RECHERCHE SCIENTIFIQUE

جـامعة محمد خيضـر بـسكـرة

UNIVERSITÉ MOHAMED KHIDER BISKRA

كليـة العلـوم الدّقيقـة وَ علـوم الطّبيعـة وَ الحيـاة

Faculté des Sciences Exactes et des Sciences de La Nature et de la Vie

قسـم الإعـلام الآلـي

Département D'informatique

مخبـر الذّكـاء المعـلومـاتي

Laboratoire de l'INFormatique Intelligente (LINFI)

THÈSE

Présentée en vue de l'obtention du diplôme de

Doctorat 3ème cycle (LMD) en Informatique

Option : **Intelligence Artificielle**

Par : **Aicha KORICHI**

Titre

Recognizing Visual Object Using Machine Learning Techniques

Soutenue publiquement le : 29/05/2022 devant le jury composé de:

Président	*Pr. K. REZEG*	*Université de Biskra*
Directrice de la Thèse	*Dr. S. SLATNIA*	*Université de Biskra*
Examinateur	*Pr. B. LEJDEL*	*Université de Eloued*
Examinateur	*Pr. S. BENHARZALLAH*	*Université de Batna 2*
Examinateur	*Pr. L. KAHLOUL*	*Université de Biskra*
Invité	*Dr. O. AIADI*	*Université de Ouargla*

Année Universitaire: 2021-2022

N° d'ordre : ..../..../....

الجمهـــورية الجـــزَائرية الديمـــقراطية الشّعـــــبية

DEMOCRATIC AND POPULAR REPUBLIC OF ALGERIA

وزَارة التـعليم العـالي وَ البـحث العـلمي

MINISTRY OF HIGHER EDUCATION AND SCIENTIFIC RESEARCH

جـامعة محمد خيضـر بـسكـرة

UNIVERSITY OF MOHAMED KHIDER BISKRA

كليـة العلـوم الدّقيقـة وَ علـوم الطّبيعـة وَ الحيـاة

Faculty of Exact Sciences and Natural and Life Sciences

قسـم الإعـلام الآلـي

Computer Science Department

مخبـر الذّكـاء المعـلومـاتي

Intelligent Computing Laboratory (LINFI)

THESIS

Presented with a view to obtaining the

Doctoral degree 3rd cycle (LMD) in Computer Science

Option : **Artificial Intelligence**

By : **Aicha KORICHI**

Title

Recognizing Visual Object Using Machine Learning Techniques

Publicly defended on : 29/05/2022 in front of the jury composed of :

President	*Pr. K. REZEG*	*University of Biskra*
Supervisor	*Dr. S. SLATNIA*	*University of Biskra*
Examiner	*Pr. B. LEJDEL*	*University of Eloued*
Examiner	*Pr. S. BENHARZALLAH*	*University of Batna 2*
Examiner	*Pr. L. KAHLOUL*	*University of Biskra*
Guest	*Dr. O. AIADI*	*University of Ouargla*

Academic year: 2021-2022


بسم الله الرحمن الرحيم


وقل رب زدني علما


صدق الله العظيم

I dedicate this work to...

My self,

To all those who were giving me any kind of support,

Especially, to my Mother,

To my Father,

To all my family,

Aicha KORICHI

Acknowledgment

First of all, I thank ALLAH the almighty, for allowing me to reach this modest scientific level and for giving me the courage and the patience to carry out the work done in this thesis.

It has been an auspicious journey for me to arrive at this point. I am really shorted of words now, but I take this opportunity to express my sincere thanks to my thesis supervisor ***Dr. SLATNIA Sihem*** *for accepting to supervise me, and for their incontestable instructions, patience, skills, and remarks that have me benefited.*

I would like to express my sincere thanks to ***Dr. AIADI Oussama****, who had a great virtue in accomplishing this thesis. Thank you Dr. Oussama for your support, patience, and encouragement during the realization of this work.*

I would particularly like to thank all the jury members who have accepted to preside and read this work. ***Pr. REZG Khaled*** *from Biskra University as a jury president,* ***Pr. LEJDEL Brahim*** *from Eloued University,* ***Pr. BENHARZALLAH Saber*** *from Batna 2 University,* ***Pr. KAHLOUL Laid*** *from Biskra University.*

I also extend my sincere acknowledgments to the previous LINFI laboratory head ***Pr. KAZAR Okba****, for encouraging me to do research at the LINFI Laboratory and also to providing me with infrastructure and a variety of other resources necessary for my research.*

Finally, my thanks go to all those who contributed in any way to the outcome of this work.

Abstract

Nowadays, Visual Object Recognition (VOR) has received growing interest from researchers and it has become a very active area of research due to its vital applications including handwriting recognition, diseases classification, face identification ..etc. However, extracting the relevant features that faithfully describe the image represents the challenge of most existing VOR systems.

This thesis is mainly dedicated to the development of two VOR systems, which are presented in two different contributions. As a first contribution, we propose a novel generic feature-independent pyramid multilevel (GFIPML) model for extracting features from images. GFIPML addresses the shortcomings of two existing schemes namely multi-level (ML) and pyramid multi-level (PML), while also taking advantage of their pros. As its name indicates, the proposed model can be used by any kind of the large variety of existing features extraction methods. We applied GFIPML for the task of Arabic literal amount recognition. Indeed, this task is challenging due to the specific characteristics of Arabic handwriting. While most literary works have considered structural features that are sensitive to word deformations, we opt for using Local Phase Quantization (LPQ) and Binarized Statistical Image Feature (BSIF) as Arabic handwriting can be considered as texture. To further enhance the recognition yields, we considered a multimodal system based on the combination of LPQ with multiple BSIF descriptors, each one with a different filter size.

As a second contribution, a novel simple yet efficient, and speedy TR-ICANet model for extracting features from unconstrained ear images is proposed. To get rid of unconstrained conditions (e.g., scale and pose variations), we suggested first normalizing all images using CNN. The normalized images are fed then to the TR-ICANet model, which uses ICA to learn filters. A binary hashing and block-wise histogramming are used then to compute the local features. At the final stage of TR-ICANet, we proposed to use an effective normalization method namely Tied Rank normalization in order to eliminate the disparity within block-wise feature vectors. Furthermore, to improve the identification performance of the proposed system, we proposed a softmax average fusing of CNN-based feature extraction approaches with our proposed TR-ICANet at the decision level using SVM classifier.

Keywords: *Visual object recognition, Machine learning, Pattern recognition, Deep learning, Ear recognition, Arabic handwriting recognition*

Résumé

Actuellemnt, la reconnaissance visuelle des objets (Visual Object Recognition -VOR- en Anglais) suscite un intérêt croissant chez les chercheurs et il est devenue un domaine de recherche très actif grace à ses essentielles applications, notamment la reconnaissance de l'écriture manuscrite, la classification des maladies, l'identification des visages, etc. Cependant, l'extraction des caractéristiques pertinentes qui décrivent précisément l'image représente le défi de la plupart des systèmes VOR.

Cette thèse est principalement destinée à développer deux systèmes VOR présentés dans deux contributions différentes. Dans la première contribution, nous proposons un nouveau modèle générique de pyramide multiniveau indépendant des caractéristiques (GFIPML) pour extraire les caractéristiques à partir des images. GFIPML traite les défauts de deux schémas existants, à savoir le multi-niveau (ML) et la pyramide multi-niveau (PML), et donc tire parti de leurs avantages. Comme son nom l'indique, le modèle proposé est utilisable par n'importe quel type de caractéristiques parmi la grande variété des méthodes d'extraction des caractéristiques existantes. Nous avons appliqué le modèle GFIPML à la tâche de reconnaissance des montants littéraux en arabe. En effet, cette tâche est difficile en raison des caractéristiques spécifiques de l'écriture arabe. Alors que la plupart des travaux littéraires ont pris en compte les caractéristiques structurelles qu' ils sont sensibles aux déformations des mots, nous avons choisi alors l'utilisation de la quantification de phase locale (LPQ) et de la caractéristique d'image statistique binarisée (BSIF) car l'écriture arabe peut être considérée comme une texture. Pour améliorer encore les rendements de reconnaissance, nous avons envisagé un système multimodal basé sur la combinaison entre le descripteur LPQ et les multiples caractéristiques BSIF, chacune avec une taille de filtre différente.

La deuxième contribution est un nouveau modèle TR-ICANet simple, efficace et rapide pour extraire les caractéristiques d'image d'oreille sans contrainte. Pour se débarrasser les conditions de contraintes (par exemple, les variations d'échelle et de position), on a suggéré de normaliser d'abord toutes les images en utilisant CNN. Puis, les images normalisées sont ensuite introduites dans le modèle TR-ICANet, qui utilise l'ICA pour apprendre les filtres. Un hachage binaire et un histogramme par blocs sont ensuite utilisés pour calculer les caractéristiques locaux. A la phase final de TR-ICANet, nous avons proposé d'utiliser une méthode de normalisation efficace, à savoir la normalisation Tied Rank, afin d'éliminer la disparité dans les vecteurs de caractéristiques par bloc. En outre, pour améliorer les performances d'identification du système proposé, nous avons proposé une fusion la moyenne softmax des approches d'extraction de caractéristiques basées sur CNN avec TR-ICANet proposé au niveau de la décision en utilisant un classifieur SVM.

Mots clés:*Reconnaissance visuelle des objets, apprentissage automatique, reconnaissance des formes, reconnaissance des oreilles, reconnaissance de l'écriture arabe.*

الملـــــخّص

في الوقت الحاضر، حظي التعرف على الأشياء من خلال محتواها المرئي (VOR) باهتمام متزايد من الباحثين وأصبح مجالًا نشطًا للغاية للبحث بسبب تطبيقاته الحيوية بما في ذلك التعرف على خط اليد، وتصنيف الأمراض، وتحديد الوجه.. إلخ. ومع ذلك، فإن استخراج الميزات ذات الصلة التي تصف الصورة يمثل تحديًا لمعظم أنظمة VOR.

هذه الأطروحة مخصصة بشكل أساسي لتطوير نظامين للتعرف على الأشياء من خلال محتواها المرئي VOR مقدمين في مساهمتين مختلفتين. كمساهمة أولى، نقترح نموذجًا هرميًا متعدد المستويات (GFIPML) جديدًا ومستقلًا عن الميزات العامة لاستخراج الميزات من الصور. يتعامل GFIPML مع أوجه القصور لنموذجين موجودين سابقا هما متعدد المستويات (ML) وهرمي متعدد المستويات (PML) ويستفيد من مزاياهم. كما يشير اسمه، فإن النموذج المقترح قابل للاستخدام بأي نوع من مجموعة كبيرة ومتنوعة من طرق استخراج الميزات الموجودة. طبقنا GFIPML لمهمة التعرف على قيم المبالغ المكتوبة باليد للغة العربية. في الواقع، هذه المهمة صعبة بسبب الخصائص الجوهرية للكتابة اليدوية العربية. في حين أن معظم الأعمال السابقة قد نظرت في السمات الهيكلية التي بدورها الحساسة لتشوهات الكلمات، فإننا نختار استخدام تكميم المرحلة المحلية (LPQ) وميزة الصورة الإحصائية الثنائية (BSIF) حيث يمكن اعتبار الكتابة اليدوية العربية نسيجًا. لزيادة تعزيز عائدات التعرف، اقترحنا نظام متعدد الوسائط يعتمد على مزيج من LPQ مع ميزات BSIF متعددة، ولكل منها حجم مرشح مختلف.

كمساهمة ثانية، تم اقتراح نموذج TR-ICANet جديد بسيط وفعال وسريع لاستخراج الميزات من صور الأذن المأخوذة من ظروف غير مقيدة. للتخلص من الظروف غير المقيدة (على سبيل المثال، اختلافات المقياس والوضعية)، اقترحنا أولاً تطبيع جميع الصور باستخدام .CNN ثم يتم تغذية الصور إلى نموذج TR-ICANet ، الذي يستخدم ICA لتعلم المرشحات. ثم يتم استخدام تجزئة ثنائية ومدرج تكراري للكتل لحساب الميزات المحلية. في المرحلة النهائية من TR-ICANet ، اقترحنا استخدام طريقة تطبيع فعالة وهي تسوية التصنيف المقيد من أجل القضاء على التباين داخل نواقل الميزات الكتلية. علاوة على ذلك، لتحسين أداء تحديد النظام المقترح، اقترحنا متوسط softmax لدمج أساليب استخراج الميزات المستندة إلى CNN مع TR-ICANet المقترح على مستوى القرار باستخدام مصنف SVM.

الكلمات المفتاحية: التعرف المرئي على الأشياء، التعلم الآلي، التعرف على الأنماط، التعلم العميق، التعرف على الأذن، التعرف على الكتابة اليدوية العربية.

Contents

List of Figures

List of Tables

Abbreviations

AHDB	:	Arabic Handwriting Data Basen	AWE	:	Annotated Web Ear
BSIF	:	Binarized Statistical Image Features	CMC	:	Cumulative match characteristic curve
CNN	:	Convolutional Neural Network	DCT	:	Discrete Cosine Transform
DCTNet	:	Discrete Cosine Transform Network	FN	:	False Negative
FNR	:	False Negative Rate	FP	:	False Positive
FPR	:	False Positive Rate	GFIPML	:	Generic Feature-Independent Pyramid Multi-Level
HOG	:	Histogram of Gradient	ICA	:	Independent Components Analysis
ICANet	:	Independent Components Analysis Network	KNN	:	K-Nearest Neighbors
LBP	:	Local Binary Pattern	LDA	:	Linear Discriminant Analyses
LPQ	:	Local Phase Quantization	ML	:	Multi Level
PCA	:	Principal Component Analysis	PCANet	:	Principal Component Analysis Network
PML	:	Pyramid Multi Level	ROC	:	Receiver Operating Characteristic curve
SVM	:	Support Vector Machine	TN	:	True Negative
TP	:	True Positive	TR	:	Tied Rank
VOR	:	Visual Object Recognition		:	

Chapter 1

General Introduction

1.1 Introduction

Due to the revolutionary progress achieved in imaging technology, image acquisition devices, such as smartphones, have become available for most people. Indeed, ordinary persons can easily and rapidly perform different analysis actions on images. However, the performance of human analysis is proportional to the number of images in hand (i,e., in the case of a huge amount of images, human performance declines dramatically). Thus, this task is becoming quite complicated and much expensive in terms of time and effort as well. This heightened the need to develop automatic tools that are capable to perform recognition instead of human beings.

In the last few years, image recognition based on visual content has become a very active topic in the pattern recognition domain and it attracted the attention of many researchers this is because of the increasing amount of images that are publicly available online and the necessity to automatically analyze of them. Visual object recognition refers to the task of automatically understanding a particular object (a scene, a person, a behavior, a manuscript, an emotion, a flower category, a signature, etc), which is contained within an image based on its content. Moreover, it can extend beyond object segmentation, detection, and recognition by attempting to provide a semantic description as well as explain higher-level relationships between objects such as behavior or activity [1, 2].

Machine learning is a branch of artificial intelligence and a subfield of computer science that can be defined as an ensemble of data analysis methods that automate analytical model building by using algorithms that iteratively learn from data. It enables computers to discover hidden insights without having to be explicitly programmed [3]. It is concerned with the development of computer programs that can change in response to new data. Machine learning systems are commonly used to recognize objects in images, transcribe speech to text, match news items, posts, or products to user interests, and select relevant search results [4].

According to the learning methods, machine learning paradigms are broadly divided into three types: Supervised, Unsupervised, and Semi-supervised learning. In the supervised learning approach, the training data are labeled with known labels. However, the data in unsupervised learning are unlabeled and the task is to cluster them. Semi-supervised learning falls somewhere between supervised and unsupervised learning.

1.2 Problematic

Generally, a visual object recognition (VOR) system contains three main stages which are preprocessing, feature extraction, and classification. The first stage (i,e., preprocessing) aims to ameliorate images quality by removing irrelevant information such as noise removal and normalization. However, the relevant characteristics that faithfully describe the images are extracted in the feature extraction stage. The query images are assigned to a specified class in the classification stage according to the decision of the used classifier.

The feature extraction stage is considered a key stage for any VOR system as a good feature extraction technique leads to a good classification and matching. However, even with the critical role of the feature extraction stage in well representing the images, image-preprocessing enhancements could have a good and positive effect on the recognition rates. Thus, automatic VOR performance depends mainly on the used ML techniques for images representation.

In the literature, several important applications of VOR can be found including Handwriting recognition, Biometric recognition, Date fruit recognition.. etc. This thesis is dedicated mainly to dealing with the conception of two different VOR systems which are an automatic Arabic handwriting recognition system and an unconstrained ear identification system. In the rest of this section, we illustrate, justify and demonstrate why we focused our interests on these two problems.

1.2.1 Why Arabic Handwriting Recognition?

Offline handwritten word recognition is defined as the process of translation of a hand word image into a printed word in a digital format usable by the machine. It is still one of the most challenging areas of the pattern recognition field. In recent years, it has received growing interest from researchers and it has become a very active field of research due to its important applications.

Among the other languages, in this thesis, we opt to recognize offline isolated handwriting words from the Arabic language. As the task of recognizing words from Arabic scripts is still a challenging task because of the inherent characteristics of Arabic such as cursiveness,

overlapping characters, presence of diacritical marks, etc. [5]. Moreover, unlike other scripts, Arabic is written from right to left and contains 28 letters. Each letter has more than two different shapes, depending on the letter position (isolated, initial, medial, and final). The presence of all these characteristics makes the recognition task more challenging which motived us to present a robust Arabic handwriting recognition system that overcomes the shortcoming of existing systems and covers the characteristics of the Arabic script.

1.2.2 Why Unconstrained Ear Recognition?

Automatic human identification based on their physical characteristics (i.e, Biometrics), including ear, iris, face, fingerprint .etc., is an attractive field of research in the last few years because of the huge demand for more secured automated authentication systems. The uniqueness of biometric modalities for each individual could be considered as one of the most strong points of biometric authentication systems.

The human face has been extensively studied among all biometric traits [6–9]. Despite ensuring uniqueness and ease of data collection, face recognition systems suffer greatly due to their sensitivity to various image distortions such as mask-wearing, pose variations, illumination changes, and facial expressions, etc. In addition, the face is a somewhat privacy-violating modality. On the other hand, Iris and fingerprints as biometric traits guaranteed uniqueness and protect privacy, however, it is, to some extent, difficult to collect data samples. For instance, special sensors (i.e., additional costs) are required to generate iris images that are eligible for authentication ends.

Unlike the aforementioned biometric modalities, the human ear contains reliable information that makes it a robust source of information for human identification [10] as well as it preserves a stable structure that does not show drastic changes with age [11]. Additionally, despite the simplicity of its outer structure, the ear is unique for every individual and the change between two ears is distinguishable even in the case of identical twins [12, 13]. Thus, the ear can be considered as a distinctive and promising biometric modality due to its advantages compared to other biometrics: the ear images can be easily acquired from a distance without the cooperation of the concerned individual. Also, it doesn't require expensive tools for acquisition, and it is invariant to facial expressions. Another appealing feature regarding the ear is that it is a privacy-persevering trait, as it is difficult to detect other sensible traits (e.g., face) from the ear image.

1.3 Overview On The Related Work

In the literature, several attempts have been made to tackle Arabic handwriting recognition and unconstrained ear recognition issues. For the first issue (Arabic handwriting recognition) a considerable amount of works have been proposed to automate the checks recognition process. From a point of view, existing works dealing with that problem can be categorized -according to the technique used for extracting the features- into four categories. Some researchers [14–16] have proposed to deal with Arabic words as an ensemble of structural features based on word shape and edges. Other existing works [17–19] described Arabic words using pixels distribution measurements (i, e., statistical features). For instance, authors in [17] proposed an automatic holistic system for the recognition of handwritten Arabic literal amount based on Gabor filters with Bag of Features (BoF) that encode the local features. In [19], a set of statistical features including Invariant Moments (IV), Histogram of Oriented Gradients (HOG), and Gabor filters, are used to describe images of Arabic amounts. Indeed, little attention has been devoted to dealing with the Arabic literal amount recognition issue using deep learning techniques. A Deep Convolutional Neural Network CNN architecture with 17 layers has been used by [20]. However, in [21] authors investigated the efficiency of several deep architectures for recognition of Arabic handwriting literal amounts. The fourth subcategory considered works that combined several kinds of features to describe images [22–24].

On the other hand, ear detection and recognition have become the core of many recent works where CNN-based approaches have been considered and used by several relevant states of the art. From one point of view, existing works can be classified into two categories namely handcrafted-based and deep-based schemes. As for the first category of works(i.e., handcrafted-based methods), the research community on the ear has investigated a wide broad of texture (e.g., BSIF, LPQ, HoG…etc) and geometrical features to pick out features that are capable of faithfully reflecting the rich ear structure. For instance, Hassaballah et al [25] proposed a new descriptor for ear image representation namely Robust Local Oriented Patterns (RLOP). In [26], the authors proposed a new geometrical feature extraction method based on ear shape structure.

Otherwise, most recent works tend to consider deep-based schemes. For instance, Omara et al [27] proposed to learn Mahalanobis distance from features extracted using pre-trained models (i.e., VGG-s, VGGverydeep16, and ResNet). Authors in [28] proposed a simple CNN architecture for ear recognition with several parameters including learning rate, kernel size, epochs, and activation functions.

It should be worth noting that we only provide an overview of the related work in this section, while we devote an entire chapter (Chapter 3) to detailing the various aspects of methods concerned with the two issues.

1.4 Motivation

In spite of the significant number of works that are concerned with the automation of checks literal amount recognition, much more efforts have to be done to address limitations of the existing methods. To sum up, these limitations can be summarized as follows:

1. The existing works consider the whole image at the features extraction stage and gather all image features in a compact feature vector. Nevertheless, considering different parts of an image on different scales and levels could significantly improve classification outcomes. In the literature, there are several schemes to effectively extract image features including Multi-Level (ML) and Pyramid Multi-Level (PML). While ML and PML have shown to be effective than extracting features from the entire image, they suffer, in turn, from several cons. As an instance, shifted letters represent a serious issue leading to a minor performance of ML. More details about the limitations of ML and PML are presented in Section 4.2.1.
2. Certain methods are relying on structural features which are sensitive to word deformation. Indeed, the same word may be written differently (e.g., line deviation or missed PAWs) by the same writer multiple times. This leads to maximizing the intra-class variation and increases the confusion between words belonging to different classes.
3. Most existing studies have focused on using structural features to describe literal amounts, while too little attention has been paid to using texture features. Different homogeneous and repetitive parts of the word (PAWs) that are spreading over the image can be considered as texture. Thus, handwriting literal amounts can be characterized using the large arsenal of robust texture features due to the inherent nature of handwriting.

On the other hand, human identification based on the ear print has achieved a significant performance in the last year. However, it is still at the level of experiment and more effort remains to be done. The limitations of existing works proposed for ear recognition can be summarized as follows:

1. Nevertheless, the performance of handcrafted features is generally limited to constrained scenarios (i.e., where the image is taken under controlled conditions without significant variations of illumination, pose, or scale). Besides, the performance of

those features tends to considerably decrease as the size of the dataset increases (i.e., they are not scalable).

2. Most recent works tend to consider deep learning-based schemes. However, deep-based schemes are known to be data-hungry and they may require a great deal of time to perform features learning and extraction, especially for the networks with a high number of stacked layers and a huge number of parameters.

Given the aforementioned limitations, we are motivated to propose new solutions to both issues. The following section summarizes our contributions and demonstrates how we addressed the issues raised above.

1.5 Contributions

The main and major contributions of this thesis, which are published and validated by two different international journals [29, 30], can be summarized as follows:

1. To overcome the shortcoming of existing works dealing with Arabic handwriting recognition, we opt to propose a novel feature-independent scheme that is usable by any kind of feature, and which is capable to represent images faithfully. The proposed model termed as Generic Feature-Independent Pyramid Multi-Level model (GFIPML) combines the pros of both ML and PML.
2. Motivated by the fact that handwriting can naturally be described using texture features, as handwritten words are constituted of homogeneous and repetitive parts of the word (PAWs), we investigate, the performance of two robust and well-known texture features namely LPQ and BSIF.
3. To further enhance the recognition yields, we considered a multimodal system for Arabic handwriting recognition based on combining LPQ with multiple BSIF features, each with a different filter size.
4. We propose TR-ICANet a simple, yet efficient and speedy, network for automatic ear recognition.
5. To alleviate the disparity in ICANet histograms, we propose using a normalization technique namely Tied Rank Normalization (TR Normalization) for each histogram block.
6. To get rid of unconstrained conditions e.g., scale and pose variations, we suggest normalizing ear images using CNN.
7. To further enhance the recognition outcomes, we considered a soft-max average fusion of CNN-based schemes, which is then fused with CNN-like networks at the decision level using SVM classifier.

1.6 Thesis Structure

This **1*st* Chapter** describes the problems that we are dealing with, as well as the drawbacks of existing solutions. This chapter also describes the objective of the thesis. The major contributions of this thesis are also mentioned briefly. This chapter is followed by four other chapters in succession, the contents of which are briefly described below:

Chapter 2, introduces the background and the general context of the work. We review the important notions related to Visual Object Recognition (VOR) systems. These notions include supervised, unsupervised, semi-supervised learning, and the explication in detail of the main components of a VOR.

The **3*rd* Chapter** is devoted to surveying the literature methods concerned with both issues we are working on. Firstly, we review several state-of-the-art methods that are close to ours and deal with Arabic literal amounts recognition. The presented methods belong to three different approaches according to the technique used for extracting the features. In the second part of this chapter, we provide an overview of the state-of-the-art methods that have been proposed for ear recognition where the existing approaches are divided into five categories.

The **4*th* Chapter** presents our contributions as well as the two systems we developed for Arabic handwriting recognition and human ear identification. The different components constituting the systems are explained and detailed. We start by introducing the proposed GFIPML model for feature extraction. Then, we present our second contribution TR-ICANet deep scheme for unconstrained ear recognition.

Chapter 5 deals with the presentation of the obtained experimental results. This chapter is divided into three parts. The first part concerned the presentation of the databases we have used for evaluating the performance of our systems. In the second part, the results of the first contribution are presented in which we present the experimental results of ML, PML, and GFIPML models, as well as the implementation details of our proposed system for handwritten literal amount recognition. However, the results and all the experimental setups of the proposed TRICANet are presented and discussed in the third part. Moreover, to assess the performance of the proposed systems, at the end of each part, we provide a comparison against other related studies that are using the same databases.

At the end of the thesis, we draw the main conclusions of the work and we introduce some perspectives and future works.

Chapter 2

General Background: Machine Learning and Visual Object Recognition

2.1 Introduction

Visual object recognition refers to the task of automatically recognizing an object based on its content. This thesis is devoted mainly to dealing with two different visual objects which are Arabic Handwriting Recognition and Ear Recognition as they are still one of the most challenging tasks of visual object recognition because of the intra-class variability caused by differences in lighting, misalignment, non-rigid deformations, occlusion, and corruptions. In the literature, several machine learning techniques have been proposed to deal with such problems.

In this chapter, we start by presenting a general background of machine learning paradigms. Next, we explain in detail the structure and the stages of a recognizer system by introducing the idea behind each method used in the conception of the proposed systems. Finally, we summarize some system performance evaluation metrics and other existing comparison model techniques.

2.2 Learning Paradigms

Machine learning is a subfield of computer science and a branch of artificial intelligence that can be defined as the ensemble of data analysis methods that automates analytical model building using algorithms that iteratively learn from data. It allows computers to find hidden insights without being explicitly programmed [3]. Machine learning focuses on

the development of computer programs that can change when it is exposed to new data. Generally, machine learning systems are used to recognize objects in images, transcribe speech to text, match news items, posts, or products to users' interests, and select relevant search results [4]. Its roots may be traced back to the 1950s artificial intelligence movement, and it focuses on practical goals and applications, including prediction and optimization.

Machine learning paradigms can be broadly categorized into three types based on learning methods (i.e., model fitting) whether is "supervised", "semi-supervised" or "unsupervised". In the following, the principles of each category as well as the methods that pertain to each of them are explained.

2.2.1 Supervised Learning

A supervised learning algorithm analyzes the training data and produces an inferred function, which can be used for mapping new examples. In this approach, algorithms are trained using labeled examples. So the data are already assigned to a set of pre-defined classes (training set), the task is to determine which class (label) a new data belongs to. In another simple way, supervised learning is looking as learning with a teacher [31] or learning through induction [32].

The general scheme of supervised learning algorithms is illustrated in Figure 2.1.

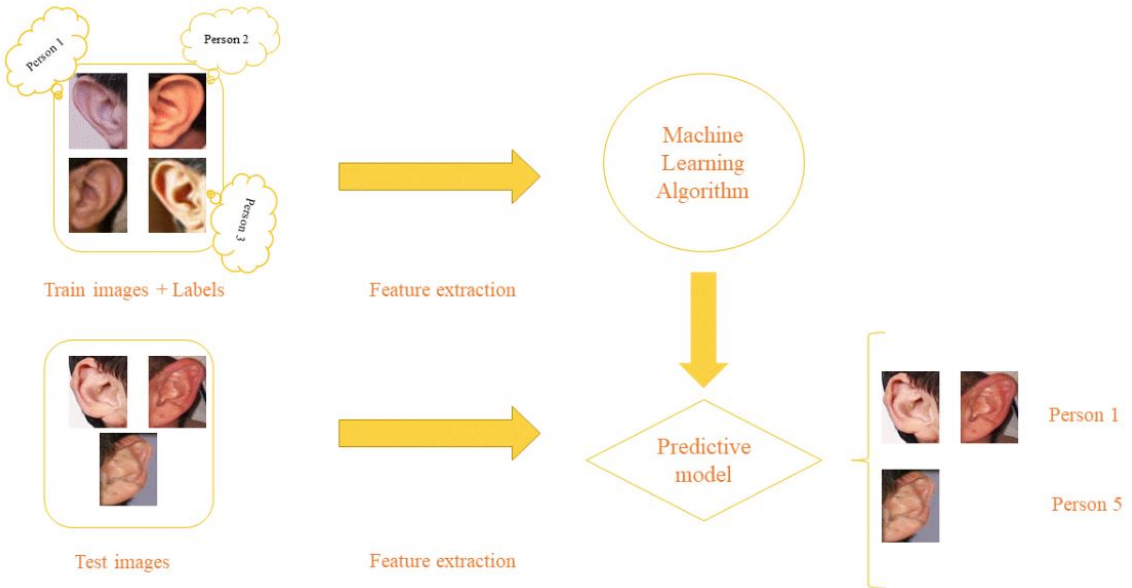


FIGURE 2.1: General scheme of supervised learning algorithms.

Supervised learning tasks can be further classified as:

— **Classification:** The ability to perform correct classifications is one of the most prominent and widely studied tasks in supervised learning [33]. Classification models are

used to solve problems in which the output variable can be classified (i,e., discrete variable).

— **Regression:** Regression models are used to solve problems where the output variable is a real value [34]. It is most commonly used to predict numerical values based on previous data observations. Some of the more well-known regression algorithms are logistic regression, linear regression, polynomial regression, and ridge regression.

In the literature, there are several practical applications of supervised learning algorithms including: Face Detection and Identification, Signature recognition, Text categorization, Customer discovery, Weather forecasting, Predicting housing prices, Stock price predictions..etc.

2.2.2 Unsupervised Learning

Unlike supervised algorithms principal, the input data is unlabeled in the unsupervised approaches. It means that no training data can be provided, and the goal is to find the similarity between the input instances (Figure 2.2). Thus, the machine must be able to classify the data without any prior knowledge of it [35].

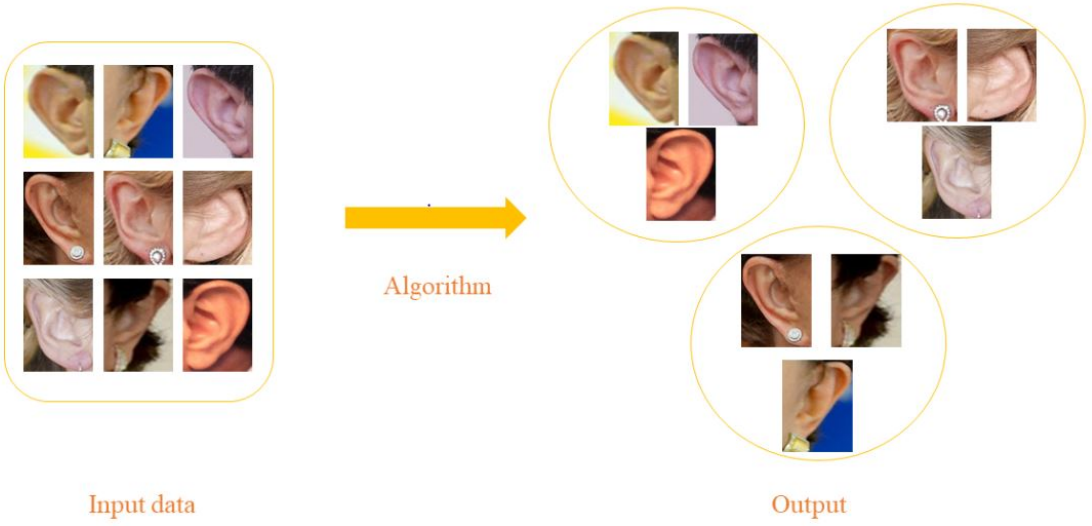


FIGURE 2.2: General scheme of unsupervised learning algorithms.

Unsupervised learning tasks can be further classified as:

— **Clustering** is one of the most widely used unsupervised learning techniques [36]. It aims to organize the unlabeled data into similar groups known as clusters. As a result, a cluster is a collection of data that are similar where the goal is to find similarities in the data and group them into clusters.

— **Association** is an unsupervised learning method for identifying relationships between variables in a large database. It identifies the group of items in the dataset that occur together [37].

In the literature, there are several practical applications of unsupervised learning algorithms including: Malware detection, Fraud detection, Human errors identification during data entry..etc.

2.2.3 Semi-supervised Learning

Semi-supervised learning is a type of learning that falls somewhere between supervised and unsupervised learning. In this type of learning, both labeled and unlabeled data are used, with the amount of labeled data being significantly less than the amount of unlabeled data. In comparison to the two previous categories (supervised and unsupervised), there are just a few semi-supervised approaches in the literature. A general scheme of semi-supervised learning algorithms is illustrated in Figure 2.3.

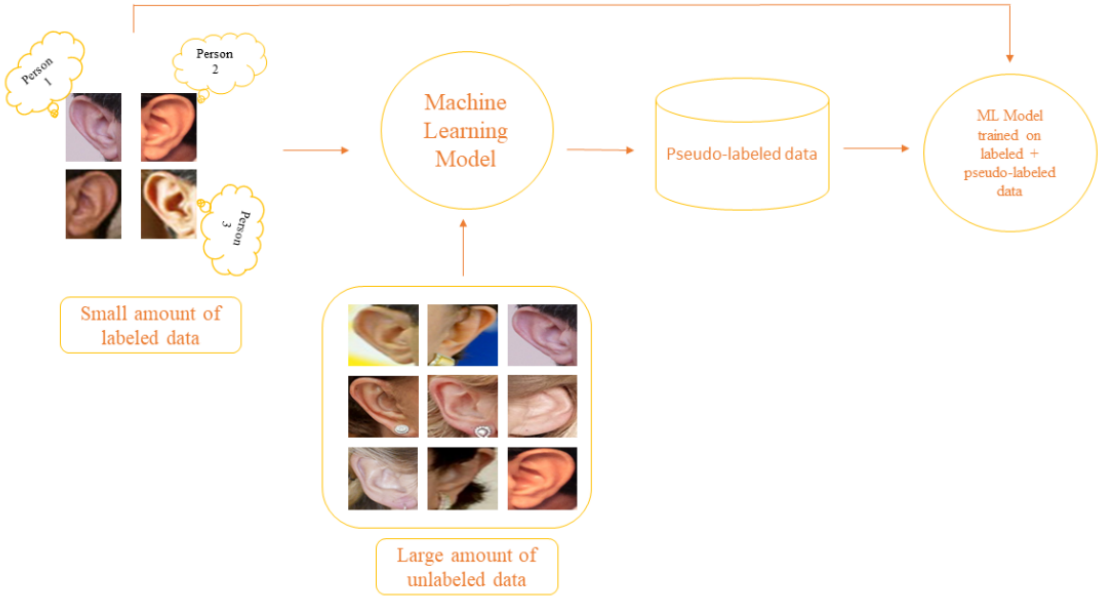


FIGURE 2.3: General scheme of semi-supervised learning algorithms.

2.3 Recognition Systems

Recognition is a typical scenario of analysis, which may consist in recognizing a scene, a person, a behavior, a manuscript, an emotion, a flower category, a signature, etc. However, with the growing number of images, this task is becoming quite complicated and much expensive in terms of time and effort as well. This heightened the need to develop automatic tools that are capable to perform recognition instead of human beings. The term "recognition" refers to the process of learning and understanding the input data and making wise decisions based on the supplied data. Visual object recognition refers to the task of automatically

recognizing an object based on its content. This goal is achieved by using machine learning techniques.

In the literature, existing approaches can roughly be classified into two categories namely holistic and segmentation approach [38]. As its name indicates, the segmentation approach consists in dividing the target images into specific parts depending on the object treated, meanwhile, the holistic approach considers the whole image i.e., without segmentation. Moreover, according to the data acquisition manner, two types of recognition systems are existing [39]: Offline and Online recognition systems. The Offline recognition system deals with previously acquired images (i.e., images obtained from a scanner). However, specific tools and capture devices are used in the Online recognition systems to get the data in real-time.

Generally, a recognition system consists of three main phases: preprocessing, feature extraction, and classification. The general scheme of a supervised recognition system is illustrated in Figure 2.4.

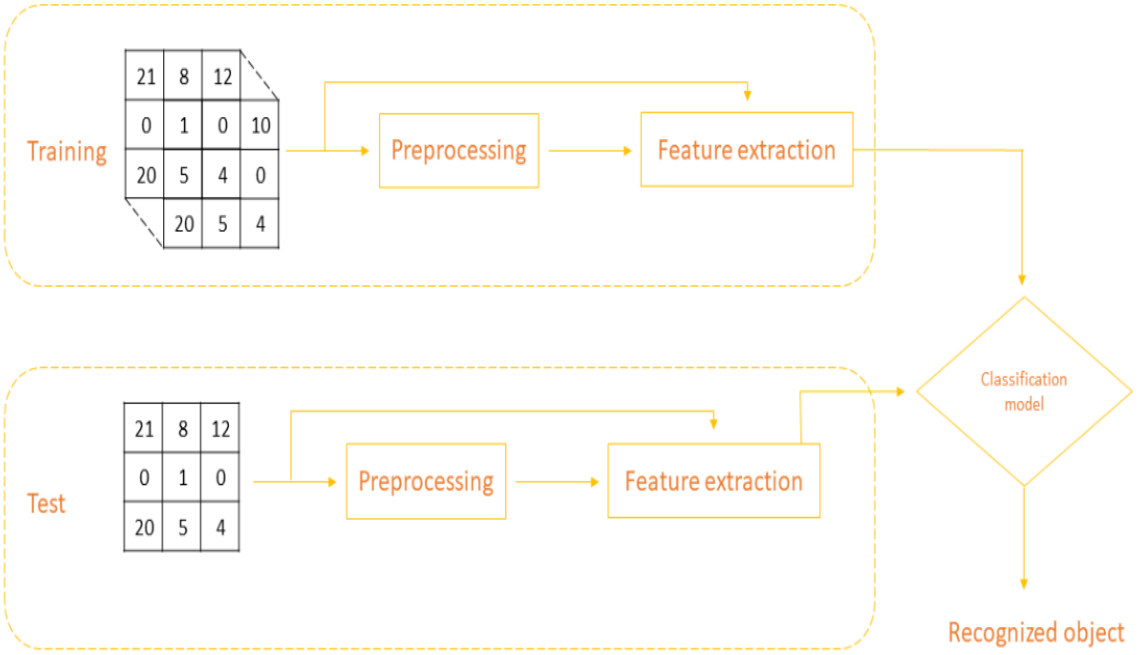


FIGURE 2.4: General scheme of a supervised recognition system.

As it's clearly shown in Figure 2.4, a recognition system comprises two stages: Training and Testing. The training stage contains the set of images used for learning where the correct class for each image is known. The images on the test, on the other hand, are the images to be recognized, and the performance of the learned model is evaluated based on how well the images are recognized. Next, we will explain the outlined phases of a recognition system:

2.3.1 Preprocessing

The reprocessing phase is the first step in the development of recognition systems. It aims to remove irrelevant information, which may harm the overall recognition [40], in order to enhance the image. This stage employs techniques such as noise reduction, thresholding, and normalization. However, even with the critical role of the other phases (i.e., feature extraction and classification), reshaping image-based on specific preprocessing methods is considered a vital step that may have a positive effect on the recognition outcome. Below, we mention some preprocessing techniques that we have used in the development of the systems presented in this thesis.

2.3.1.1 Histogram Normalization

Histogram normalization is one of the basic and simple preprocessing techniques. Sometimes the pixel values are heterogeneous because of noise, for example. Normalization aims to change the values of pixel intensity in order to ameliorate the image contrast. Several normalization techniques have been proposed in the literature, including Min-Max normalization, Quadratic-line-Quadratic (QLQ) normalization, Z-score normalization, and Hyperbolic tangent (HTan) normalization.

2.3.1.2 Gaussian Smoothing

Gaussian smoothing is one of the well-known methods for image preprocessing. It aims to blur or convolute the image with a gaussian function to remove noises and unuseful details. The Gaussian function formula in two dimensions is given as follows:

$$G(x, y, \sigma) = \frac{1}{2\pi\sigma^2} \exp(-(x^2 + y^2)/2\sigma^2) \tag{2.1}$$

Where x and y are the pixel coordinates. σ is the Gaussian distribution's standard deviation.

The resulting blurred image L is obtained by convolving the image I with the Gaussian function G as shown follows:

$$L(x, y, \sigma) = G(x, y, \sigma) * I(x, y) \tag{2.2}$$

2.3.2 Feature Extraction

For any recognition system, feature extraction represents the most important stage and it is considered as a key step, as a good feature extraction technique leads to a good classification and matching. The main aim of this stage is to extract the relevant characteristics to faithfully describe the image. Depending on the technique used, feature extraction methods can be classified generally into two categories: handcrafted and deep-based feature extraction techniques. For the handcrafted subcategory, the algorithms used in this approach are manually designed and extracted. The term "Handcrafted" refers to properties derived using various algorithms from the information contained in the image itself. From one point of view, the algorithms of this approach could be divided into three main types according to the characteristics of the image to be extracted which are texture, shape, and color features. In our work, we focused our interests on the two first types as the studied objected can't be differences from the color. However, deep learning methods do not use human engineers to design features; instead, they are learned from data using a general-purpose learning procedure [4]. In another word, unlike handcraft methods which extract features manually, features are learned automatically.

2.3.2.1 Texture-based Techniques For Feature Extraction

Texture can be defined as the structure of a region, with repetitive patterns containing elements that are arranged according to a set of rules [41]. In the literature, several texture features descriptions have been proposed, they can be roughly divided into three approaches [42]: frequency [43, 44], statistical [41, 45], and geometrical [46, 47]. Next, we provide detailed explanations for texture features we used in this thesis.

❶ Local Binary Patterns (LBP)

The original Local Binary Patterns LBP operator was proposed by [48], it is designed specifically for face recognition [49]. The main idea behind the LBP is to label the pixels of an image with decimal numbers, called Local Binary Patterns or LBP codes, which encode the local structure around each pixel [50]. The general principle is to compare the intensity level of each pixel with the levels of its neighbors by subtracting the central pixel value. The resulting values that are strictly negative are encoded with 0 and the remaining with 1.

A binary number is obtained by concatenating all these binary codes in a blockwise direction starting from the top-left. The corresponding decimal value is used for labeling. Figure 2.5 shows the principle of the LBP descriptor:

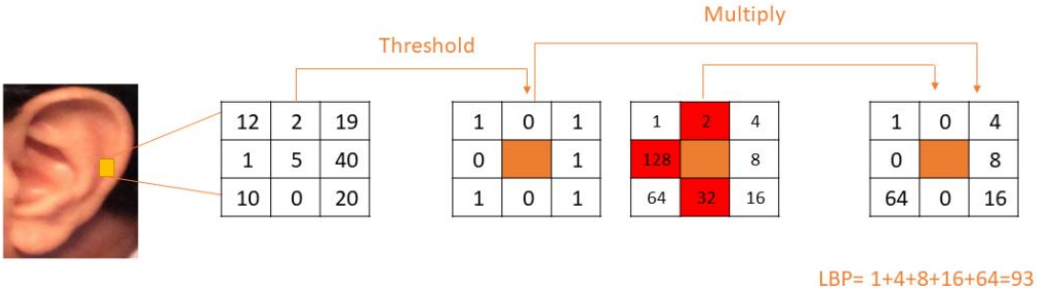


FIGURE 2.5: The principle of the LBP encoder.

❷ Local Phase Quantization (LPQ)

The Local Phase Quantization LPQ operator [51] is a textural descriptor that is widely used for face recognition and writer identification. It is based on the local phase information extracted using 2-D DFT or, more obviously, a short-term Fourier transform (STFT) that is computed over a rectangular M×M neighborhood at each coordinate x within the image given by equation 2.3:

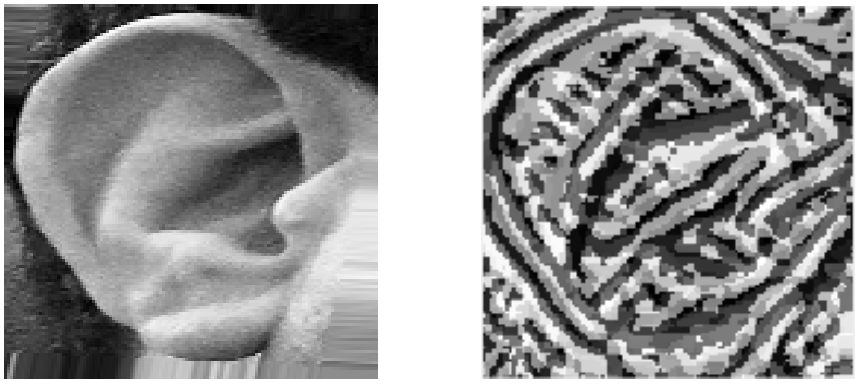
$$F(u,x) = \sum_{y \in N_x} f(x-y) \exp(-j2\pi u^T y) = w_u^T f_x \tag{2.3}$$

Where

— w_u is the basis vector of the 2-D DFT at frequency u.

— f_x is a vector containing all M^2 neighborhoods.

An example of an image before and after applying LPQ is illustrated in Figure 2.6:



(A) Before (B) After

FIGURE 2.6: An image before and after applying LPQ

❸ Binarized Statistical Image Features (BSIF)

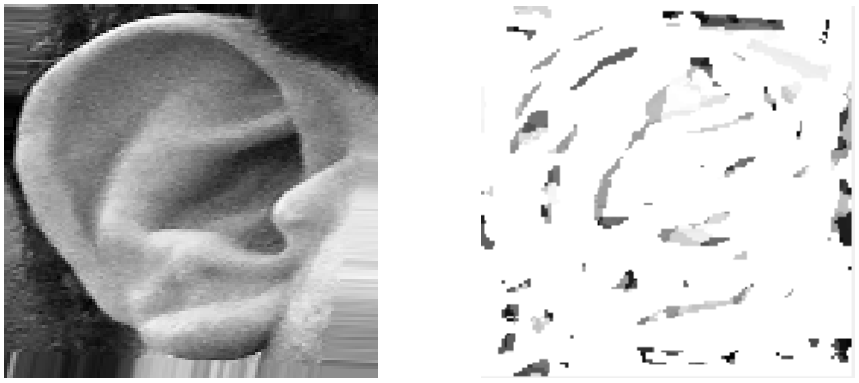
The Binarized Statistical Image Feature (BSIF) is a local texture encoder proposed by [52] inspired from LBP and LPQ methodologies. It is an efficient descriptor, which is applied firstly for face recognition. The main principal of BSIF is to encode each image pixel with

a binary code that is obtained by first convolving of each pixel's neighbors with a set of linear filters. In contrast to other methodologies, the filters used by BSIF are learned from statistics of a small set of natural image patches by maximizing the statistical independence of the filter responses (ICA) [53]. The filters responses are binarized according to a threshold fixed at 0. If the filter response $Si > 0$ the pixel is encoded with 1 and with 0 otherwise. The code length produced corresponds to the number of filters used.

More formally, for a given sub image I of size l × l pixels and a linear filter Fi of the same size with I, the filter response Si is obtained by:

$$\mathrm{Si} = \sum_{u,v} Fi(u,v)I(u,v) \tag{2.4}$$

An example of an image before and after applying BSIF is illustrated in Figure 2.7:



(A) Before (B) After

FIGURE 2.7: An image before and after applying BSIF

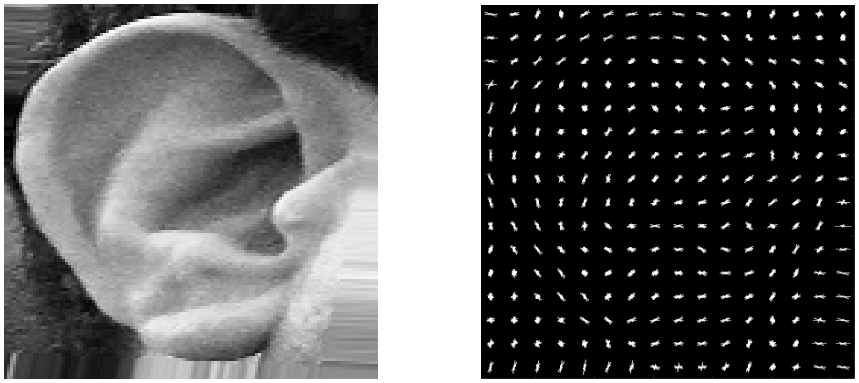
2.3.2.2 Geometrical-based Techniques For Feature Extraction

As its name indicate, geometrical features studied the geometrical characteristics of the input image including edge, corners, blobs, and ridges. Below we mention some known geometrical feature extraction methods.

❶ Histogram of Gradient (HOG)

The histogram of gradient (HOG) was proposed firstly by Dalal and Triggs [54] for human detection purposes. Furthermore, it is widely used in image processing and computer vision tasks to detect objects. The main idea behind the HOG descriptor is to divide the input image into small squared cells. Then, from each cell count the histogram occurrences of gradient orientation by convolving them with a 1 D gradient filter mask [-1 0 1]. The HOG descriptor is produced by combining the histograms.

An example of an image before and after applying HOG is illustrated in Figure 2.8:



(A) Before (B) After

FIGURE 2.8: An image before and after applying HOG

❷ Gabor filters responses

In pattern recognition and image processing, Gabor filter features are used both for texture analysis and object edge detection. It is an efficient feature that proved its ability to well represent the Arabic handwritten [55]. The response features of the Gabor filter are obtained by the convolution of the word image with a set of Gabor filters with different orientations and scales. A 2-D Gabor filter g(x,y, λ,θ) can be expressed mathematically by equation 2.5:

$$g(x, y, \lambda, \theta) = \frac{1}{2\pi(k\lambda)^2}(\exp(-\frac{x'^2 + y'^2}{2(k\lambda)^2})(\exp(2\pi j \frac{x\cos\theta + y\cos\theta}{\lambda}) \tag{2.5}$$

Where:

- λ and θ are respectively the orientation in degrees of the carrier frequency and the wavelength (the scale) in pixels.
- k is a scalar factor.
- x'= x cos θ + y cos θ.
- y'= -x sin θ + y cos θ.

An example of applying Gabor filters on an image is illustrated in Figure 2.9:

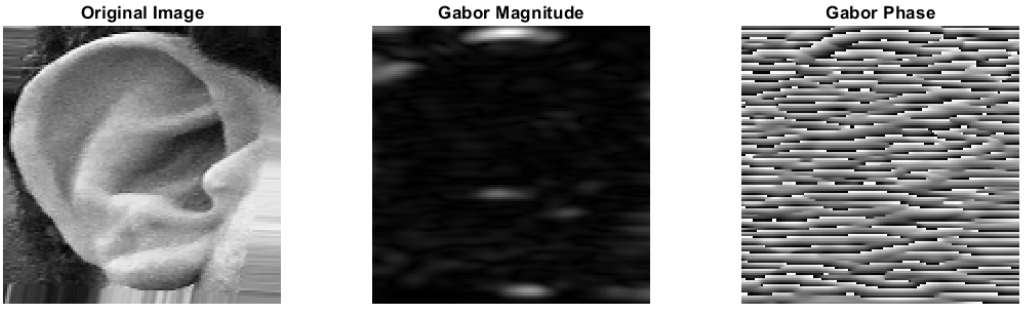


FIGURE 2.9: An image before and after applying gabor filters.

2.3.2.3 Deep learning-based Techniques For Feature Extraction

In recent years, an incredible growth for many computer vision and pattern recognition tasks have appeared and proposed where deep learning techniques are at the core of solutions due to their efficiency in well capturing complex characteristics from the low level to the high level.

Convolutional Neural Networks CNNs [56] are one of the most popular and commonly used deep learning techniques. Moreover, they proved their performance and efficiency especially for handwriting recognition problems, and they are considered as top solutions in such issues [17, 57–60]. Commonly, CNNs are divided into three main layers which are convolutional layer, pooling layer, and fully connected layer (FCL). The convolutional layer aims to extract features based on filters that stride over the input image that produces the feature map. Pooling layer aims to down sample feature maps by summarizing the number of features in each patch of the feature map. The two most often used pooling techniques are average and max pooling. Finally, the recognition process is performed by the fully connected layer.

It is worth noting that there are two types of deep learning techniques: supervised and unsupervised. The main difference between them is whether a backpropagation phase exists or not. In general, because backpropagation is not used, the unsupervised learning process speeds up the network's computation. Bellow, we mention some deep learning techniques used for feature extraction including both unsupervised (PCANet) and supervised ones (VGG, Googlenet, Resnet):

❶ Principal Component Analysis Network (PCANet)

The Principal Component Analysis deep network (PCANet) [61] was proposed firstly for face recognition. It is intended to be a lightweight convolutional neural network (CNN) in which PCA is used to learn the filters of the convolution layers. PCANet shared the same structure as the CNN; however, it involves an unsupervised learning process in which the PCA is used to learn the filters of the convolutional layers from image patches rather than the iterative process for adjusting weights. It contains mainly three stages: Filter learning via PCA, Binary hashing, and a Block-wise histogram stage. Unlike CNN, PCANet does not perform nonlinear operations between layers; instead, the operation is performed only at the output layer. In PCANet, the nonlinear operation is the binary thresholding operation that converts the filter responses into a binary map. The spatial relationship between blocks is then encoded using block-wise histograming [62]. Finally, all block-wise histograms are concatenated to form the output feature vector.

A detailed block diagram of two-stage PCANet is illustrated in Figure 2.10:

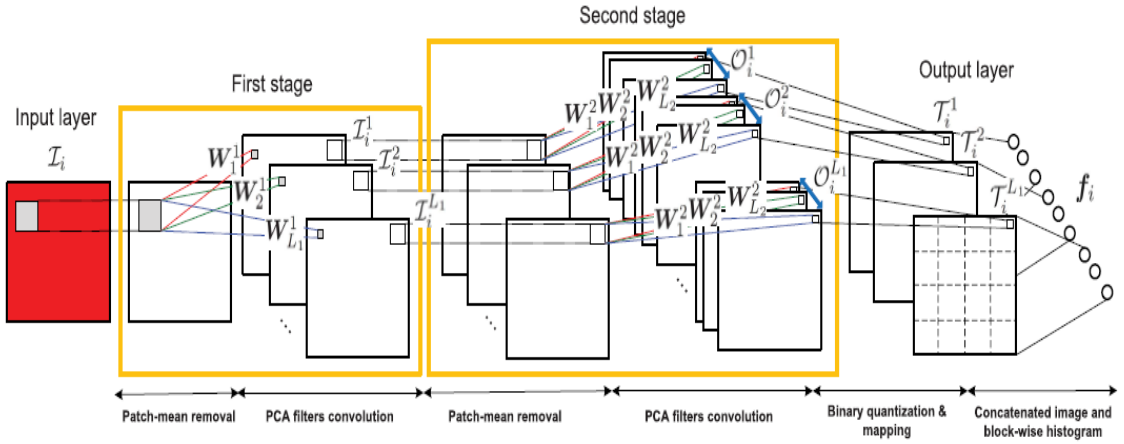


FIGURE 2.10: A detailed block diagram of two-stage PCANet [61].

❷ Visual Geometry Group (VGG)

Visual Geometry Group (VGG) network was firstly proposed by Simonyan et al.[63]. The work contribution is presented to prove the common idea that says going deeper through a network will give better accuracies [64]. The architecture consists of two convolutional layers using ReLU as an activation function followed by a single max-pooling layer and several fully connected layers that are using the ReLU activation function also. A Softmax layer for classification purposes finalizes the VGG network. Several VGG versions according to the number of layers exist including VGG-m [65], VGG16 and VGG-19 [63] where each model contains respectively 8, 16, and 19 layers.

A visualization of the VGG architecture is given in Figure 2.11:

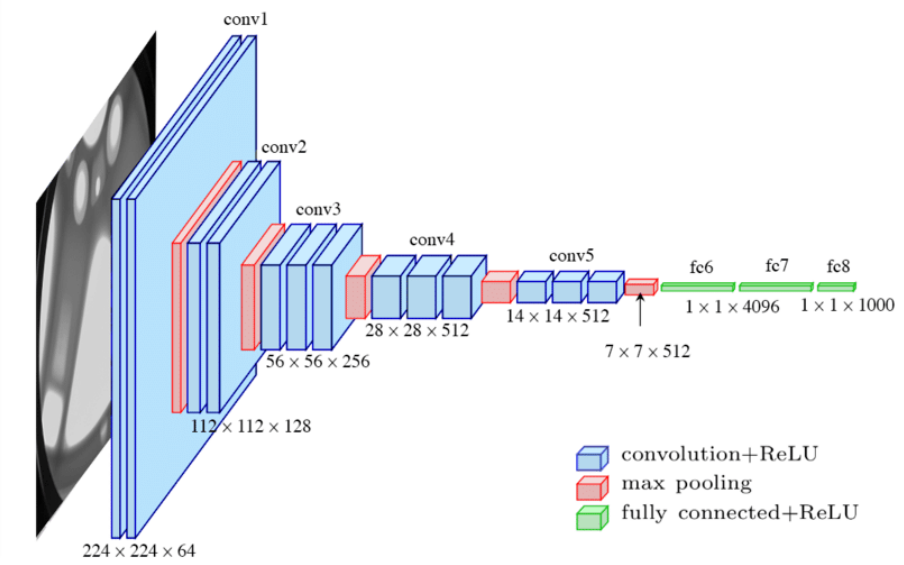


FIGURE 2.11: The standard VGG16 Architecture [63].

❸ Googlenet

Googlenet (The winner of the ImageNet Large Scale Visual Recognition Challenge ILSVRC 2014) was proposed by Christian Szegedy of Google et al. [66] where the main objective is to reduce the computational complexity compared to the traditional CNN. Googlenet model based on incorporating "Inception Layers" possesses seven million parameters and contains nine inception modules. It contains four convolutional layers, four max-pooling layers, three average pooling layers, five fully connected layers, and three softmax layers for the main auxiliary classifiers in the network [67]. In total, Googlenet contains 22 layers with a dropout regularization parameter in the fully connected layer and it uses ReLU as an activation function in all of the convolutional layers.

A visualization of the Googlenet architecture is given in Figure 2.12:

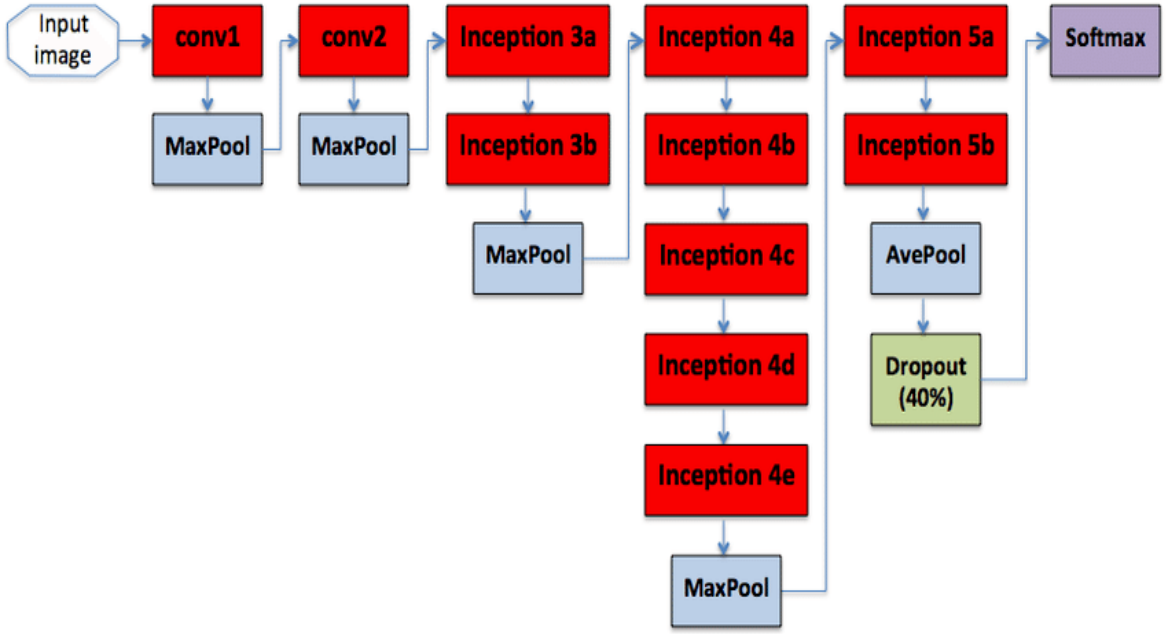


FIGURE 2.12: Googlenet architecture [66].

❹ Deep residual networks (Resnet)

Deep residual networks Resnet (The winner of the ImageNet Large Scale Visual Recognition Challenge ILSVRC 2015 with a 3.57% error rate) was proposed by He et al. [68]. Resnet has been proposed to solving vanishing gradient problem -which could be appeared with the high number of layer- by skipping one or more layers by a residual mapping. In other term, adding a parameter to the output from the previous layer to the layer ahead. In the literature, there are several Resnet versions available, each with a different number of layers.

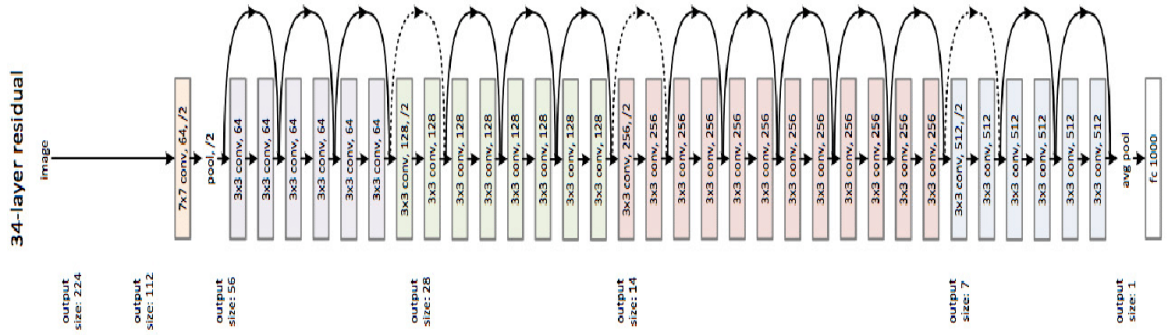


FIGURE 2.13: A 34-parameter-layer residual network [68].

2.3.3 Feature Extraction Schemes

The way of extracting image features plays a decisive and critical role in typical image recognition tasks. Over the last decade, several images feature extraction schemes such as Multi-Level (ML) and Pyramid Multi-Level (PML) have been proposed. In the rest of this section, we explain in detail the idea behind the aforementioned schemes.

2.3.3.1 Multi-Level (ML)

The main idea behind multi-level representation [69] is to extract features from different blocks, those features are then combined, which yield a final histogram constituted of features extracted from different blocks (Figure 2.14). In the level M, the image is divided into M×M blocks (M= 1, 2, 3,... M). In the next levels, the image is iteratively divided into M-1×M-1, M-2×M-2,..., 1×1, respectively. Features extracted from each level (i.e., histograms) are combined together, which produce a feature vector of $1^2 + 2^2 + 3^2 + ... + M^2$ dimensions.

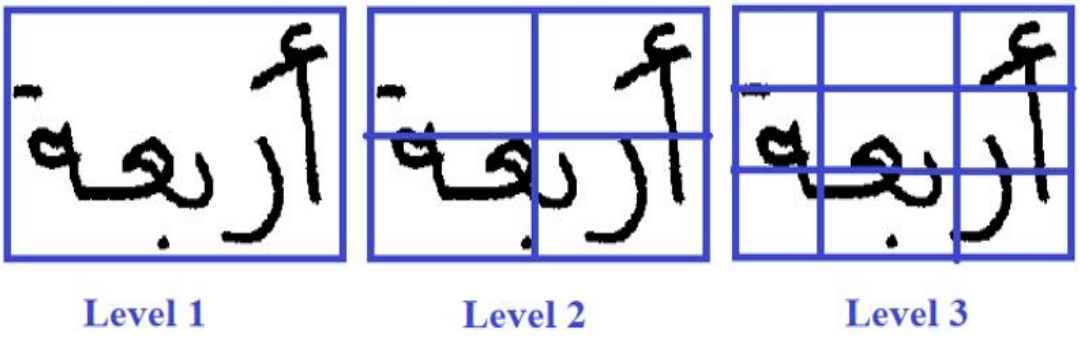


FIGURE 2.14: An example of Multi-Level representation with 3 levels.

2.3.3.2 Pyramid Multi-Level (PML)

Pyramid Multi-Level (PML) architecture was proposed firstly by Bekhouche et al [70] to extract features for facial demographic estimation. The main idea behind PML representation is to extract features from several blocks of the same size before iteratively progressing to the next level, where the image size is half of the previous image (Figure 2.15). One of the strongest points of this representation is that interested in the size of the image and studies it on several scales (Several objects can be differenced in other scales). Features extracted from each scale (i.e., histograms) are combined together, which produce a feature vector of $1^2 + 2^2 + 3^2 + ... + M^2$ dimensions (the size of the resulting feature vector is the same with ML vector size).

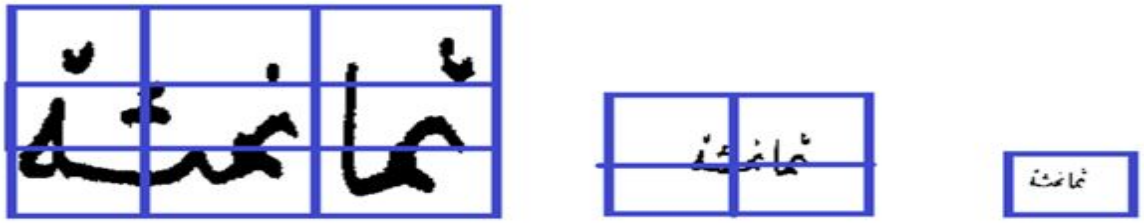


FIGURE 2.15: Pyramid Multi Levels PML representation.

2.3.4 Classification

The main aim of the classification stage is to assign a class label to each input test image. In the literature, several classifiers have been used by recognition systems. We mention below some well-known classifiers:

2.3.4.1 K-Nearest Neighbors Classifier

The K-nearest neighbors (KNN) classifier is a basic classfier that retains all available data (training set) and classifies new data (test set) by comparing it to the most similar samples in the training set using a similarity metric (Figure). The steps of KNN algorithm could be summarized as follow:

- Determine the parameter K= number of nearest neighbors (Thumb Rule can be used to determine K).
- Calculate the distance between the query instance (test example) and all the training samples.
- Sort the distance and determine the nearest neighbors based on k minimum distance.
- The predicted class of the test instance is affected based on the majority vote (The most frequent class) of the k closest neighbors selected in the previous step.

A distance function is required to compare the similarity of feature vectors. We can list numerous metrics of similarity among them:

$$\text{Euclidean distance} = \sqrt{\sum_{i=1}^{n} (\mathrm{x_i}\text{-}\mathrm{y_i})^2} \tag{2.6}$$

$$\text{Manhattan distance} = \sum_{i=1}^{n} |x_i - y_i| \tag{2.7}$$

$$\text{Minkowski distance} = \sqrt[p]{\sum_{i=1}^{n} |x_i - y_i|^p} \tag{2.8}$$

Where (x_i, y_i) represent the coordinates of the data point.

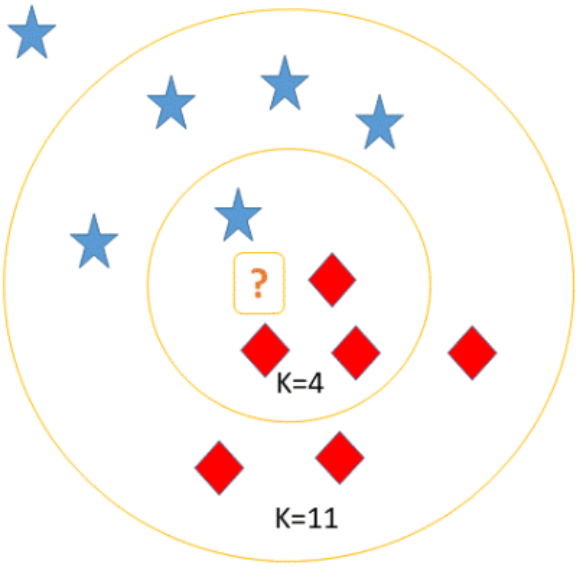


FIGURE 2.16: An example of KNN classification.

2.3.4.2 Naive Bayes Classifier

Naive Bayes classifier is an algorithm that employs Bayes' theorem to classify objects under the premise of predictor independence. It assumes that the presence of one feature in a class has no bearing on the presence of any other feature. Given an instance (i.e., sample) to be classified and represented by a vector X=(x1„,xn) encoding some n features (independent variables), Naive Bayes assigns to this instance probabilities: P(Ck|x1„,xn) for each of classes Ck. The new simple will be assigned to the class that has a high probability of belonging.

2.3.4.3 Support Vector Machine (SVM)

Support Vector Machine (SVM) algorithm was originally introduced by Vapnik [71] in his work on structural risk minimization to solve the pattern classification and regression problems. It was initially designed for binary separation problems.

SVM views a data point as an n-dimensional vector, in n-dimensional space R^n and the objective is to know the ability to separate such points with an $(n-1)$ dimensional hyperplane (Canonical plane). This is called a linear classifier [72]. Many hyperplanes could classify the data; one reasonable choice as the best hyperplane is the one that represents the greatest separation, or margin, between the two classes, because the larger the margin, the lower the classifier's generalization error. The support vectors and margins are used to find the hyperplane.

Figure 2.17 illustrates the principle of SVM algorithm:

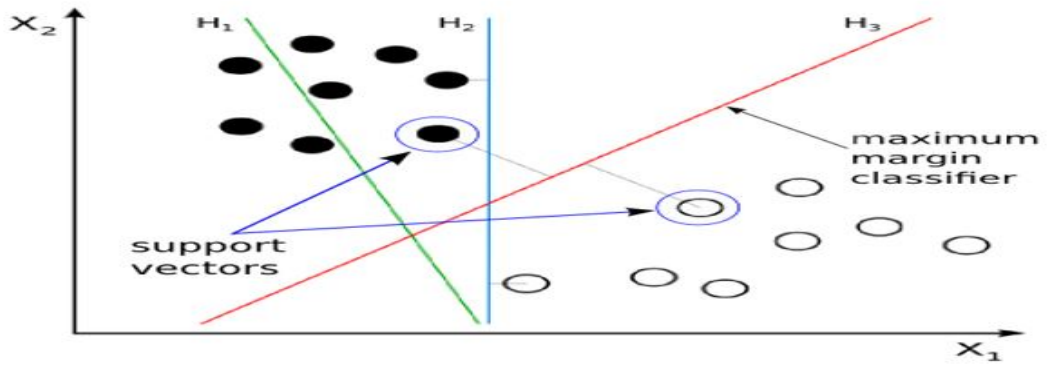


FIGURE 2.17: The principle of SVM algorithm.

If the data isn't linearly separable, SVM map the new data to another high dimensional feature space where the data is linearly separable (Figure 2.18).

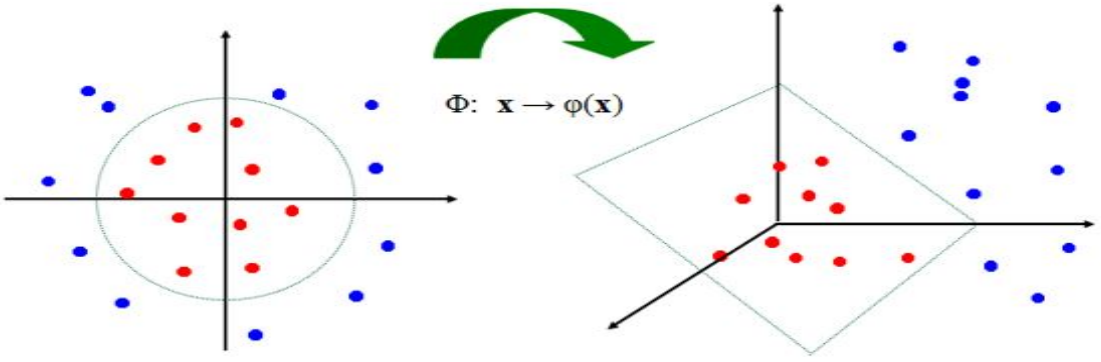


FIGURE 2.18: Nonlinear classification using SVM.

However, the majority of real-world problems are multiclass. Currently, there are two types of approaches for multi-class SVM [73]:

1. By constructing and combining several binary classifiers: In this approach, several methods have been proposed such as one-against-all, one-against-one, and dagsvm methods.
2. By directly considering all data in one optimization formulation.

2.3.4.4 Linear Discriminant Analyses (LDA)

Linear Discriminant Analyses (LDA) classifier has been successfully applied in the field of pattern recognition [74, 75] for both dimensionality reduction and classification.

To perform prediction, LDA calculates for each test sample, the cost of belonging to each class. Then, the sample is assigned to the class having obtained the smallest misclassification cost. The cost function is given as follow:

$$y = Arg \min \sum_{k=1}^{k} P(k\,|x\,)C(y\,|k\,) \tag{2.9}$$

Where

— y is the predicted class.
— k is the total number of classes.
— $P(k|x)$ is the posterior probability of the class k for the observation x.
— $C(y|k)$ is the cost of labeling an observation with y when its true class is k.

2.4 Multi-modal Systems Based On Classifier Combination Schemes

Because of the nature of the data, some classifiers produce good results while others do not because of the strong dependence between the classifier performance and the data [76]. As a result, researchers have recently attempted to overcome this problem by combining multiple classifiers (sequential or in parallel) which reduces this dependency and makes the system more generally. Thus, the final decision is taken from several classifiers rather than relying on one classifier which helps the system to be more accurate (Figure 2.19).

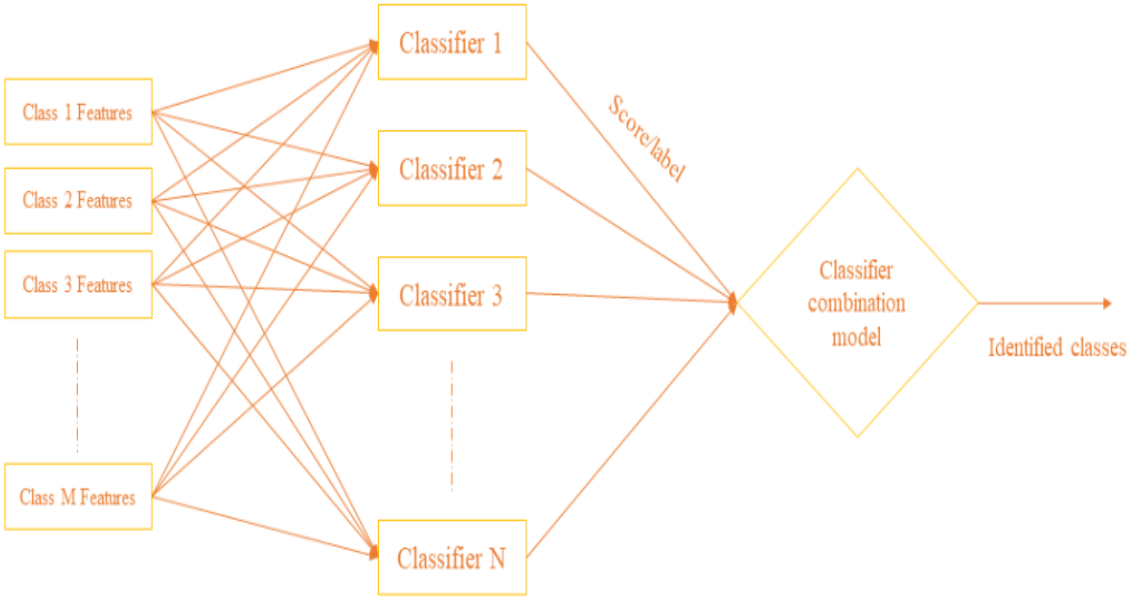


FIGURE 2.19: Classifiers combination principal.

Several simple and efficient combination methods are existing including Majority vote (The final label decision is assigned to the class label that has been voted for most), Product, Sum, Min, Max, Mean and Median.

Supposing we are given L classifiers (j=1...L), for each input image, the probability of belonging to each of the N classes (i=1...j), denoted as Pij, is computed. The equations (5-10) present, respectively, the formula of each of the aforementioned methods:

$$\text{Product rule: } Ci = \prod_{j=1}^{L} P_{i,j} \tag{2.10}$$

$$\text{Sum rule: } Ci = \sum_{j=1}^{L} P_{i,j} \tag{2.11}$$

$$\text{Min rule: } Ci = Min_{j=1}^{L} P_{i,j} \tag{2.12}$$

$$\text{Maximum rule: } Ci = Max_{j=1}^{L} P_{i,j} \tag{2.13}$$

$$\text{Mean rule: } 1/L \sum_{j=1}^{L} P_{i,j} \tag{2.14}$$

$$\text{Median rule: } Ci = Med_{j=1}^{L} P_{i,j} \tag{2.15}$$

2.5 System Evaluation Metrics

Assessing the performance of the system is a fundamental issue in the conception of the object recognition/identification system. In this section, we review some methods used for evaluating the efficiency of a system. Before mentioning the methods, we should clarify some abbreviations that will be used then.

Supposing a system with two classes:

Table 2.1: Evaluation metric concepts explanation.

		Predicted class	
		Positive	Negative
Actual class	Positive	True Positive	False Negative
	Negative	False Positive	True Negative

— True Positive (TP): is the number of instances that belong to the positive class and the system correctly classified them.
— True Negative (TN): is the number of instances that belong to the negative class and the system correctly classified them.
— False Positive (FP): is the number of instances that belong to the negative class and the system misclassified them to the positive class.
— False Negative (FN): is the number of instances that belong to the positive class and the system misclassified them to the negative class.

The concepts could be generalized easily for multi-classification systems.

2.5.1 Recognition/identification rate (Accuracy)

Accuracy is one of the most used and basic metrics for evaluating the performance of a recognition/identification. The accuracy is defined as the quotient of the correctly predicted classes over the total testing classes.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{2.16}$$

2.5.2 False Positive Rate (FPR)

False Positive Rate (FPR) is the proportion of negative cases in the data that were mistakenly identified as positive cases.

$$FPR = \frac{FP}{FP + TN} \tag{2.17}$$

2.5.3 False Negative Rate (FNR)

False Negative Rate (FNR) is the proportion of positive cases in the data that were mistakenly identified as negative cases.

$$FNR = \frac{FN}{FN + TP} \tag{2.18}$$

2.5.4 Sensitivity and Specificity

Sensitivity, Recall, or True Positive Rate (TPR) is the metric that evaluates the ability of a model to predict the true positives of each available class. However, Specificity or True Negative Rate (TNR) is the metric that evaluates the ability of a model to predict the true negatives of each available class. The mathematical formulas of calculating Sensitivity and Specificity are given as follow :

$$Sensitivity = \frac{TP}{FN + TP} \tag{2.19}$$

$$Specificity = \frac{TN}{TN + FP} \tag{2.20}$$

2.5.5 Precision

Precision expresses the accuracy of a model's positive prediction. It is calculated by dividing the number of true positives by the total number of positive predictions.

$$Precision = \frac{TP}{FP + TP} \tag{2.21}$$

2.5.6 Statistical Significance Tests For Models Comparison

Having a high recognition/identification isn't enough to decide whether a model is better than another. In some cases, even the rate is high but the performance of the system is low in many other terms. In another word, it is not known if the results are real or they are a result of a statistical fluke. Statistical analysis is one of the methods that can be used to compare the performance of two models. The results show and prove wherever the

difference between them is statistically significant or not. In this context, the statistical method ANalysis Of the VAriance (ANOVAs) is considered one of the simple and efficient methods. ANOVA compares the means of different samples to determine the impact of one or more factors. A statistically significant result is reported by ANOVA if any group differs significantly from the overall group mean. The F statistic, which is the ratio of the mean sum of squares to the mean square error, is used to calculate significant differences between group means .

2.5.7 Curve-based Methods For System Performance Evaluation

The performance of recognition systems can be expressed graphicly by using some specific curves. Moreover, When comparing recognition systems with similar performance, the logarithmic scale is sometimes used to make them clearer and more readable. As a result, we find:

2.5.7.1 Receiver Operating Characteristic (ROC)

The Receiver Operating Characteristic curve is one of the standard methods for comparing the performance of classification models. It is a graphical representation that plots TPR against FPR at different classification thresholds. Moreover, calculating the value of Area Under Curve is important. AUC represents the degree or measure of separability. It expresses how well the model can distinguish between classes. Whenever AUC is near to 1, the classifier can well distinguish between classes. An example of this curve is shown in Figure 2.20

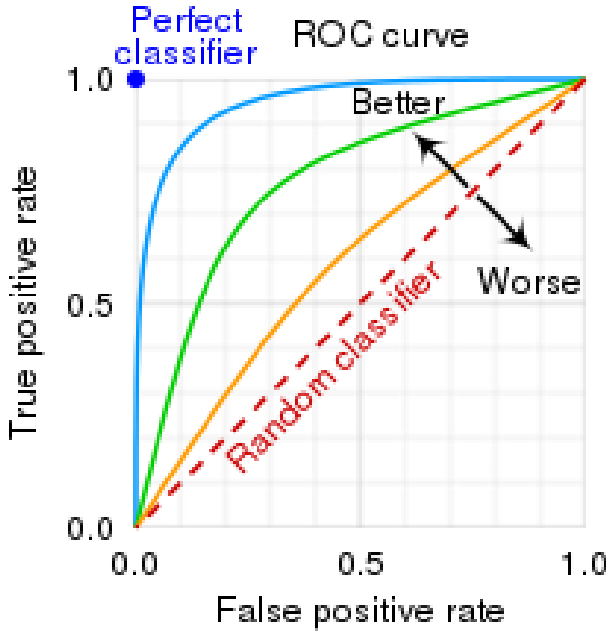


FIGURE 2.20: Receiver Operating Characteristic curve (ROC).

2.5.7.2 Cumulative match characteristic curve (CMC)

Cumulative match characteristic curve (CMC) is one the most popular graphical evaluation metrics for recognition systems. CMC curve plots the recognition/identification rate against the rank. An example of this curve is shown in Figure 2.21.

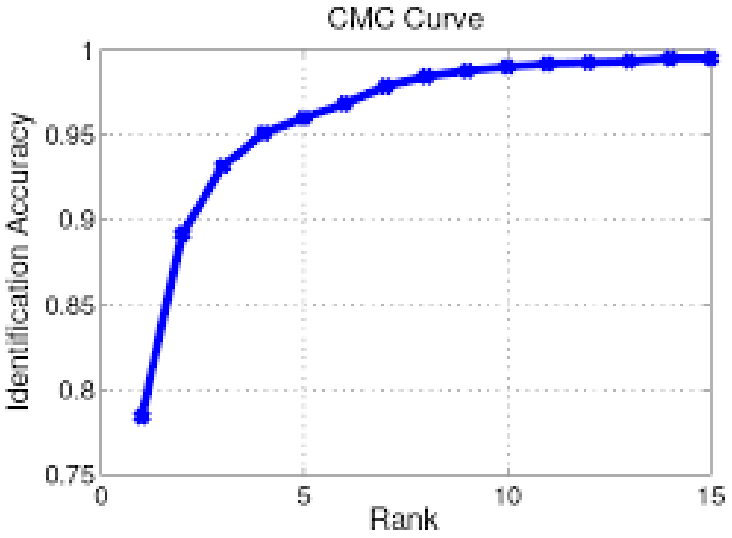


FIGURE 2.21: Cumulative match characteristic curve (CMC).

2.6 Conclusion

This second chapter is devoted mainly to offering a general background of machine learning with the intention to introduce briefly its paradigms and explain the stages and the structure of a recognizer system. To that end, the fundamental concepts and their principles have been described. As it is considered a critical stage in any recognition system, this chapter included a discussion and explanation of some features descriptors followed by presenting some existing features extraction schemes. At the end of the chapter, we mentioned the existing evaluation metrics and the comparison models techniques. In the next chapter, we will present an overview of the state-of-the-art related works on the specifically focused objects (Arabic Handwriting Recognition and Ear Recognition).

Chapter 3

State Of The Art Methods

3.1 Introduction

This thesis is dedicated mainly to dealing with two important applications of visual object recognition which are Arabic handwriting recognition and ear recognition.

In this chapter, we present an overview of the relevant state-of-the-art for each sub-field (i.e., Arabic handwriting recognition and ear recognition). We categorized the previous works of each sub-field according to the used technique for extracting features. We summarized all the relevant state-of-the-art in a table at the end of each part.

3.2 Related Work For Arabic Handwriting Recognition

Because of its nature, inherent characteristics, and writing style, Arabic handwriting recognition is still a challenging task. The existing researches in this context can be divided into two categories. Online Arabic handwriting recognition and Offline Arabic handwriting recognition. For each subcategory, some researchers are dealing with the recognition of isolated characters, and some others are dealing with the recognition of whole words. For this work, we focus our interests on the recognition of Offline Arabic handwriting words and more specifically on the recognition of Arabic handwritten Literal amounts words as the researches efforts in this context have not yet yielded satisfactory results because of the complex nature of words belonging to this kind of databases.

The existing state-of-the-art (Table 3.1) can be divided according to the nature of extracted feature methods into three categories namely structural, statistical and hybrid. The first subcategory (i.e., structural features), includes both topological and geometrical characteristics of an image [77]. It can include the total number of dots and their position relative to the baseline, as well as the end of points and loops [78, 79]. However, statistical features

are numerical measures, which are obtained from pixels distribution. The hybrid features combined several kinds of features.

It should be worth noting that in the literature little attention has been devoted to dealing with such issues using deep learning techniques. The authors of [20] used a deep Convolutional Neural Network CNN architecture with 17 layers on 50 classes extracted from the AHDB database in which a recognition rate of 97.8% has been achieved with data augmentation and 96.8% without data augmentation. However, in [21], authors investigated and tested several deep learning architectures for recognizing Arabic handwritten literal amounts, including ALexNet and RNNs. The authors presented two recognition models in their work: one based on a holistic word recognition model and the other on a character recognition model.

3.2.1 Arabic Handwriting Literal Amount Recognition Based On Structural Features

Some researchers have proposed to characterize Arabic handwritten literal check amounts by considering a set of structural features. In this context, Farah et al [15] proposed to use some structural features (ascenders, descenders, loops, etc.) which were extracted from 4800-word images representing handwritten Arabic literal amounts. To further enhance the performance of their system performance, the authors combine the results of three classifiers, which are multilayer neural network (MLP), k-nearest neighbor (KNN), and fuzzy k- nearest neighbor (FKNN), where a recognition rate of 94% has been achieved. An enhanced structural perceptual feature extraction model (PFM) to recognize handwritten Arabic literal amount has been proposed by [16]. The authors combined two types of structural features, the first type includes components and dots features and the second type involves loops and character shapes features. The AHDB database was used to evaluate their proposed method, where a recognition rate of 92.13% has been obtained. The same authors in another work [14] proposed two combination schemes, the first scheme is based on features combination, where four features namely angular, distance, horizontal and vertical span features are considered. In the second scheme, three ELM classifiers are combined using the majority vote technique. The same database (i.e., AHDB) was used to evaluate their proposed method, and the recognition rate was 64.63%.

To some extent, structural features have performed well. However, considering handwritten words as shapes may be affected by word distortion. Because words are most often written spontaneously, the writer may make critical changes to the word shape, such as letter deviations, missed diacritics, and PAWs. This increases the probability of confusing words from different classes.

TABLE 3.1: Related work for Arabic Handwriting recognition.

Feature type	Reference	Features	Classifier	Database	Recognition rate %
Structural	Farah et al [15]	Ascenders, Descenders, Loops,	MLP+ KNN+ FKNN	4800 images from AHDB	94
	Al-Nuzaili et al [16]	Components and dots features shapes features	ELM	AHDB	92.13
	Al-Nuzaili et al. [14]	Angular, distance, horizontal vertical span features	ELM+ majority vote	AHDB	64.63
Statistical	Assayony and Sabri [17]	Gabor filters with Bag of Features (BoF)	KNN	CENPARMI	86.44
	Hassen and Al-Maadeed [19]	Invariant Moments (IV) Histogram of Oriented Gradients (HOG) Gabor filters	SMO	3045 images from AHDB	91.59
	Qawasmeh et al. [80]	SIFT+PCA	SVM	12 classes from AHDB	58.55
	Korichi et al. [81]	LPQ with PML	SVM	AHDB	91.5
	Asmae Lamsaf et al. [82]	Pixels distribution + N-gram	KNN	AHDB	92.7
	Al-Nuzaili et al. [18]	Quadratic angular model based on pixels distribution	ELM	AHDB	83.06
Hybrid	Menasria et al. [22]	Structtural+Statistical features	SVM	AHDB	95.91
	Hassan and Kadhm [24]	Transformation-based and statistical features	ANN	AHDB	95
	Almaadeed et al. [23]	Structural and Local statistical features	HMM	4700 images from AHDB	45
Deep Learning	El-Melegy et al.[20]	CNN with 17 layer	SVM	50 classses of AHDB	96.8
	Eltay et al.[21]	ALexNet and RNNs	SVM	9074 images from AHDB	95.07

3.2.2 Arabic Handwriting Literal Amount Recognition Based On Statistical Features

More recently, some researchers have proposed to describe Arabic handwriting amounts by using statistical features. For Instance, authors in [17] proposed an automatic holistic system for the recognition of handwritten Arabic literal amount based on Gabor filters with Bag of Features (BoF). In order to extract local features, images were filtered by a set of Gabor filters with several scales and orientations. The responses of Gabor filters are arranged then delivered to BoF frameworks. Their proposed method was tested on a database namely CENPARMI which encompasses Arabic handwritten checks, where a recognition rate of 86.44% has been scored. In [19], several statistical features, including Invariant Moments (IV), Histogram of Oriented Gradients (HOG), and Gabor filters, are used to describe images of Arabic amounts. The sequential Minimal Optimization (SMO) method, which is an improvement of the Support Vector Machines (SVM), is utilized for classification. A subset of 3045 images from the AHDB database has been used to assess the proposed system. An average recognition rate of 91.59% was achieved. In [18], an enhanced quadratic feature model was presented, where the achieved recognition score was 83.06% using AHDB database. A new model based on integrating N-gram was proposed in [82]. This system is composed of two parts: the first part concerning word recognition, while, the second part integrates the N-gram model to improve the accuracies obtained from the first step. Authors have used the distribution of pixels as features and KNN for classification. The experiments were carried out on the AHBD database and a recognition rate of 92.7% was achieved. A holistic approach based on SIFT descriptor and PCA for dimensionality reduction was proposed by [80] for recognizing handwritten literal amounts. 58.55% of words from 12 classes of the AHDB database were correctly classified by SVM. In [81] authors investigated the effectiveness of several CNN networks against some statistical descriptors based on the PML model. 91.52% is the best recognition rate that was achieved by the LPQ descriptor.

The state-of-the-art results using statistical features show a good representation compared to structural features. However, pixel distribution measurements are sensible to rotations conditions and sub-word positions.

3.2.3 Arabic Handwriting Literal Amount Recognition Based On Hybrid Features

Certain other researchers have combined several types of features (i.e., structural, statistical, and transformation-based features) for describing Arabic handwritten amounts. Menasria et al. [22] proposed to describe Arabic handwriting literal amounts by combining the statistical feature extracted from the whole image word including local chain code histograms (CCH),

zoning, Zernike moment invariants (ZMI), and the density profile histograms (DPH) with some structural features extracted from different image part (i.e., PAWs). From 61 class in AHDB, 95.91% has been successfully classified using SVM. In [24], images were firstly pre-processed, and then certain transformation-based and statistical features including Discrete Cosine Transform features (DCT) and Histogram of Oriented Gradient (HOG) are extracted from images. 95% of words from the AHDB database have been successfully recognized using Neural Nets classifier. In [23], both structural and local statistical features were extracted and then fed to HMM classifier. A recognition rate of 45% has been attained using a subset of 4700 words from the AHDB database.

3.2.4 Limitations Of Existing Works

We can note that most of the previous works consider the entire image for features extraction and stack all image features in a dense feature vector, which leads to poor feature representation. However, considering different image regions at multiple scales and levels allows extracting a more faithful feature vector by incorporating more discriminative features. Besides, in the last years, several works based on Convolutional neural networks (CNN) have been proposed for improving feature extraction. For instance, authors in [83] proposed a hierarchical feature selection architecture to select just the useful features. In [84], a spectral difference CNN for learning features for super-resolution images is proposed. However, in our work, we opt to propose a feature-independent scheme that is usable by any kind of feature, and which is capable to represent images faithfully by combining advantages of two previous schemes namely ML and PML. While the majority of previous works have considered structural and/or statistical features, we investigate, in this study, the performance of two robust and well-known texture features namely LPQ and BSIF. This is motivated by the fact that handwriting can naturally be described using texture features, as handwritten words are constituted of homogeneous and repetitive parts of the word (PAWs).

3.3 Related Work For Ear Recognition

Currently, ear-based human recognition has become an active area of researches within the biometric community due to its wide broad of potential applications including security, surveillance, applications, and forensic science. Ear recognition is the process of identifying the identity of an individual from his/her ear. Extracting the relevant characteristics play a vital role in well-distinguishing ears and it is considered the most challenging step for any identification ear system [85, 86]. From one point of view, features considered for ear recognition can be classified into five classes namely texture, geometrical, holistic, deep, and hybrid-based methods.

3.3.1 Texture-based Techniques For Ear Recognition

Motivated by their performance, texture-based features have been widely used for describing the ear (Table 3.2). For instance, Benzaoui et al [87] have used LBP, LPQ, and BSIF for describing ear images. Similarly, to enhance this representation, authors in [88] have used Multi-Level Binarized Statistical Image Features (ML-BSIF). In [89], the authors suggest adopting Scale Invariant Feature Transform (SIFT) descriptor for feature extraction, where ear images were first preprocessed using an artificial bee algorithm. A new local descriptor for unconstrained ear images based on a scattering wavelet network (ScatNet) is proposed by [90]. More recently, Hassaballah et al. [91] evaluated in a comparative study the performance of local binary patterns and its variants for ear recognition under controlled conditions. However, the performance drops dramatically in the case of ear images acquired from uncontrolled conditions (e.g., AWE dataset). Authors in [25] have proposed a new descriptor namely robust local-oriented patterns (RLOP), which is supposed to be rotation-invariant.

Despite their promising performance shown in the works mentioned above, most texture-based local features have shown a weak performance in the case of unconstrained scenarios. Moreover, considering global information, which is ignored by this kind of feature, could have a positive effect on the recognition performance.

3.3.2 Geometrical and Holistic-based Techniques For Ear Recognition

On the other hand, some researchers [92–94] have considered geometrical features to represent the ear in which the shape, edges, and outer ear structure measurements have been computed to represent the ear. Recently, authors in [26, 95] showed that considering the shape and ear contour-based features can achieve good performance. Geometrical features are conceptually simple, however, they have limited performance in uncontrolled conditions, where there are variations in terms of illumination and occlusion. Besides, it is very challenging to extract edges in the case of noisy images.

Certain other researchers have used the holistic approach where the ear images are treated as a whole. Several works have been proposed in this context [96–100], where different holistic features including Linear Discriminative Analysis, Eigenear, Haar wavelet, and 1D or 2D Gabor filters are employed.

3.3.3 Deep-Learning Techniques For Ear Recognition

More recently, ear recognition systems have been shifted from handcrafted-based methods to deep learning-based methods (Table 3.3) motivated by their impressive performance in a

multitude of recognition tasks(i.g., share price prediction [101], corn crop disease detection [102], students stream selection [103].. etc.). Dodge et al. [104], proposed an ear recognition system using various Convolutional Neural Network CNN to extract features with transfer learning which are fed then to a shallow classifier. An ensemble of deep learning models with a combination scheme is proposed by [105]. Authors investigated different models including models trained with random weights, pre-trained models, and fine-tuning pre-trained models, where fine-tuned models have proved their efficiency against the other models. In [28], authors designed a simple CNN architecture for ear recognition with varied parameters including learning rate, kernel size, epochs, and activation functions tested on ear images taken from both controlled and uncontrolled environments. Authors in [106] proposed a new framework for grayscale ear images recognition using generative adversarial networks under unconstrained conditions. In [107], authors proposed an ear recognition system based on deep unsupervised active learning tested in both controlled and uncontrolled conditions. A spectral-spatial features based on CNN to describe ear images is proposed by [108]. The authors proposed then an embedding algorithm to fuse multilevel spectral information from the image extracted at each deep layer of the CNN network. The performance of their proposed system was tested using ear images taken from uncontrolled conditions. Omara et al. [27] proposed to learn Mahalanobis distance from features extracted using pre-trained models (i.e., VGG-s, VGGverydeep16, and ResNet). In [109], a comprehensive survey about ear detection and recognition and modified deep learning models based on Faster-RCNN and VGG-19 is presented using several databases from controlled and uncontrolled conditions. More recently, authors in [110] constructed a Deep Residual Networks (ResNet) model with three different transfer learning strategies and SVM classifier tested on several databases including images taken from both constrained and unconstrained conditions. Indeed, it is commonly known that deep CNN-based approaches require a huge amount of data to be trained perfectly.

3.3.4 Hybride-based Techniques For Ear Recognition

To take advantage of the above-mentioned approaches, other researchers have opted for a hybrid approach that combines different types of features (Table 3.4). For instance, authors in [111] proposed a hybrid ear recognition system based on Elliptical Local Binary Patterns (ELBP) extracted from several blocks where the Haar wavelet was used to reduce features dimension. In [112], an ear recognition system based on the combination of Discrete Cosine Transform (DCT) and weighted wavelet transform is proposed. Morales et al. [113] combined local features extracted using SIFT descriptor with some global features to describe ear images. Authors in [114] proposed a combination of local and global features in the frequency domain to describe ear images taken from constrained conditions. More recently, authors in

[115] proposed an ear recognition system for unconstrained conditions. Authors have used a fusion of CNN-based methods and a set of well-known handcrafted descriptors.

3.4 Conclusion

As they are considered one of the most important applications of visual object recognition, we focused our interests in this thesis on two applications which are Arabic handwriting recognition and unconstrained ear recognition. This chapter was divided into two parts. The first part gave an overview of the related state-of-the-art works for Arabic Literal Amount recognition. However, the second part summarized the existing works for ear recognition. Each part is subgrouped according to the nature of the features type. In the next chapter, we will present in detail our contributions for Arabic handwriting recognition and human identification based on the ear print.

TABLE 3.2: Texture based technique for ear recognition.

Feature type	Reference	Features	Classifier	Database	Recognition rate %
Texture	Benzaoui et al. [87]	BSIF	KNN	USTB	98.5
				Delphi 1	97.3
				Delphi 2	97.3
	Korichi et al. [88]	ML-BSIF		IIT Delhi	95.2
	Ghoualmi et al. [89]	SIFT		IIT Delhi	100
				USTB 1	97.2
				USTB 2	94.8
	Hassaballah et al. [91]	AECLBP	Chi-square dissimilarity	IITD 1	96.15 ± 3.85
				IITD 2	98.61 ± 0.73
				AMI	73.57 ± 2.26
				AWE	49.60 ± 3.40
		DRLBP		IITD 1	97.16 ± 1.35
				IITD 2	96.34 ± 2.09
				AMI	71.43 ± 2.56
				AWE	23.50 ± 1.05
		CLBP		IITD 1	96.14 ± 2.28
				IITD 2	98.23 ± 0.84
				AMI	73.71 ± 2.61

TABLE 3.3: Deep Learning techniques for ear recognition.

Feature type	Reference	Features	Classifier	Database	Recognition rate %
Texture	Hassaballah et al. [25]	RLOP	Chi-square dissimilarity	IITD 1	95.13 ± 2.37
				IITD 2	97.98 ± 0.84
				AMI	72.29 ± 1.65
				AWE	54.10 ± 2.01
Deep Learning	Dodge et al. [104]	CNN+Transfer Learning	SVM	AWE	69.25
	Omara et al. [27]	ResNet	Learning Mahalanobis distance	AWE	78.13
	Khaldi and Benzaoui [106]	DCGAN+ CNN		AMI	96
				AWE	50.5
	Khaldi et al. [107]	Deep Unsupervised Active Learning with VGG16		USTB2	100
				AMI	98.3
				AWE	51.3
	Alshazly et al. [110]	Deep Residual Networks		WPUT	81.9
				AMIC	98.6
				AMI	99.6
				AWE	67.3

TABLE 3.4: Hybride based techniques for ear recognition.

Feature type	Reference	Features	Classifier	Database	Recognition rate %
Hybrid	Benzaoui et al. [111]	ELBP+ Haar wavelet	Chi-square dissimilarity	IIT Delhi	94
	Tian Ying et al. [112]	DCT+ weighted wavelet transform	KNN	750 ear images	98.13
	Morales et al. [113]	SIFT+ Global features	KNN	FEARID	91.3
	Shabbou Sajadi [114]	Local and Global features	KNN	USTB-1	100
				IIT125	99.2
				IIT221	97.13
	Hansley et al. [115]	CNN+ HOG	KNN	AWE	75.6

Chapter 4

GFIPML & TRICANet Models For Arabic Handwriting Recognition and Unconstrained Ear Recognition

4.1 Introduction

NOWADAYS , visual object recognition has become the core of recent researches and it has received growing interest by researchers, this is due to its important applications including handwriting recognition, biometric recognition, date fruit classification...etc. This thesis is mainly dedicated to proposing new solutions for two different objects which are Arabic handwriting and unconstrained ear.

In this chapter, we present in detail our proposed methods. We start by introducing the proposed Generic Feature Independent Pymarimd Multi-Level (GFIPML) model for Arabic handwriting recognition. GFIPML is inspired from ML and PML. Thus, we start by citing the limitations of those models. In the second part of this chapter, we present the second contribution under the title " TR-ICANet: a fast unsupervised deep-learning based scheme for unconstrained ear recognition". TRICANet is an unsupervised deep network based on ICA for learning the filters of the convolutional layers and Tied Rank (TR) for histogram normalization.

4.2 Contribution N°1: A Generic Feature Independent Pyramid Multi-Level (GFIPML) Model For Arabic Handwriting Recognition

The main aim of a handwriting recognition system is to simulate human reading capacity by maximizing the recognition rate. More recently, it has received a growing interest by researchers and it has become a very active field of research due to its vital applications including automatic postal mail sorting, historical handwritten documents digitization, automatic checks recognition…etc. Arabic handwriting recognition is considered one of the most challenging tasks in the handwriting recognition domain due to the nature of the Arabic script that makes the recognition task very difficult.

Arabic literal amounts recognition is a typical task of Arabic handwriting recognition, as checks are considered as the fundamental mean in performing finical transactions in several countries. Despite the human capabilities in perceiving checks' literal amount, the error rate is proportional to the number of checks in hand. Therefore, relying on human agents in such a crucial situation could produce several errors in checks processing. In the literature, a considerable amount of works have been proposed to automate the checks recognition process. From one point of view, some existing works [14–16, 18] that deal with this issue considered the structural features, whereas, certain others [17, 19] distinguished literal amounts based on statistical features. A third category includes works [22–24] that combine the aforementioned features alongside with other features such as transformation-based features.

In spite of the significant number of works that are concerned with the automation of checks literal amount recognition, much more efforts have to be done to address limitations of the existing methods. To sum up, these limitations can be summarized as follows:

— It is worth noting that the above cited works consider the whole image at features extraction stage and gather all image features in a compact feature vector. Nevertheless, considering different parts of image on different scales and levels could significantly improve classification outcomes. In the literature, there are several schemes to effectively extract image features including Multi-Level (ML) and Pyramid Multi-Level (PML). While ML and PML have shown to be effective than extracting features from the entire image, they suffer, in turn, from several cons. As instance, shifted letters represent a serious issue leading to a minor performance of ML. More details about limitations of ML and PML are presented in Section 4.2.1.

— Certain methods are relying on structural features which are sensitive to word deformation. Indeed, the same word may be written differently (e.g., line deviation or missed PAWs) by the same writer multiple times. This leads to maximizing the

intra-class variation and increases the confusion between words belonging to different classes.

— Most existing studies have focused on using structural features to describe literal amounts, while too little attention has been paid to use texture features. Actually, different homogeneous and repetitive parts of word (PAWs) that are spreading over image can be considered as texture. Thus, handwriting literal amounts can be characterized using the large arsenal of robust texture features due to the inherent nature of handwriting.

To overcome the aforementioned shortcomings, we propose a novel feature–independent (can be used with the vast majority of existing features) model for Arabic literal amount recognition. The proposed model termed Generic Feature-Independent Pyramid Multi-Level (GFIPML) combines the pros of two existing models namely: Multi-Level (ML) and Pyramid Multi-Level (PML).

4.2.1 Limitations of Multi-Level and Pyramid Multi-Level representations

Over the last decade, several image features extraction schemes such as multi-level (ML) and pyramid multi-level (PML) have been proposed. Generally speaking, features extracted using ML and PML are considered more faithful than those extracted by considering the entire image. In spite of that, numerous issues are encountered when using these schemes for features extraction.

By remembering the idea of Multi-Level representation which is aiming to extract features from several subblocks of the image according to the number of levels, we can notice that ML is suitable for Arabic handwriting recognition as several words differ just on specific blocks (Figure 4.1).

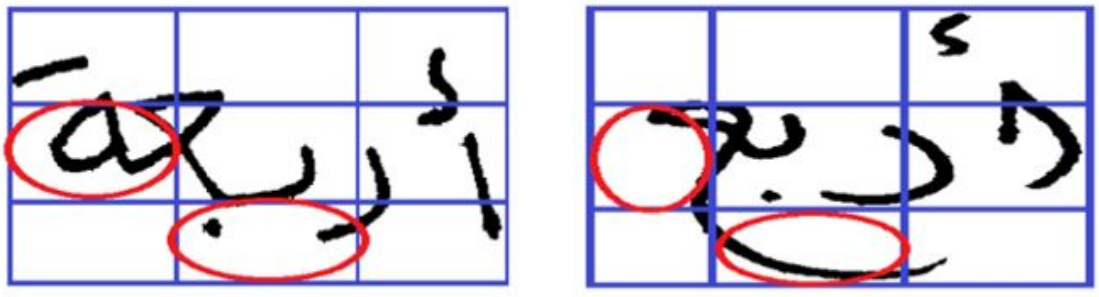


FIGURE 4.1: The importance of dividing word vertically and horizontally.

Moreover, by observing Figure 4.1, it is clearly shown the importance of dividing the image vertically and horizontally. However, considering multiple levels is generally better than considering a single level, as it allows capturing image parts that well-discriminate images.

In spite of the strengths of this ML, it neglects image size which plays a critical and a sensitive role in distinguishing words (similar to human vision where reading is not the same thing from near and far). Moreover, on certain scales, words can be confused with its sub-words (i.e., sub-words contained in the image). As it is shown in Figure 4.2, two other word class (ست and مئة) are totally integrated on the class word سمئة.

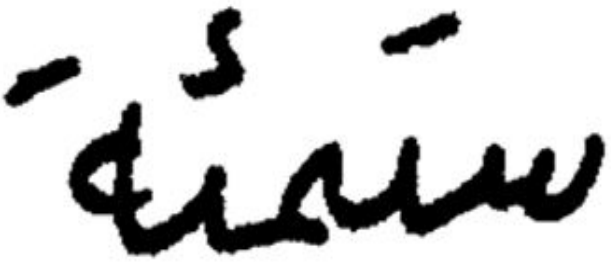


FIGURE 4.2: Multiclass word.

This integration certainly leads to confuse these classes during the recognition process. Hence, to avoid such a problem, the word image is considered at different scales by minimizing the image size every time, and thus, overlapped letters become more distinguishable (Figure 4.3).

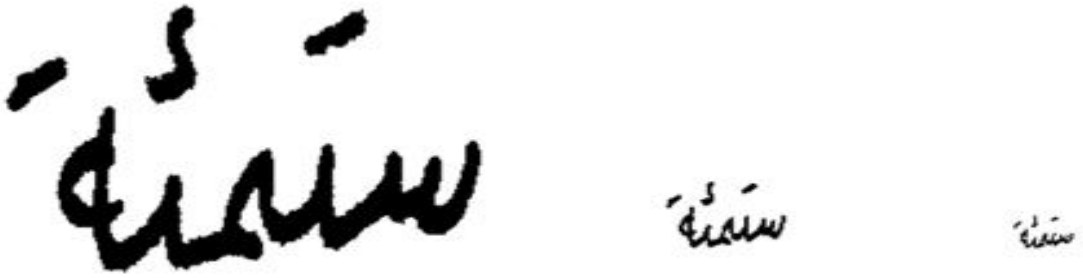


FIGURE 4.3: A handwriting word from different scales.

Another concern with ML is the shifting of character position within the same word, which is considered among the major limitations of ML scheme (Figure 4.4).

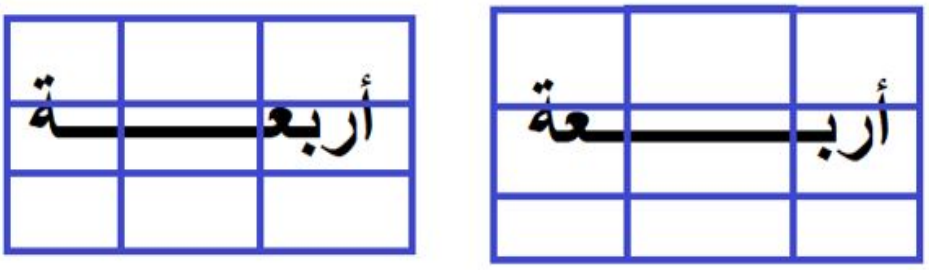


FIGURE 4.4: ML representation limitation example.

Similar to ML, the PML scheme is also suitable for Arabic handwriting recognition because it is interested in the size of the image and studies it on several scales. However, the major limitation of PML in the case of Arabic handwriting recognition is that it takes into account only a single level at each scale (i.e., at each scale only one level is considered).

4.2.2 Generic Feature-Independent Pyramid Multi-Level (GFIPML) model

There is no doubt that the human vision from near is not the same as from a distance which may have a significant impact on identifying the nature of the target object. Furthermore, distinguishing between two objects at the same visibility distance can be difficult if the difference is only in small zones. In this work, we are trying to simulate and study the effect of human eyes in analyzing, perceiving and recognizing objects. To achieve such a goal, we proposed an efficient model that covers eye properties in perceiving objects. Our model, which is called Generic Feature-independent Pyramid Multi-Level model (GFIPML), is based on multi-scale perception to extract features in addition to multi-level decomposition to alleviate confusing different objects. GFIPML takes advantages of both ML and PML by considering different scales, where each scale is associated with multiple levels.

The proposed GFIPML is used for recognizing Arabic handwriting which is motivated by the fact that Arabic words can be distinguished from specific blocks (as proved by using ML in the context of Arabic words) and the importance of zoom scaling in well distinguishing between words (as proved by using PML in the context of Arabic words). As its name indicates, GFIPML is is feature-Independent and can be used with the large arsenal of existing features.

Figure 4.5 illustrates the architecture of GFIPML, where the transition level-to-level is done by reducing the size of the image to half.

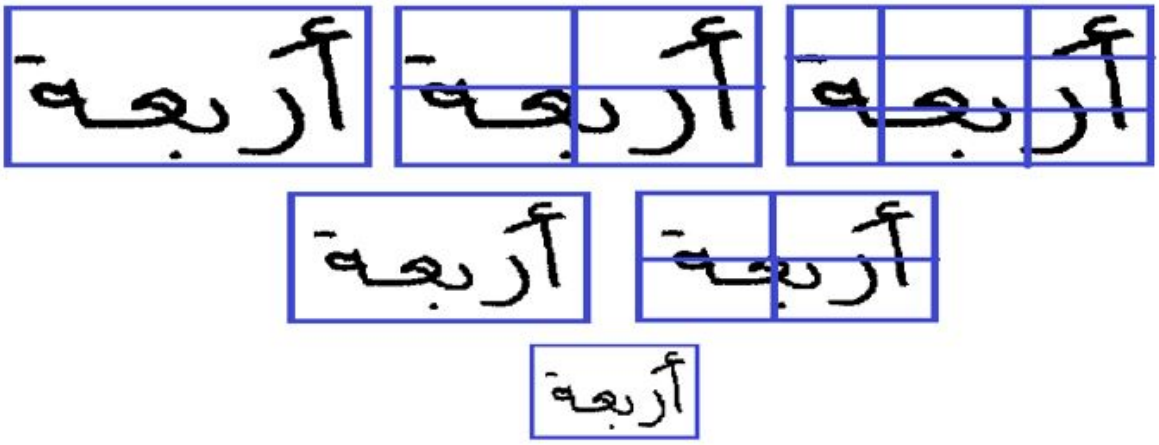


FIGURE 4.5: The architecture of the Generic Feature-independent Pyramid Multi-Level model (GFIPML).

The following steps summarize the principle of the proposed GFIPML model, noting that the required number of levels N should be defined priorly:

1. Divide the input image into (N × N, (N-1) × (N-1), ..., 1×1) blocks by keeping the same size.
2. Extract and save the features from each block. In this stage, the required features of level N are done.
3. Move to the next level by resizing the input image in half and decreasing N by 1. Supposing X is the size of the input image: X ← X/2 and N← N-1.
4. Repeat the above steps until N =1.

Generally speaking, both ML and PML as well as our GFIPML produce a high dimensional feature vector, making matching test probes somewhat computationally heavy. However, compared to ML and PML, GFIPML can achieve a faithful representation of image content and a better performance from only a few number of levels (i.e., a small value of N). A few number of levels imply producing a lower-dimensional features vector (compared to ML and PML), and thus, reducing the processing time required for matching.

The size of the resulting feature vector is computed as follows:

$$|\text{feature vector}| = \sum_{N=1}^{M} \sum_{i=1}^{N} \text{Si}^2 \tag{4.1}$$

Where

— N is the level corresponding to each scale. For example on scale 3, construct a ML architecture with 3 levels.
— S represents the size of the descriptor used for extracting features.
— M is the number of scales.

4.2.3 The Proposed System For Arabic Handwriting Recognition

A handwriting word recognition system is an automatic system that simulates human reading capacities. The proposed system consists of two main stages: features extraction and classification.

As it has been previously mentioned, feature extraction is a key step in any recognition system, as extracted features significantly affect the system yields. It aims to extract relevant characteristics that faithfully represent the image content could improve the system's performance.

For this work, we propose to take advantage of textural features because of their low computational complexity and high speed [116]. More precisely, we opt for using Binarized

Statistical Image Feature (BSIF) [52] and Local Phase Quantization (LPQ) [51] descriptors. Indeed, those features have proven to be very efficient in several recognition tasks including face and iris [52, 117], or palmprint [118]. Hence, we propose to use these two features (i.e., BSIF and LPQ), for the first time, to the best of our acknowledge, in the context of Arabic literal amounts recognition as handwritten words are constituted of homogeneous and repetitive parts of the word (PAWs). To further improve image representation, we consider the proposed GFIPML model for automatic Arabic handwriting recognition.

The next algorithm summarizes the principle of our proposed system for automatic Arabic handwritten recognition where the principle of GFIPML is given details and it can be used with a variety of feature extraction methods. For the sake of clarity, we explain hereafter some used functions that are involved in the algorithm.

— floor(X/Y) is a function that returns the euclidean quotient of X by Y.

— mod(X/Y) returns the remaining of the euclidean division of X by Y.

— S(a:b,c:d) is a function that returns a submatrix (from the line a to b and from column c to d) from the matrix S.

Algorithm 1 Handwriting recognition

Background: a set of images from AHDB database

Input: Handwriting descriptors : HV = HV1 , HV2 , ….HVn, Images , Number of levels : M, Combinaison scheme models

Output: Digital format of the image content : Dg

```
for all image ∈ AHDB test do
    RV(image)← []
    for i=1 to M do
        [X,Y]← size(image)
        for j=1 to i do
            H ← floor(X/j)
            W ← floor(Y/j)
            HL ← mod(X/H)
            WL ← mod(Y/W)
            for xx=1 to X-HL step=H do
                for yy=1 to Y-WL step=W do
                    block ← image(xx:xx+H-1,yy:yy+W-1)
                    RV(image) ← RV(image)+HV(block)
        image ← resize(image,X/2,Y/2)
    First decisions ← fed each RV for the same classifier
    Second decisions ← apply combinaison schemes(First decisions)
    Final decisions Dg ← Majority voting of (Second decisions)
```

As it is shown in the algorithm, the input images of our proposed system are firstly virtually segmented into sub-blocks according to a predefined number of levels. As explained in section 4.5, at the level N, the image is subdivided into NxN blocks the features of each sub-block is combined with the features of the same level (features taken from blocks (N-1) × (N-1), ..., 1×1) as well as with feature of the next levels where the image is firstly resized into the half. Shortly, using GFIPML for extracting features. Several features extraction methods can be used. The obtained features are fed to a specific classifier where the first decision is produced. The second decision is obtained by feeding the probability scores produced by each method used for extracting the features, to several combination schemes where each scheme will give its own decision. The final decision is the result supported by the majority of schemes.

Indeed, we considered a multimodal system in which we fuse the outcomes of unimodal systems to further improve the recognition yields. More precisely, we considered the combination of LPQ features with three variations of BSIF each with a different filter size (13× 13, 15×15 and 17×17), wich we denote respectively BSIF13, BSIF15, and BSIF17. The computed features are fed to the same LDA classifier, the final decision is obtained by the combination of the results of each descriptor at the decision level. We use several combination methods including Product, Sum, Min, Max, Mean and Median, as they are considered to be simple and efficient as well. Moreover, each of these methods has its advantages; the obtained decision is not necessarily the same by all. Thus, to ensure that the combination will give the best and the highest rates, we propose to assign the input image to the class selected based on the majority voting.

The general scheme of the proposed system is illustrated in Figure 4.6:

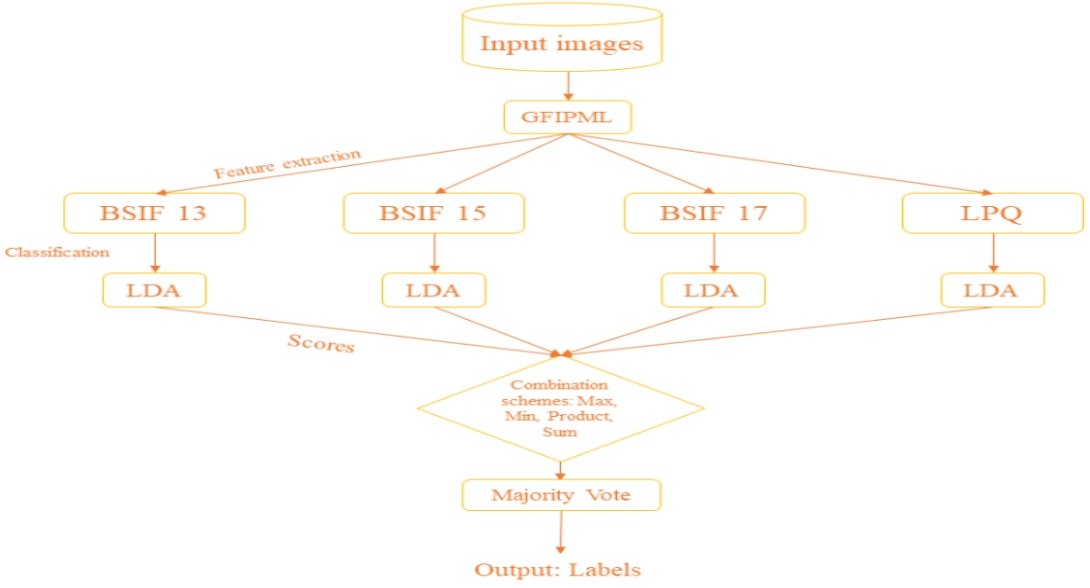


FIGURE 4.6: The general scheme of the proposed system for AHR.

4.3 Contribution N°2: TR-ICANet: A Fast Unsupervised Deep-Learning-Based Scheme for Unconstrained Ear Recognition

As a second contribution, we shifted our interests from Handwriting recognition to Biometric recognition and more specifically on Unconstrained ear recognition. The main goal of an ear recognition system is to determine the identity of a person from the ear print. The term " Unconstrained" refers to the task of recognizing ear images taken from the wild (i,e., images taken from uncontrolled conditions). It should be worth noting that moving from controlled to uncontrolled scenarios or in the wild with more challenging conditions represents the shortcomings of most current ear recognition systems.

As it has been previously mentioned, existing works dealing with ear identification can be divided into two categories: handcrafted-based schemes and deep-based schemes. Despite their performance, handcrafted features are generally limited to constrained scenarios (i.e., where the image is taken under controlled conditions without significant variations of illumination, pose, or scale). Furthermore, as the size of the dataset grows larger, the performance of those features tends to deteriorate significantly (i.e., they are not scalable). On the other hand, most recent works tend to consider deep learning-based schemes. However, as it is commonly known, deep-based schemes may require a great deal of time to perform features learning and extraction, especially for networks with a high number of stacked layers and a huge number of parameters. Thus, in our work, we opted to propose an unsupervised deep model for feature learning with a minimum number of layers and parameters. The proposed unsupervised CNN-like network shares the same structure as the CNN; however, it involves an unsupervised learning process (i.e, without backpropagation) in which the ICA is adopted to learn the filters of the convolutional layers from image patches rather than the iterative process for adjusting weights. The unsupervised learning process makes the network computationally fast. Moreover, the efficiency of the model depends on the learned features used for the convolutional layer. Thus, we proposed using ICA to learn filters, as ICA is considered a useful tool to estimate the filters. Shortly, we propose TR-ICANet, a simple yet efficient and computationally fast CNN-like network for recognizing ear print. TR refers to using Tied rank normalization at the final stage of our network motivated by its great influence proved by other architectures [119] in alleviating the negative effect of bursty features and disparity within blocks histograms. We show that TR-ICANet can outperform CNN architectures in terms of performance while using a low computational budget and processing time.

We can summarize our contribution in the following steps:

— We propose TR-ICANet a simple, yet efficient and speedy, network for automatic ear recognition.

— In order to alleviate the disparity in ICANet histograms, we propose using a normalization technique namely Tied Rank Normalization (TR Normalization) for each histogram block.

— To get rid of unconstrained conditions (e.g., scale and pose variations), we suggest normalizing ear images using CNN.

— To further enhance the recognition outcomes, we considered a soft-max average fusion of CNN-based schemes with CNN-like networks at the decision level using SVM classifier.

In the reset of this section, we present in detail the proposed TRICANet and the steps of the proposed system for unconstrained ear recognition.

4.3.1 ICANet Network For Filter Learning and Feature Extraction

Similar to PCAnet [61], the proposed ICANet contains three main stages (Figure 4.7): learning filters which are used as detectors for the input images, binary hashing, and block-wise histogramming for computing the features. However, we considered using the Independent Components Analysis (ICA) to learn filters instead of PCA, as ICA is considered a useful tool to estimate the filters [52]. Unlike PCA, which assumes some assumptions on the data, there are no criteria to determine the number of components used in ICA, all components are equally important, and they are not necessarily uncorrelated as PCA [120, 121]. Moreover, to alleviate the negative effect of bursty features and disparity within blocks histograms, we propose using Tied Rank (TR) normalization in the final stage of TR-ICANet.

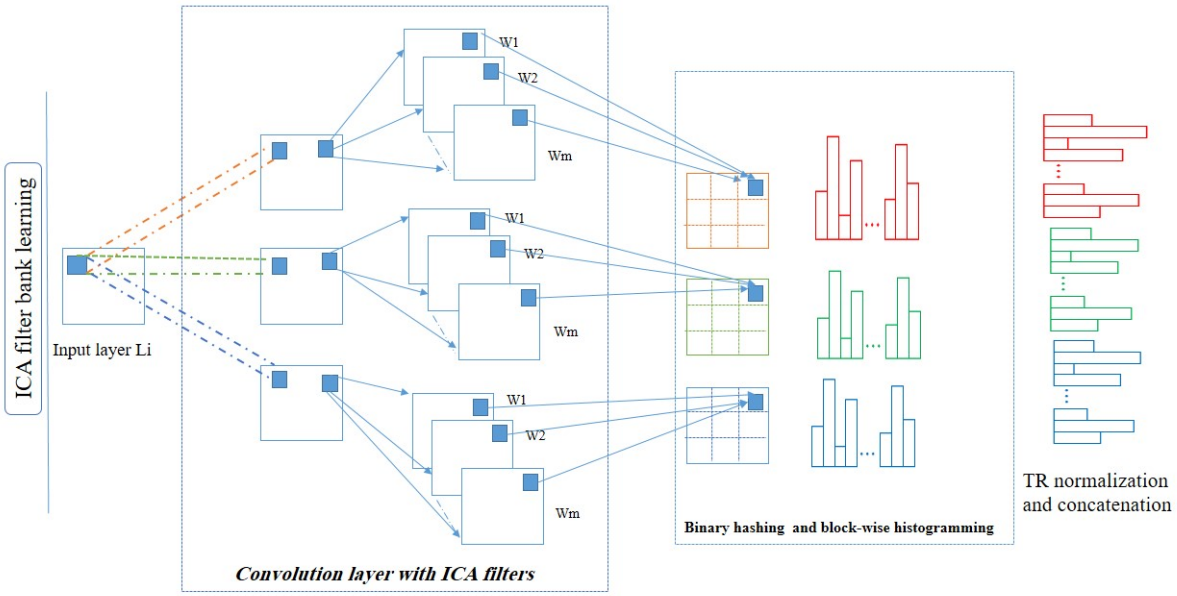


FIGURE 4.7: A detailed block diagram of the proposed TR-ICANet network.

4.3.1.1 Filter Bank Learning

The technique used for generating the filters plays a critical role in the performance of the proposed network where good generated filters produce well image features representation which leads to a good classification. In order to estimate useful and powerful filters, we opt for learning filters via independent component analysis (ICA) as in [52, 122, 123]. However, unlike the cited works, and in order to learn well from the data and generate suitable filters, we suggest first applying data augmentation on the training patches then passing to generating the filter (the parameters used for data augmentation are given in Section 5.2.3.1). In short, the process of generating filters is described as follow:

Given an image patch L of size m × m and set of filters Wi of the same size. The filters are estimated by maximizing the statistical independence of their responses Ri (equation 4.2), which guarantees efficient features for image processing [47].

$$\mathrm{Ri} = \sum_{u,v} Wi(u,v)L(u,v) \tag{4.2}$$

Where u and v represent the pixel coordinates.

Decomposing the filter matrix W into two parts (equation 4.3) is one of the purposes of using standard independent component analysis (ICA) to estimate the independent components.

$$R = WL = UVL = UZ \tag{4.3}$$

As it is clearly shown from the above equation that Z= VL. U is an m × m square matrix that will be estimated by ICA and V performs the canonical preprocessing [124] (i.e., simultaneous whitening and dimensionality reduction of the training samples L) using principal component analysis. The procedure of generating the matrix V and U is summarized in the following steps:

— Given a set of image patches, which are selected randomly from the training image samples.
— Patch mean removal (i.e. from each patch subtract its mean intensity).
— Reduce the dimensionality by keeping only the m first principal components.
— To get the whitened data samples Z, the previously produced principal components are divided by their standard deviation.
— For more details, let C is the covariance matrix of the training samples L. The eigendecomposition of C can be written as follow: $C = EDE^T$. V is defined by

V= $(D^{-1/2}E^T)_{1:m}$. The main diagonal of D contains the eigenvalues of C in descending order. Noting also that the $(.)_{1:m}$ means the m first rows of the matrix in parenthesis.

— Once the whitened data samples are generated and the mean is removed (the same with step 2), the standard independent component analysis methods can be used to estimate an orthogonal matrix U that allows yielding the independent components of the training samples.

— Finally, since the two matrices V and U are generated, the filter matrix W could be easily computed. The obtained filters are used directly by the convolutional layer.

— A sample of 11 filters of size 11 × 11 learned by ICA is shown in Figure 4.8 . It's worth noting that the value of each pixel is multiplied by 10 to make the filters shown clearly.

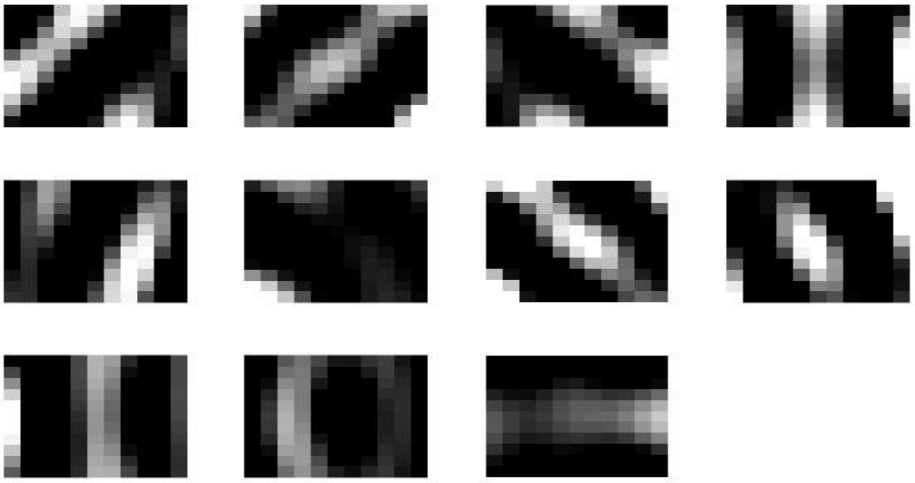


FIGURE 4.8: A sample of 11 learned filters of size 11 × 11

4.3.1.2 Binary Hashing and Block-Wise Histogramming

For each input image L_i, there are D × m outputs corresponding to the number of convolved W_m filters with the D channel of the input image (D=1 if the input image is gray) L_i $\{L_i^d * W_j\}_{j=1}^m$ d=1..D. The outputs are binarized according to a binarization function F where:

$$\mathrm{F(x)} = \begin{cases} 0, & \text{if x<0} \\ 1, & \text{otherwise} \end{cases} \tag{4.4}$$

Afterward, a decimal value D_i is associated to the string code corresponding to each pixel filter response where each pixel value is an integer in the range [0,2^{m-1}].

$$\mathrm{Di} = \sum_{j=1}^{m} 2^{m-1} F(x) \tag{4.5}$$

Then, the image is divided into B non-overlapped blocks. For each block, compute a histogram, which is denoted by H_b^d b=1...B,d=1..D noting that each histogram has 2^m bins. The final feature vector is obtained by concatenating the histograms of all blocks.

4.3.1.3 Tied Rank (TR) Normalization

The main aim of this stage is to alleviate the disparity and the negative effect of bursty features within each block histogram. To do so, each block histogram H_b^d b=1..B,d=1..D is ranked using a tied rank principal which produces a vector $\overline{H}_b^d$ that ranges from 1 to the length of H_b^d. Then, a square root is added to each $\overline{H}_b^d$ to make them more uniformly distributed which forming $V_b^d = \sqrt{\overline{H}_b^d}$. A L_2 norm normalization is applied then for each V_b^d which producing a vector $\hat{V}_b^d$. The final normalized histogram is obtained by concatenating all $\hat{V}_b^d$. The algorithm below summarizes the above steps:

Algorithm 2 Histogram normalization based on Tied Rank (TR) principal

Input: Extracted block-wise histograms of an image : H, Number of channels D, Number of blocks B

Output: TR normalized histogram : v

```
V=[]
for d=1 to D do
    for b=1to B do
        Compute H̄_b^d the the tied rank of H_b^d
        V_b^d = sqrt(H̄_b^d)
        V̂_b^d = L2 normalization of V_b^d
        V=V+V̂_b^d
```

4.3.2 The Proposed System For Unconstrained Ear Recognition

The main goal of an ear recognition system is to determine the identity of a person from the ear print. Generally, any identification system contains two stages, training and testing stage. As it has been previously mentioned, the main aim of the second stage is to test the efficiency of the trained model by classifying the test images based on the extracted features. Our proposed system consists mainly of three phases (See Figure 4.9) which are

respectively preprocessing, features extraction, and classification. More details about each phase are given below:

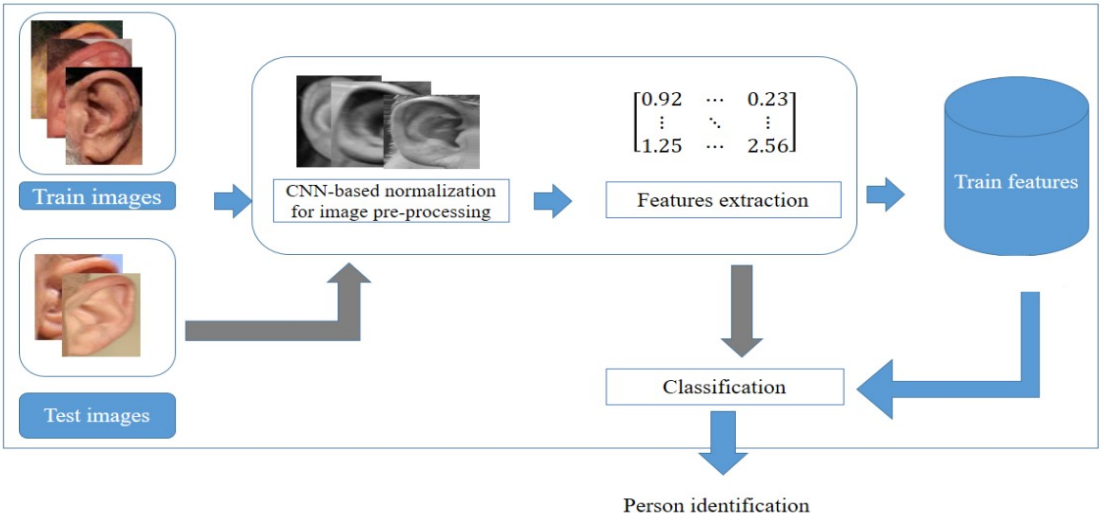


FIGURE 4.9: The flowchart of the proposed ear recognition system

4.3.2.1 CNN-based Image Preprocessing

Despite the importance of the feature extraction step in accurately representing images, image preprocessing enhancements may have a positive effect on recognition rates. Motivated by this, we recommend in this study that images be reshaped into a unified format using PCA and CNN-based image normalization where the ear images are aligned first to get rid of unconstrained environmental conditions. To do so, we use the same technique used by [115]. In short, collections of landmarks are detected at first based on a CNN trained on a collection of images and annotations taken from the ITWE database where ADAM optimization technique was used as an optimizer. We used a training dataset of 15500 images, with a batch size of 36 images, and carried the training for 2000 epochs. To alleviate overfitting caused by the limited data, data augmentation based on PCA was used. Moreover, dropouts were added after all max pooling and the first fully connected layers. The data augmentation procedure is as the following: first, PCA is used on the 2D coordinates of landmarks to acquire the upright orientation of the ear in each image in the training. Then, numerous images are created by rotating the upright from -45° to +45° with a step equal to 3°. Furthermore, each image is also transformed by a random scale where the change is up to 20% of the original image in both axes. This procedure produces 15500 training images which are resized into 96 × 96 pixels and feed to the CNN network (Table 4.1). After having the landmarks, a set of geometrical normalizations including pose and scale are used by applying PCA.

A sample images from the AWE database before and after CNN normalization are shown in Figure 4.10:

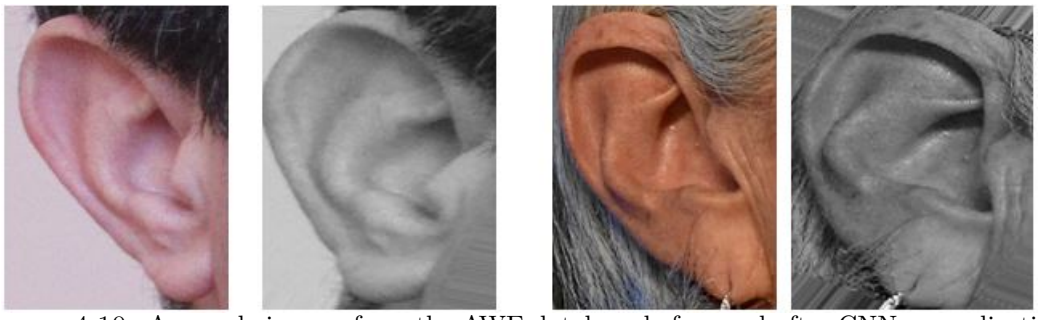


FIGURE 4.10: A sample images from the AWE database before and after CNN normalization

TABLE 4.1: CNN architecture for landmark detection

N°	Type	Input	Filter	Stride	Drop	Output
1	Conv/Relu	96×96×1	3 × 3 × 1 × 32	1	10%	96×96×32
2	MaxPool	96×96×32	2 × 2	2		48×48×32
3	Conv/Relu	48×48×32	2 × 2 × 32 × 64	1	20%	48 × 48 × 64
4	MaxPool	48 × 48 × 64	2 × 2	2		24 × 24 × 64
5	Conv/Relu	24 × 24 × 64	2 × 2 × 64 × 128	1	30%	24 × 24 × 128
6	MaxPool	24 × 24 × 128	2 × 2	2		12 × 12 × 128
	Flattening	12 × 12 × 128				18,432
7	Fc/Relu	18,432			50%	1000
8	Fc/Relu	1000				1000
9	Fc	1000				110

It's clearly noticed that in this study we used two different data augmentation techniques. The first one is based on shift, rotation, and zoom as described in the data augmentation section 5.2.3.1. Data augmentation based on these parameters has been investigated in [104] and high performance was obtained. However, the second one is based on PCA (in the preprocessing section 4.3.2.1) where authors in [115] proved the effectiveness of this technique in the context of ear recognition. Motivated by the achieved success of these two techniques, we suggest including both of them in the proposed model. In order to generate useful and powerful filters used in the filter bank learning stage and since the available data are limited, we augment the data training images using the first technique. The second technique based on PCA was used for data augmentation to avoid overfitting in the CNN trained part.

4.3.2.2 Feature Extraction and Classification

For any recognition system, feature extraction stage represents the most important step where the relevant characteristics are extracted to faithfully describe the image. In this study, the deep features are extracted from images using the proposed TR-ICANet model as described in the above section. To perform the identification, the Support Vector Machine (SVM) classifier with a linear kernel is used.

4.3.3 Multimodal Scheme For Human Ear Identification

Motivated by the high success of multimodal systems compared to unimodal ones, we propose a multimodal system based on fusing several models at score level. Figure 4.11 illustrates the general scheme of the proposed multimodal system:

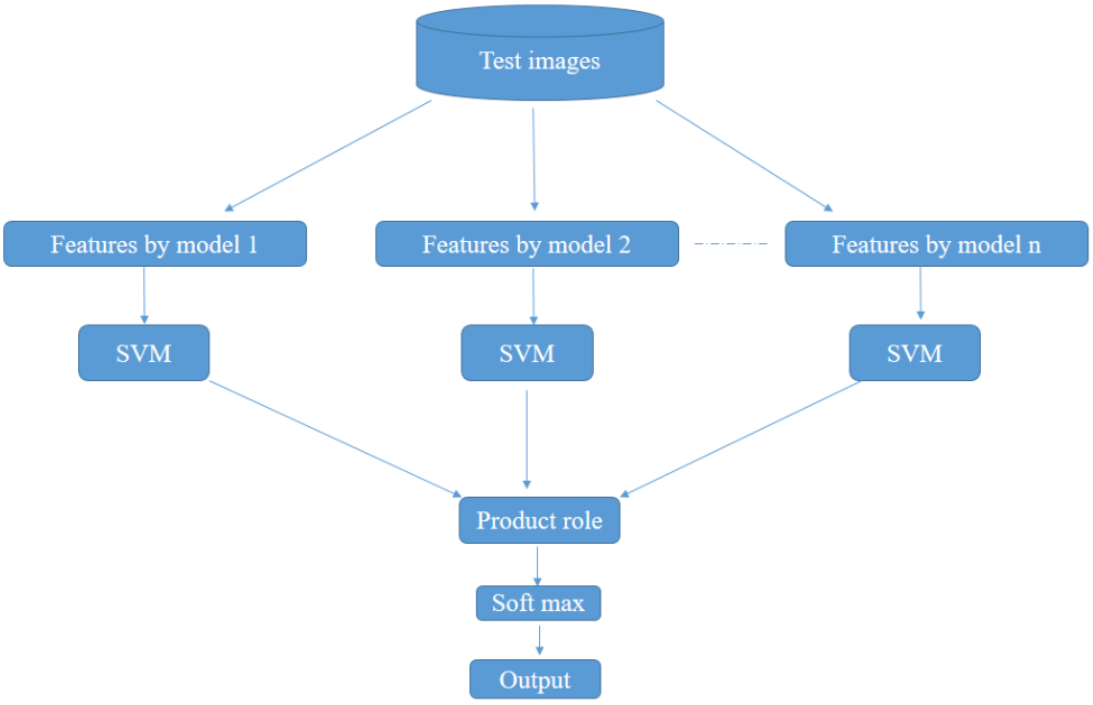


FIGURE 4.11: Multimodal scheme for human ear identification

For the sake of clarity, after extracting the relevant features using different models (ICANet, PCANet, ..etc.), each of those extracted features are fed then to linear SVM for the identification which generates a score representing the probability that an ear corresponds to a person x. The final decision is obtained by fusing scores of all SVMs. More precisely, considering N SVM classifier (n=1..N) where N is the number of models, for each input image feature, the score of belonging to one of the M persons denoted as $S_{i,n}$ is computed. The final decision is the max value of C_i (i=1..M) where Di = $\prod_{n=1}^{N} S_{i,n}$.

In this study, we tested the proposed multimodal system based on four feature extraction methods which are TR-ICANet, TR-PCANet, Alexnet, and VGG19. TR-PCANet is a modified version of PCANet, where we propose to add a TR normalization stage at the final.

4.4 Conclusion

Despite the advanced solutions that have been presented, Arabic handwriting recognition and Unconstrained ear recognition are still from the challenging tasks of visual object recognition and they are still at the level of the experiment where several efforts can be done. In this chapter, we present in detail our proposed methods for feature extraction where we proposed two models. The first model is GFIPML for Arabic handwriting feature extraction which is based on studying the image in several levels in different zooming scales. However, in the second contribution, we proposed a new unsupervised deep network-based ICA for filter learning and Tied Ranks for histogram normalization. In the next chapter, we will present the obtained results with the parameters of each model.

Chapter 5

Experimental Results and Discussion

5.1 Introduction

Due to its indispensable importance and the exquisite solutions that it gave to our daily life problems, the pattern recognition domain has become the core of recent researches and it has received a growing interest in the last two decades. It should be worth noting that that the way and the methods used for feature extraction play a decisive role in the recognition outcomes. In this thesis, we focused our interest on two important applications of visual object recognition: Arabic handwriting recognition and Unconstrained ear recognition. In the first application, we proposed a new Generic Feature Independent Pyramid Multi-Level GFIPML model for features extraction and it can be used for a variety of feature extraction methods. For the second contribution, we proposed a new simple yet efficient supervised deep learning network for feature extraction.

This chapter is divided into three parts. The first part concerned the presentation of the databases used for evaluating the performance of our systems. Then, the results of the first contribution are presented. In this part, we will present the experimental results of ML, PML, and GFIPML models, and the implementation details of our proposed system for handwritten literal amount recognition. However, the results and all the experimental setups of the proposed TRICANet are presented and discussed in the third part. Moreover, to assess the performance of the proposed systems, at the end of each part, we provide a comparison against other related studies that are using the same databases. followed by the analysis of the second contribution results.

5.2 Experimental results and discussion

5.2.1 Databases

5.2.1.1 AHDB Database For Arabic Literal Amount Recognition

The recognition of Arabic handwriting literal amounts has attracted the attention of many researchers and it has become an active subject of research and a challenging task during the last two decades. In this study, the public AHDB database [125] was used to evaluate the performance of our proposed model for recognizing Arabic handwritten amounts. The database contains the most frequently used words in checks. It is made up of 70 classes representing different words, where each word is written by 105 different writers. We divide the dataset, which is made up of 6615 images, into three folds each of which contains 2189, 2183, and 2243 images, respectively. In each cross-validation iteration, two folds are used for training and one for testing.

A sample of each class from the database is shown in Table 5.1. An instance from the database images is shown in Figure 5.1

TABLE 5.1: Arabic words used to express amounts on checks extracted from AHDB database

N°	Arabic name	N°	Arabic name	N°	Arabic name	N°	Arabic name	N°	Arabic name
1	أحد	14	خمسة	27	تسعة	40	ستة	53	ثلاثمئة
2	احدى	15	خمسمئة	28	تسعمئة	41	ستمئة	54	ثلاثمائة
3	واحد	16	خمسمائة	29	تسعمائة	42	ستمائة	55	عشرون
4	ثمان	17	أربع	30	تسعون	43	ستون	56	عشرين
5	ثمانية	18	أربعة	31	تسعين	44	ستين	57	اثنان
6	ثمانمئة	19	أربعمئة	32	لا	45	عشر	58	اثنين
7	ثمانمائة	20	أربعمائة	33	سبع	46	عشرة	59	مئتين
8	ثمانون	21	أربعون	34	سبعة	47	ثلاثون	60	مائتين
9	ثمانين	22	أربعين	35	سبعمئة	48	ثلاثين	61	ألفان
10	اثنى	23	مئة	36	سبعمائة	49	ألف	62	ألفين
11	خمسون	24	مائة	37	سبعون	50	آلاف	63	غير
12	خمسين	25	مليون	38	سبعين	51	ثلاث		
13	خمس	26	تسع	39	ست	52	ثلاثة		

FIGURE 5.1: A sample images from AHDB database.

5.2.1.2 AWE Database For Unconstrained Ear Recognition

For the second contribution, the public Annotated Web Ear (AWE) dataset [126] was used to evaluate the performance of our proposed model for human ear identification. The AWE dataset contains mainly 1000 human ear images from 100 different persons (10 images per person) distributed between males and females of different ethnicity and ages. The images of the AWE are taken from uncontrolled conditions in the wild with different sizes, varying illumination, and different viewing angles. Sample images from the AWE dataset are shown in Figure 5.2:

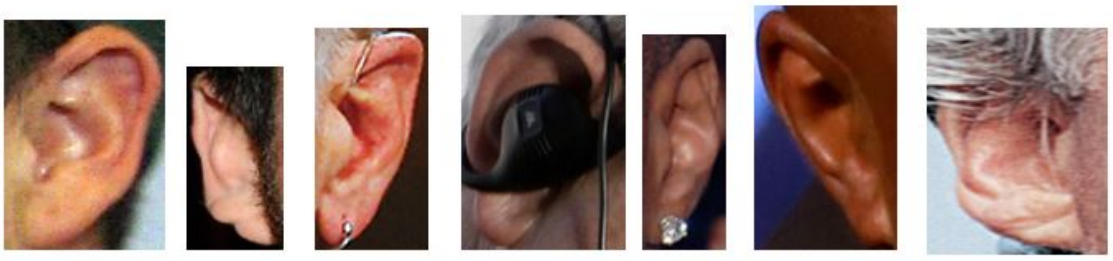


FIGURE 5.2: A sample images from AWE database.

5.2.2 Experimental Results For The First Contribution: A Generic Feature Independent Pyramid Multi-Level (GFIPML) Model For Arabic Handwriting Recognition

The existing literature shows that the majority of Arabic handwriting recognition researchers follow the holistic approach where features are extracted from the whole image. The main objective of this paper is to construct a new efficient model for extracting features from images regardless of the method used for computing and extracting features.

In order to assess the performance of our proposed model in improving the recognition rates, we carry out several experiments by considering two existing models namely Multi-Level (ML) and Pyramid Multi-Level (PML). Noting that ML and PML architectures were proposed and tested just on some biometric recognition systems.

As we have said in the previous chapter that BSIF descriptor is based on convolving images with a set of filters. After having to experiment with different filters, three filters with different sizes have yielded the best results using a fixed number of bits (11 bits). Those filters are 13 × 13, 15 × 15 and17 × 17, which we have denoted respectively by BSIF13, BSIF15and BSIF17. Those three filters were selected from 7 filters because they yielded the best results after conducting several experiments with different filter size. The obtained results is the average of all sizes are illustrated in Figure 5.3. Noting that all the results reported in this section are obtained by using LDA classifier. A series of experiments were conducted using four classifiers, among which LDA is selected, as it has shown better performance compared to the three remaining classifiers namely KNN, Naïve Bayes and SVM.

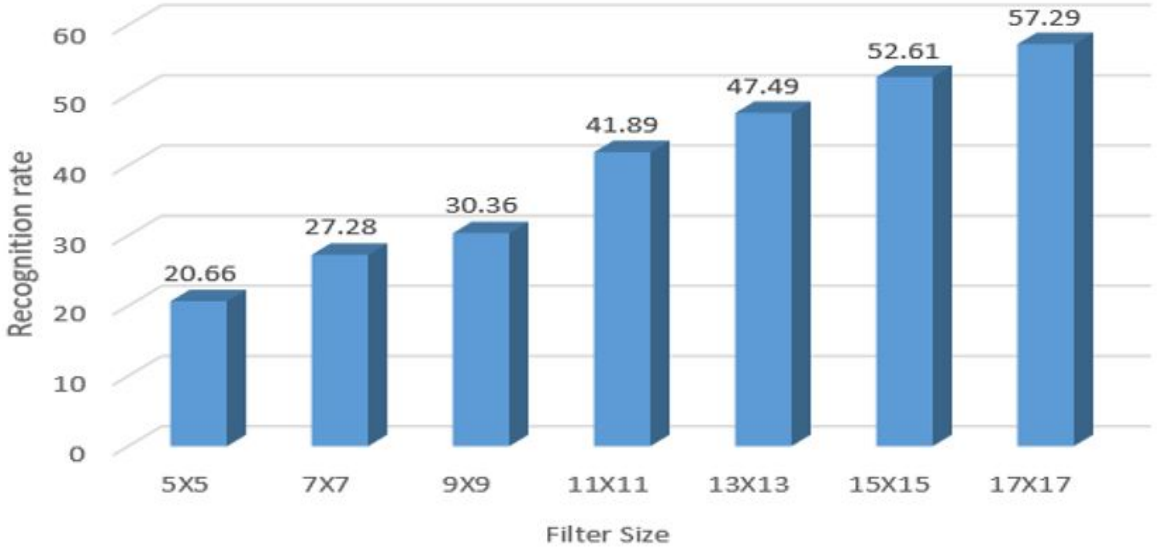


FIGURE 5.3: The experimental results of several filter size from BSIF.

It is clearly shown that from size 11×11 the recognition rate starts decreasing, which can be explained by the fact that the low filter sizes, captures poor information from small zones that can't represent even a word diacritic or a character part. The medium and the high size (13×13, 15×15 and 17×17) can be lying on a character or a sub word which made the word well represented.

In addition to these descriptors, the LPQ is also used in our experiments in order to validate the superiority of our model.

5.2.2.1 Multi-Level (ML) experiments

In this part, we will discuss and analyze the efficiency of the ML architecture on Arabic handwriting recognition. To do so, several tests were carried out by considering different levels with proposed descriptors. The reported results represent the average recognition rates of the three folds, as shown in Table 5.2:

Table 5.2: Experiment results using ML representation with different levels.

Level	Recognition Rates %			
	LPQ	BSIF13	BSIF15	BSIF17
1	47.13	47.49	52.61	57.29
2	75.69	73.34	77.72	80.97
3	86.68	84.95	87.53	88.87
4	**88.78**	**87.72**	89.12	90.19
5	88.02	88.42	**89.60**	**91.04**
6	87.22	87.54	88.93	90.56

The first level without any division of the image that represents the global architecture (features extracted from the whole image) gives low recognition rates compared to the remaining levels. This is may be attributed to the fact that key features of the image are pooled in same feature vector.

Increasing the number of levels play a decisive role in distinguishing different classes, as it has been explained above. It is clearly observed in the table that the recognition rate increases by increasing the number of levels until the fourth level. However, the results stabilize after the fourth level. Increasing the number of levels makes the block size very small in the advanced levels, which may not be useful in distinguishing different classes. In addition, adding more levels incurring more computation cost.

5.2.2.2 Pyramid Multi-Level (PML) Experiments

The main principal of the PML model is to extract features from several blocks of an image in different sizes. To reach such an aim, we conduct several experiments with different levels (level on PML model means the scale of zooming).

The obtained results are shown in Table 5.3, noting that the mean size of our images is 256×256, for this reason, we stopped in the fourth level where images are of size 32×32 because after this size images will very small and the words that contain them disappear.

It is clearly shown that there is a remarkable improvement in recognition rates for all levels compared to the ML model. This is mean that considering different scales of images play an important role in improving the recognition rate; each scale detects new characteristics in the new decreased image size that complete the characteristics obtained by the previous

Table 5.3: Experiment results using PML representation with four levels.

Level	Recognition Rates %			
	LPQ	BSIF13	BSIF15	BSIF17
1	54.17	65.82	67.22	66.81
2	81.04	85.47	84.96	83.85
3	90.83	90.52	90.05	89.92
4	**91.52**	**91.36**	**92.05**	**91.56**

ones, which make the features that represent the image very riches with the most pertinent information to well distinguish it from the other images.

5.2.2.3 Generic Feature-independent Pyramid Multi-Level model (GFIPML) Experiments

Our proposed GFIPML model is based on considering the advantages of the two former models. The idea behind the GFIPML model is to capture the characteristics of Arabic words by studying the images on several scales, and considering different levels at each scale i.e., Multi-Level (ML).

Table 5.4: GFIPML model results by using LPQ and BSIF with three derivation descriptors.

Level	Recognition Rates %			
	LPQ	BSIF13	BSIF15	BSIF17
1	54.17	65.68	67.91	66.81
2	85.41	88.95	88.92	88.18
3	**94.17**	**94.18**	94.34	94.41
4	93.87	93.56	**94.5**	**94.54**

Several words can be confused by considering a single copy of them on each scale, increasing the number of copies (our proposed model by constructing a ML representation at each zooming scale) allow differentiating more between these words, this was proved by the obtained results done in Table above. The obtained results proved the efficiency of GFIPML against PML and ML models. Furthermore, despite the complex architecture of GFIPML compared

to the other architectures, GFIPML produces a lower-dimensional feature vector, making the matching process faster compared to ML and PML. This is attributed to the capability of GFIPML to capture the relevant information from a few number of levels, whereas, other architectures require a higher number of levels to approach the GFIPML performance.

Based on the results reported in Table 5.4, it has been shown that the best recognition rates for GFIPML were obtained from the third level. Thus, features are extracted from 20 blocks $(3\times3+2\times2+1)+(2\times2+1)+1$. However, regarding ML and PML, the best recognition rates are obtained in the fifth and fourth levels, respectively (features extracted from 50 blocks for ML and 30 blocks for PML). This significant difference in the number of blocks makes very different the processing time required by the ML and PML architectures and ours.

Figure 5.4 summarizes some comparison statistics between our model and the two others:

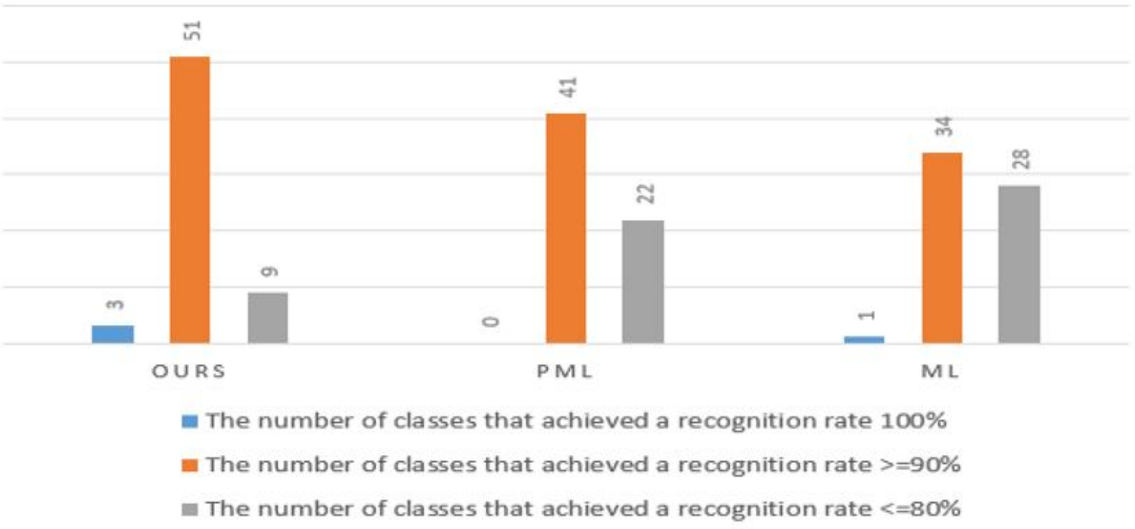


FIGURE 5.4: Some comparison statistics between our model, PML and ML model.

The problems related to the nature of the Arabic language that made confusion in the recognition process are well treated by our architecture.

Table 5.5 illustrates some confused samples in which our proposed method has success to recognize them where the other models fail.

Finally, to summarize the comparison between ML, PML, and GFIPML model in terms of accuracy, all the obtained results are summarized in Table 5.6 where it's clearly shown the high performance of the proposed GFIPML.

TABLE 5.5: Some misclassified samples where our model correctly classified them

Samples	ML Architecture	PML Architecture	GFIPML
	ستة	خمس	خمسون
	خمسة	خمسون	خمسين
	أربعة	أربعة	أربعمئة
	اثنين	اثنين	مائتين
	عشرة	عشرون	عشرين
	ستون	سبعين	ستين

TABLE 5.6: Comparison between ML, PML, and GFIPML in terms of accuracy

Model	Best recognition rate per descriptor %			
	LPQ	BSIF13	BSIF15	BSIF17
ML	88.78	87.72	89.6	91.04
PML	91.52	91.36	92.05	91.56
GFIPML	**94.17**	**94.18**	**94.5**	**94.54**

5.2.2.4 Multi-modal System Results

As we have mentioned in chapter 2, in order to improve the recognition rate outcomes, we opt for using a multimodal system in which we fuse the outcomes of unimodal systems. More precisely, each input image descriptor (BSIF13, BSIF15, BSIF17, and LPQ) is fed to the same LDA classifier. After having obtained the probability scores of the individual classifiers and fed them to the decision-level rules, the appropriate class is determined based on the majority voting technique over the decisions provided by these rules (i,e., Product, Sum, Min, Max, Mean, and Median).

For the sake of clarity, we illustrate the need for using majority-voting technique by the following example (Table 5.7). For three classes namely C1, C2 and C3 and three classifiers denoted as D1, D2 and D3, we detect the final decision using the majority voting.

From Table 5.7, we notice that the final decision yielded by the Max, Sum and Mean methods is C2, whereas that obtained by Min and Product methods is C3. On the other hand, we

TABLE 5.7: The principal of some combination schemes.

	D1	D2	D3	Max	Min	Sum	Median	Mean	Prod
C1	0.0	0.6	0.5	0.6	0.0	1.1	0.5	0.37	0.0
C2	0.8	0.0	0.4	0.8	0.0	1.2	0.4	0.4	0.0
C3	0.2	0.4	0.1	0.4	0.1	0.7	0.2	0.23	0.008
Max				0.8	0.1	1.2	0.5	0.4	0.008

can see that only the Median method has decided to assign the input image to C1.

Each of these methods has its own advantages; the obtained decision is not necessarily the same by all. Thus, to ensure that the combination will give the best and the highest rates, we propose to assign the input image to the class selected based on the majority voting. In the example, the actual class of the image is C2 and the final decision is C2. However, if the majority vote was chosen as a combination scheme firstly without applying any method from the aforementioned ones, the final decision will be the class C1 (It is chosen by classifier D2 and D3). We didn't choose the majority vote technique at first because it has several limitations.

The detailed results obtained by each descriptor before and upon the combination for the three folds are illustrated in Table 5.8.

The obtained results proved that the decision taken by more than one descriptor is the best decision. As instance, for the first fold, the best rate by following the unimodal strategy is 93.92% which has been obtained by the LPQ descriptor, this rate is low compared by those of multimodal strategy where the lowest rate was 95.01% and the highest rate was 96.5% by taking product combination scheme. Moreover, it clearly showed that our proposition by taking the majority chosen class has given the best results.

Due to the nature of features, some combination methods yield good results whereas some others don't. Thus, it is not convenient to say that a method is better than a method because it depends on the nature of the data so taking the decision obtained by the majority will ensure the best decision that makes our system efficient and outperforms very recent researches using the same AHDB database.

Moreover, to confirm the generalization capability of the proposed system, we check the over-fitting issue. A cross-validation procedure is adopted such that in each iteration 70%, 50%, and 30%, respectively, are used for training, and the remaining (i.e., 30%, 50%, and 70%) are used for testing. The obtained results are shown in Fig.5.5:

TABLE 5.8: Experiment results using our model with combination scheme on AHDB database.

Methods	Folds			Average
	1	2	3	
LPQ	93.92	93.49	95.1	94.17
BSIF 13	93.31	94.37	94.88	94.18
BSIF 15	93.32	94.5	95.2	94.34
BSIF 17	93.81	94.19	95.23	94.41
Max	95.01	95.34	96.16	95.51
Min	95.38	96.3	96.26	95.98
Sum	95.34	96.25	97.01	96.2
Mean	95.34	96.25	97.01	96.2
Prod	96.5	96.4	96.75	96.25
Median	95.24	96.2	97.11	96.18
Majority of methods	95.84	96.62	97.06	**96.5**

FIGURE 5.5: Experimental results against over fitting situation.

The obtained results proved the solidity and the efficiency of our proposed method against over-fitting issue.

5.2.2.5 Comparison With State Of The Art

❶ With structural/statistical-based techniques

In order to prove the superiority of our proposed Arabic handwriting literal amount recognition system, an experimental comparison with other recent and relevant works has been conducted, the results are summarized in Table 5.9.

The most closest work to us is that proposed by [22]. Authors have used two folds to evaluate their system; by following the same way as them, we obtained 94.2% as a recognition rate. Thus, our system significantly outperform theirs by 5.07%.

Additionally, their system failed to recognize six words سبعمائة, تسعمائة, ثلاث, ثلاثة, خمسون, تسعون as each pair of those words strongly resemble each other. In the contrary, our system has successfully recognizing the same words (more than 97%). This is due to the proposed model where the probability of confusion between words is very low excluding the word سبعمائة that yielded the lowest recognition in our system with a recognition rate of 88.5%. This is may be attributed to the fact that several images that belong to this class written in a manner (distorted character) that makes very difficult, even for humans, to differentiate them.

The other aforementioned works are used three folds for cross validation phase (the universal number of folds) where we achieved 96.5%. The only work that achieved a recognition rate more than 92% is [127]. However, authors used just few samples from the database to evaluate their system. It's also clear that our system based on GFIPML model has the highest accuracy rate on the AHDB database compared to which are recently published [82].

❷ With deep learning-based techniques

In the literature, little attention has been devoted to using deep learning-based techniques for the recognition of Arabic handwritten literal amounts. In [20], authors have used a deep Convolutional Neural Network CNN architecture based on 17 layers applied on 50 classes extracted from the AHDB database where a recognition rate of 97.8% by using data augmentation and 96.8% without using data augmentation were achieved. By applying the same experimental conditions as them (the same number of classes and without data augmentation), we have obtained a recognition rate of 98.39% by following our proposed method, which proves its efficiency. However, in [21], the authors explored and investigated several deep learning architectures including ALexNet and RNNs for recognizing Arabic handwritten literal amounts. In their work, authors presented two models for the recognition: the first one is based on a holistic word recognition model, and the second one is based on a character recognition model. For a fair comparison, we compared our work by following the same approach (holistic model for word recognition). The authors mentioned that the best recognition rate achieved is 95.07% based on AlexNet with data augmentation.

Table 5.9: Comparison with the state of the art.

Authors	Method	Model	Accuracy %
Menasria et al [22]	Some statistical & structural features	Global	89.13
Assayony and Mahmoud [17]	Gabor filters with Bag of Features	Global	86.44
HASSAN et al. [127]	(HOG,DCT) with Neural Networks	Global	95
AL-NUZAILI et al. [14]	Some structural features	Global	92.13
Lamsaf et al. [82]	Some statistical features with N-gram model	Global	92.7
Qawasmeh et al. [80]	SIFT with PCA	Global	58.55
Korichi et al. [81]	LPQ	PML	91.52
Proposed system	(BSIF with 3 filters ,LPQ)+ LDA	Proposed GFIPML	**96.5**

Table 5.10 summarizes some comparisons between our work and the other proposed by [20] and [21] :

TABLE 5.10: Comparisons with deep learning-based technique.

Authors	N° of simples used	Method	Data augmentation	Accuracy%
El-Melegy et al.[20]	5250	CNN based architecture	Yes	97.8
			No	96.8
Eltay et al.[21]	9074	AlexNet	yes	95.07
Proposed system	5250	Proposed GFIPML	No	**98.39**
	6615		No	**96.5**

To ensure a fair comparison, the experiments should be under the same conditions. From Table 5.10, it is clearly shown the superiority of the proposed method against very recent works that use deep learning techniques.

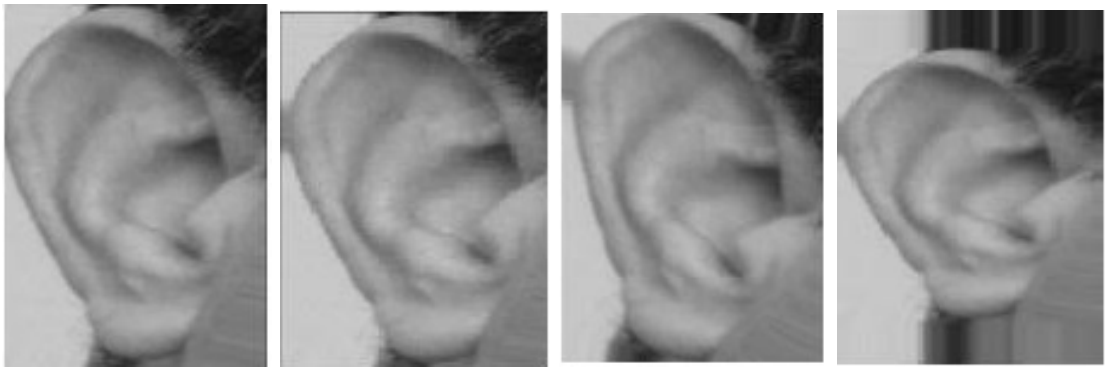
5.2.3 Experimental Results For The Second Contribution: TR-ICANet: A Fast Unsupervised Deep-Learning-Based Scheme for Unconstrained Ear Recognition

In this section, we will present in detail the parameter tuning of the proposed TR-ICANet scheme, followed by introducing our multi-modal system for human ear recognition where an extensive comparison was carried out with the deep CNN-based feature extraction and PCANet model as well as with the relevant state-of-the-art methods. To assess the performance of the proposed model, we conduct experiments on some challenging datasets taken from unconstrained conditions. it should be worth noting that the evaluation metric used in this study is Accuracy.

5.2.3.1 Data augmentation

In order to generate useful and powerful filters, we propose to augment the number of training data used in the filter bank learning stage since the available data are limited. To do so, for each normalized ear image, we propose to generate three additional copies based on three criteria which are width shift, rotation, and zoom where their ranges are respectively [-5,5], [0°,17°], and [0.7,1.8]. Figure 5.6 shows a sample of each generated copy:

It is also worth noting that all images have been filtered with the Gaussian filter and resized into the same size (175×80) to preserve the aspect ratio of image height and width.



(A) Original image (B) Shifted image (C) Rotated image (D) Zoomed image

FIGURE 5.6: Data augmentation parameters.

5.2.3.2 ICANet Parameters Tuning

The performance of ICANet is heavily dependent on well tuning the parameters. Indeed, three parameters have the greatest influence on the recognition outcomes which are the number of filters, their sizes, and the size of blocks used in block-wise histogramming stage. Thus, we conduct extensive experiments to determine the best parameters that yield high identification rates. The obtained results are reported in Table 5.11. According to the high

TABLE 5.11: Experiments for ICANet parameters tuning

N° filters	Filter size	Histogram size	Identification rate%
5	5x5	22x22	39
7	7x7	22x22	42.5
9	9x9	22x22	47.5
11	11x11	22x22	41.5
13	13x13	22x22	28

identification rates obtained, we fixed the value of each parameter as follows: Number of filters=9, Size of each filter= 9×9, and the size of each block is 22×22. Furthermore, to assess the efficiency of the proposed enhancement methods (pre-processing based on CNN and TR histogram normalization), four subset experiments were performed by changing the values in each experiment (Figure 5.7):

— The first experiment corresponds to the results obtained by using the original AWE images without CNN pre-processing and feeding the resulting block-wise histograms to SVM without TR normalization.

— The second experiment aims to investigate the impact of including TR normalization stage without any pre-processing.

— In the third experiment, images are pre-processed with CNN but block-wise histograms are not normalized with TR normalization.

— The fourth experiment represents the results by incorporating both parameters: passing images through CNN pre-processing and adding TR normalization in the ICANet final stage.

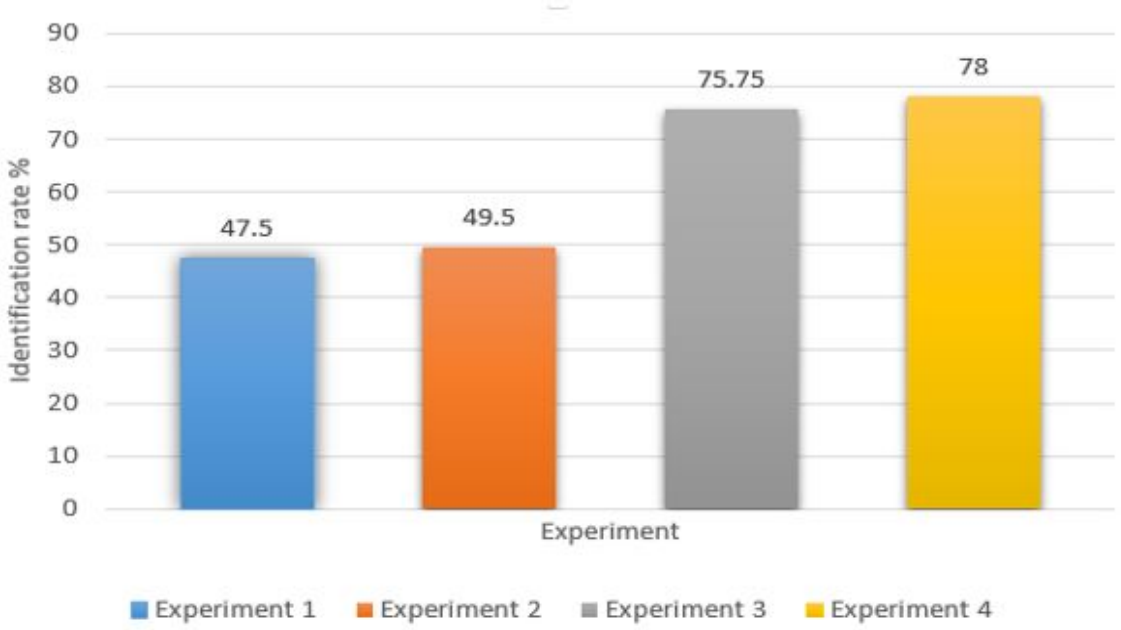


FIGURE 5.7: The influence of CNN-based normalization and TR histogram normalization.

From Figure 5.7, it is clearly shown the high positive effect of the proposed CNN normalization for pre-processing and TR histogram normalization where TR augments the results with 2% without considering CNN enhancements. However, CNN image enhancements surprisingly increase the identification rate with a percent of 28.25%. Considering both CNN for pre-processing and TR histogram normalization yields to obtain very high results that outperform those achieved based on the original images with 30.5%.

5.2.3.3 Comparaison With PCANet and Deep-Learning Models

To ensure the effectiveness of the proposed TR-ICANet model as a feature extractor, an extensive comparison with some feature extraction methods was carried out using SVM classifier. In particular, we consider comparing our method with PCANet and pre-trained models including AlexNet, VGG19, VGG16, and Resnet50. The obtained results are reported in Figure 5.8. Moreover, it should be worth noting that for all mentioned results, we used the normalized ear images.

Although the PCANet and ICANet models are the simplest, the obtained results show the outscoring of CNN-like models (i.e, PCANet and ICANet) against pre-trained models. Moreover, based on the pre-trained CNN results, it is clearly shown that the results obtained based on low layers provide the best results, which prove the effectiveness of the proposed model.

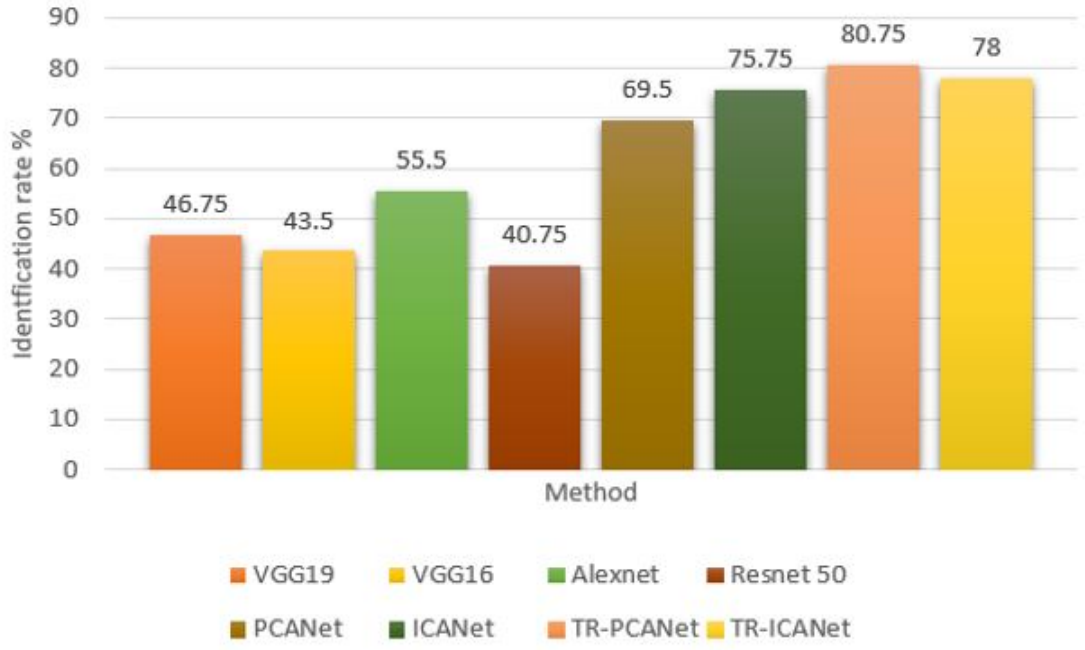


FIGURE 5.8: Comparaison results with PCANet, TR-PCANet and some pre-trained models

For a fair comparison, the proposed ICANet is compared with the original PCANet without considering the TR normalization. From Figure 5.8, it is clearly shown that the ICANet outperforms the PCANet with 6.25%. Besides, the obtained results by adding the TR normalization stage to the original PCANet show and prove that our proposition of unifying the data distribution base on TR normalization has a high positive impact.

Generally speaking, because of the vectorization during the filter learning stage that could be considered as a limitation of the proposed model, both TRICANet and TRPCANet produce a high dimensional feature vector, making matching test probes somewhat computationally heavy in the case of high images size. However, compared to TRPCANet, TRICANet can achieve a faithful representation of image content and better performance and low cost in terms of computational time. To prove this, the speediness of both methods is computed (Table 5.12) where the obtained results show the outperformance of the proposed TRICANet model.

TABLE 5.12: Comparison between the computational time for filters learning and training feature extraction time of TR-ICANet and TR-PCANet

Method	Filter learning time	Feature extraction time
TR-ICANet	4.055 s	15.117 s
TR-PCANet	12.596 s	162.542 s

To assess the performance of our modal and show the statistical significance between TRICANet and TRPCANet models as they gave the best results, we conduct a statistical analysis on their results. The variances analysis (ANOVAs) carried out on the identification rates according to the two models (TRICANet and TRPCANet) by the software SPSS ipm20, shows "non-significant" statistical results with $P = 0.403$ (Table attached). Although the non-significant statistical analysis results show that there is no difference between the two models, our TRICANet model proves a good performance in terms of complexity (execution time -Table 5.13-).

TABLE 5.13: Statistical anlysis using ANOVA results

	Squares sum	ddl	Medium square	F	P
Intergroups	0.38	1	0.038	0.704	0.403
Intragroups	10.642	198	0.054		
Total	10.680	199			

5.2.3.4 Multimodal System Results

As we have mentioned in chapter 4, in this study, we propose to use a multimodal system for human ear recognition. The performance of the proposed multimodal system is tested by fusing four feature extraction methods at score level which are TR-ICANet, TR-PCANet, Alexnet, and VGG19.

Figure 5.9 summarizes the previously obtained results based on the unimodal system and the results obtained with the multimodal system where it is clearly shown the great outperform of this later against unimodal ones:

We opt for using a hybrid model based mainly on the proposed TRICANet for feature extraction. Our proposition is justified by:

1. The positive effect of image preprocessing on the recognition rates. Thus, we suggested reshaping images into a unified format using an efficient CNN-based preprocessing method, in which PCA is used for geometric normalization of the ear. The effect of this normalization is shown in Figure 5.7.
2. To perform classification, we used SVM classifier, which has shown better performance compared to several other classifiers.
3. Moreover, the decision taken by several models is typically better than individual decisions. This is clearly shown in the results obtained by comparing unimodal and multi-modal system (Figure 5.9).

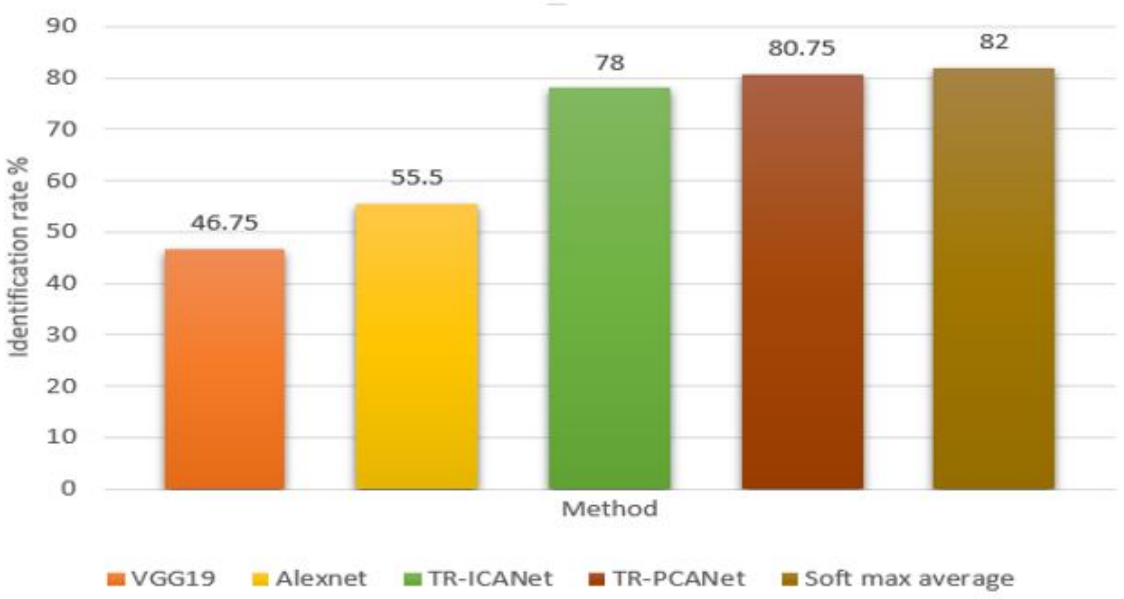
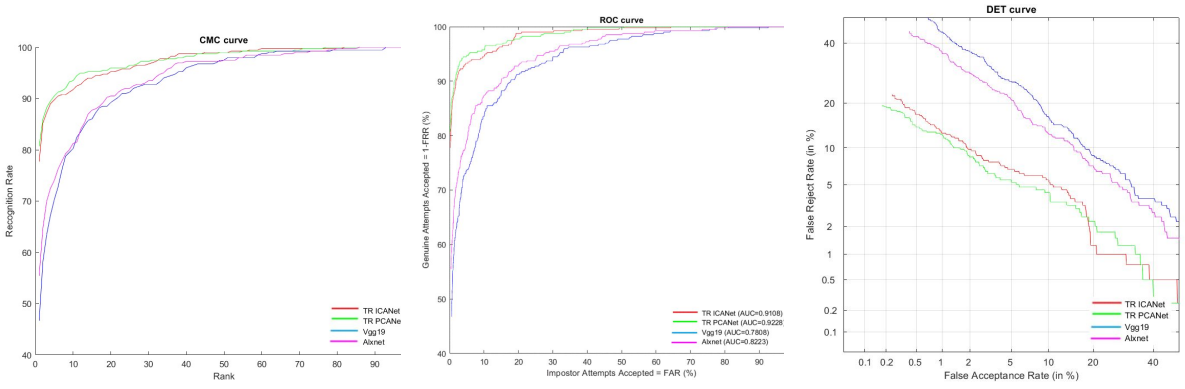


FIGURE 5.9: Experiment results of multimodal system against unimodal ones

To summarize the open set identification mode experiments , a receiver operating characteristic (ROC) which plot the FRR against FAR (Figure 5.10c) and GAR against FAR (Figure 5.10b) is given, where a commutative match curve (CMC) is plotted to explain the closed set identification mode (Figure 5.10a).



(A) (B) (C)

FIGURE 5.10: The commutative match curve CMC (a), the receiver operating characteristic ROC (b), and the detection error tradeoff DET curves.

5.2.3.5 Comparison With State Of The Art

To prove the superiority of the proposed multimodal system for human ear recognition, an experimental comparison with other recent and relevant works has been conducted. For a fair comparison, we follow the same evaluation protocol which is adopted by the compared works.

Tables 5.14 presents the experimental comparison results of the proposed ear recognition system performances and other state-of-the-art systems.

From table 5.14, it is noticed that the proposed method outperforms the performance of the majority of the state-of-the-art methods. To be more accurate, our proposed framework outperformed the approaches presented in recent studies [27, 110] respectively by 3.87 % and 14.75%. This explains and proves that the proposed unimodal (TR-ICANet or TR-PCANet) or multimodal system can play a vital role in recognizing ears in unconstrained conditions.

5.3 Conclusion

The main aim of this chapter is to evaluate the performance of the proposed systems for Arabic handwriting recognition and unconstrained ear recognition. To do so, we divided the experimental into three parts. The first part contained the description of the AHDB and AWE databases which are used for evaluating the performance of the proposed systems. In the second part, we present in detail the experimental of each model (i,e., ML and PML) used for the comparison with our proposed GFIPML model for Arabic handwriting recognition. The parameters tuning of TRICANet, as well as the corresponding experimental setups, were discussed in the third part.

Table 5.14: Comparison with state of the art

Authors	Method/ Descriptor	Identification rate %
Benzaoui et al. [87]	BSIF+SVM	48.4
Dodge et al. [104]	CNN with transfer learning	69.25
Emeršič and Peer [126]	BSIF	48.4
	LBP	43.5
	LPQ	42.8
	HOG	43.9
Hassaballah et al. [91]	AECLBP	49.6
Hassaballah et al.[25]	POEM	62.5
	RLOP	54.1
	HOG	53.9
Omara et al. [27]	ResNet + Learning Mahalanobis distance	78.13
Khaldi and Benzaoui [106]	DCGAN+CNN	50.53
Khaldi et al. [107]	Deep Unsupervised Active Learning with VGG16	51.25
Mewada et al. [20]	Wavelet features embedded CNN	78.88
Alshazly et al. [110]	Deep Residual Networks	67.25
Our proposed method	TR-ICANet	**78**
	TR-PCANet	**80.75**
	Multimodal system based on softmax average	**82**

Chapter 6

Conclusions, Perspectives, and Future Directions

Nowadays, the number of images publicly available online is incredibly increased. Thus the task of automatically analyzing them is becoming quite complicated. Ordinary people can perform various image analysis tasks quickly and easily. However, the performance of human analysis, on the other hand, is proportional to the number of images available.

Visual object recognition (VOR) refers to the task of automatically recognizing an object based on its content. In the last few years, it has become the core of researches and it attracts the attention of researchers due to its vital applications in our daily life including handwriting recognition, date fruit classification, biometric identification..etc. To deal with such issues, several machine learning techniques have been proposed, in the literature, based on several features extraction methods where the features extraction stage represents the challenge of most existing works. However, even with the critical role of features extraction, the preprocessing stage could positively influence the recognition outcomes. Thus, a VOR system performance depends mainly on the performance of the used machine learning techniques.

The work presented in this thesis has two main objectives: the first one is devoted to the conception of a robust offline recognition system for Arabic handwriting. The second objective is to develop an efficient human identification system based on the ear print.

As a first contribution, we proposed a novel Genetic Feature Independent Pyramid Multi-Level (GFIPML) model for features extraction. As its name indicates, the proposed GFIPML is feature-independent and can be used with a variety of existing features extraction methods. We apply GFIPML in the context of Arabic handwriting recognition by using two efficient texture descriptors namely Local Phase Quantization and Binarized Statistical Image Feature (BSIF). We opted for using these texture descriptors as handwritten words are

constituted of homogeneous and repetitive parts of the word (PAWs), in part, and as they proved their efficiency for many recognition tasks in another part. Moreover, we proposed a multimodal offline Arabic handwriting recognition system based on the combination of LPQ and BSIF with three filters at the decision level using LDA classifier. The experimental results conducted on the public AHDB database have proven the effectiveness of the proposed model versus ML and PML as well as the pertinent state-of-the-art methods involving deep learning ones.

On the other hand, ear-based human recognition has become an active area of researches within the biometric community where deep learning-based approaches have been considered and used by several relevant states of the art. However, deep-based schemes are known to be data-hungry and they may require a significant deal of time to perform features learning. Indeed, the ear images taken from unconstrained conditions represent the most challenge in the field of identity identification. The main aim of our second contribution is to present a new simple yet efficient and speedy CNN-like network for feature extraction. The network uses the Independent Component Analysis (ICA) for learning the filters. It contains mainly three stages namely filter learning, binary hashing and block-wise histogramming, and a Tied rank TR normalization for histogram normalization. Thus, we refer the network to as TR-ICANet. Furthermore, To get rid of unconstrained conditions e.g., scale and pose variations, we suggest using a CNN-based normalization for image pre-processing. The performance of the proposed TR-ICANet is validated against several feature extractions methods including PCANet, VGG with two extensions (VGG19 and VGG16), Resnet, and Alexnet. To further improve the yield recognition rate, we propose a multimodal system based on fusing several models at the score level. The obtained results, conducted on the public AWE database, have proven the efficiency of the proposed scheme against the relevant state-of-the-art methods.

The findings of this thesis complement those of earlier studies and make several noteworthy contributions to visual object recognition domain. In the future, further investigations can be carried out to improve the proposed models. Because the models are composed of different parameters, improvements can be done on the level of each parameter. For instance, we intend to evaluate an enhanced version of our proposed TR-ICANet model, which will be more optimized by taking into consideration its shortcoming (i.e, solve the problem of vectorization in the case of high image size, add max-pooling layers to reduce features dimension) which could be applied for other complex databases. On the other hand, we intend to use other approaches to extract features based on the GFIPML model and apply them to other domains.

Appendix A

Personal Contributions

A.1 Publications

1. Korichi, A., Slatnia, S. & Aiadi, O. TR-ICANet: A Fast Unsupervised Deep-Learning-Based Scheme for Unconstrained Ear Recognition. Arab J Sci Eng (2022). https://doi.org/10.1007/s13369-021-06375-z
2. Korichi, A., Slatnia, S., Aiadi, O. et al. A generic feature-independent pyramid multilevel model for Arabic handwriting recognition. Multimed Tools Appl (2022). https://doi.org/10.1007/s11042-022-11979-0

A.2 Chapter Books

1. Korichi, A., Slatnia, S., Tagougui, N., Zouari, R., Kherallah, M., Aiadi, O. (2022). Recognizing Arabic Handwritten Literal Amount Using Convolutional Neural Networks. In: Lejdel, B., Clementini, E., Alarabi, L. (eds) Artificial Intelligence and Its Applications. AIAP 2021. Lecture Notes in Networks and Systems, vol 413. Springer, Cham. https://doi.org/10.1007/978-3-030-96311-8_15.

A.3 International Communications indexed in the IEEE xplore database

1. Korichi, A., Slatnia, S., Aiadi, O., Tagougui, N., and Kherallah, M. (2020, November). Arabic handwriting recognition: Between handcrafted methods and deep learning techniques. *In 2020 21st International Arab Conference on Information Technology (ACIT) (pp. 1-6). IEEE.*

2. Korichi, A., Aiadi, O., Khaldi, B., Slatnia, S., and Kherfi, M. L. (2018, November). Off-line Arabic handwriting recognition system based on ML-LPQ and classifiers combination. *In 2018 International Conference on Signal, Image, Vision and their Applications (SIVA) (pp. 1-6). IEEE.*

A.4 Non indexed International Communications

1. Korichi, A., Aiadi, O., Khaldi, B., Kherfi, M. L., and Slatnia, S. (2019, March). A Comparative Study on Arabic Handwritten Words Recognition Using Textures Descriptors. *In 2019 International Conference on Artificial Intelligence and Information Technology (ICA2IT)*

Bibliography

[1] Antonio Chella, Marcello Frixione, and Salvatore Gaglio. "Understanding dynamic scenes." In: *Artificial intelligence* 123.1-2 (2000), pp. 89–132.

[2] Gerhard Sagerer and Heinrich Niemann. *Semantic networks for understanding scenes*. Springer Science & Business Media, 2013.

[3] Kevin P Murphy. *Machine learning: a probabilistic perspective*. MIT press, 2012.

[4] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. "Deep learning." In: *nature* 521.7553 (2015), pp. 436–444.

[5] Bouchra El Qacimy, Ahmed Hammouch, and Mounir Ait Kerroum. "A review of feature extraction techniques for handwritten Arabic text recognition." In: *2015 International Conference on Electrical and Information Technologies (ICEIT)*. IEEE. 2015, pp. 241–245.

[6] O Guehairia et al. "Feature fusion via Deep Random Forest for facial age estimation." In: *Neural Networks* 130 (2020), pp. 238–252.

[7] Yichun Shi et al. "Towards universal representation learning for deep face recognition." In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2020, pp. 6817–6826.

[8] Zhao Pei et al. "Face recognition via deep learning using data augmentation based on orthogonal experiments." In: *Electronics* 8.10 (2019), p. 1088.

[9] Umara Zafar et al. "Face recognition with Bayesian convolutional networks for robust surveillance systems." In: *EURASIP Journal on Image and Video Processing* 2019.1 (2019), pp. 1–10.

[10] Ping Yan and Kevin W Bowyer. "Biometric recognition using 3D ear shape." In: *IEEE Transactions on pattern analysis and machine intelligence* 29.8 (2007), pp. 1297–1308.

[11] Chiarella Sforza et al. "Age-and sex-related changes in the normal human ear." In: *Forensic science international* 187.1-3 (2009), 110–e1.

[12] Hossein Nejati et al. “Wonder ears: Identification of identical twins from ear images.” In: *Proceedings of the 21st International Conference on Pattern Recognition (ICPR2012)*. IEEE. 2012, pp. 1201–1204.

[13] Yuxiang Zhou and Stefanos Zaferiou. “Deformable models of ears in-the-wild for alignment and recognition.” In: *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*. IEEE. 2017, pp. 626–633.

[14] Qais Ali Al-Nuzaili et al. “Pixel distribution-based features for offline Arabic handwritten word recognition.” In: *International Journal of Computational Vision and Robotics* 7.1-2 (2017), pp. 99–122.

[15] Nadir Farah, Labiba Souici, and Mokhtar Sellami. “Classifiers combination and syntax analysis for Arabic literal amount recognition.” In: *Engineering Applications of Artificial Intelligence* 19.1 (2006), pp. 29–39.

[16] Qais Al-Nuzaili et al. “Enhanced structural perceptual feature extraction model for Arabic literal amount recognition.” In: *International Journal of Intelligent Systems Technologies and Applications* 15.3 (2016), pp. 240–254.

[17] Mohammed O Assayony and Sabri A Mahmoud. “Recognition of Arabic handwritten words using Gabor-based bag-of-features framework.” In: *International Journal of Computing and Digital Systems* 7.01 (2018), pp. 35–42.

[18] Qais Al-Nuzaili et al. “An enhanced quadratic angular feature extraction model for arabic handwritten literal amount recognition.” In: *International Conference of Reliable Information and Communication Technology*. Springer. 2017, pp. 369–377.

[19] Hanadi Hassen and Somaya Al-Maadeed. “Arabic handwriting recognition using sequential minimal optimization.” In: *2017 1st International Workshop on Arabic Script Analysis and Recognition (ASAR)*. IEEE. 2017, pp. 79–84.

[20] Moumen El-Melegy et al. “Recognition of Arabic handwritten literal amounts using deep convolutional neural networks.” In: *Iberian Conference on Pattern Recognition and Image Analysis*. Springer. 2019, pp. 169–176.

[21] Mohamed Eltay, Abdelmalek Zidouri, and Irfan Ahmad. “Exploring deep learning approaches to recognize handwritten arabic texts.” In: *IEEE Access* 8 (2020), pp. 89882–89898.

[22] Azzeddine Menasria et al. “Multiclassifiers system for handwritten Arabic literal amounts recognition based on enhanced feature extraction model.” In: *Journal of Electronic Imaging* 27.3 (2018), p. 033024.

[23] Somaya Alma'adeed, Colin Higgins, and Dave Elliman. "Off-line recognition of handwritten Arabic words using multiple hidden Markov models." In: *International Conference on Innovative Techniques and Applications of Artificial Intelligence*. Springer. 2003, pp. 33–40.

[24] Alia Karim Abdul Hassan and Mustafa S Kadhm. "Handwriting word recognition based on neural networks." In: *International Journal of Applied Engineering Research* 10.22 (2015), pp. 43120–43124.

[25] M Hassaballah, HA Alshazly, and Abdelmgeid A Ali. "Robust local oriented patterns for ear recognition." In: *Multimedia Tools and Applications* 79.41 (2020), pp. 31183–31204.

[26] Ibrahim Omara et al. "A novel geometric feature extraction method for ear recognition." In: *Expert Systems with Applications* 65 (2016), pp. 127–135.

[27] Ibrahim Omara et al. "A novel approach for ear recognition: learning Mahalanobis distance features from deep CNNs." In: *Machine Vision and Applications* 32.1 (2021), pp. 1–14.

[28] Ramar Ahila Priyadharshini, Selvaraj Arivazhagan, and Madakannu Arun. "A deep learning approach for person identification using ear biometrics." In: *Applied Intelligence* (2020), pp. 1–12.

[29] Aicha Korichi et al. "A generic feature-independent pyramid multilevel model for Arabic handwriting recognition." In: *Multimedia Tools and Applications* (2022), pp. 1–21.

[30] Aicha Korichi, Sihem Slatnia, and Oussama Aiadi. "TR-ICANet: A Fast Unsupervised Deep-Learning-Based Scheme for Unconstrained Ear Recognition." In: *Arabian Journal for Science and Engineering* (2022), pp. 1–12.

[31] Simon Haykin and Richard Lippmann. "Neural networks, a comprehensive foundation." In: *International journal of neural systems* 5.4 (1994), pp. 363–364.

[32] Sotiris B Kotsiantis, Ioannis Zaharakis, P Pintelas, et al. "Supervised machine learning: A review of classification techniques." In: *Emerging artificial intelligence applications in computer engineering* 160.1 (2007), pp. 3–24.

[33] Ulrike Von Luxburg and Bernhard Schölkopf. "Statistical learning theory: Models, concepts, and results." In: *Handbook of the History of Logic*. Vol. 10. Elsevier, 2011, pp. 651–706.

[34] Michael Biehl. "Supervised Learning–An Introduction." In: *Retrieved Enero* 2 (2019), p. 2021.

[35] Horace B Barlow. "Unsupervised learning." In: *Neural computation* 1.3 (1989), pp. 295–311.

[36] Saurav Kaushik. “An Introduction to Clustering and different methods of clustering.” In: *Analytics Vidhya* 3 (2016).

[37] Petr Berka and Jan Rauch. “Machine learning and association rules.” In: *University of Economics* (2010).

[38] Mohammad S Khorsheed. “Off-line Arabic character recognition–a review.” In: *Pattern analysis & applications* 5.1 (2002), pp. 31–45.

[39] Liana M Lorigo and Venugopal Govindaraju. “Offline Arabic handwriting recognition: a survey.” In: *IEEE transactions on pattern analysis and machine intelligence* 28.5 (2006), pp. 712–724.

[40] Bing Quan Huang, YB Zhang, and Mohand Tahar Kechadi. “Preprocessing techniques for online handwriting recognition.” In: *Seventh International Conference on Intelligent Systems Design and Applications (ISDA 2007)*. IEEE. 2007, pp. 793–800.

[41] Hideyuki Tamura, Shunji Mori, and Takashi Yamawaki. “Textural features corresponding to visual perception.” In: *IEEE Transactions on Systems, man, and cybernetics* 8.6 (1978), pp. 460–473.

[42] Belal Khaldi and Mohammed Lamine Kherfi. “Modified integrative color intensity co-occurrence matrix for texture image representation.” In: *Journal of Electronic Imaging* 25.5 (2016), p. 053007.

[43] John G Daugman. “Two-dimensional spectral analysis of cortical receptive field profiles.” In: *Vision research* 20.10 (1980), pp. 847–856.

[44] John G Daugman. “Uncertainty relation for resolution in space, spatial frequency, and orientation optimized by two-dimensional visual cortical filters.” In: *JOSA A* 2.7 (1985), pp. 1160–1169.

[45] Robert M Haralick, Karthikeyan Shanmugam, and Its’ Hak Dinstein. “Textural features for image classification.” In: *IEEE Transactions on systems, man, and cybernetics* 6 (1973), pp. 610–621.

[46] Song-Chun Zhu et al. “What are textons?” In: *International Journal of Computer Vision* 62.1 (2005), pp. 121–143.

[47] Zhi-Gang Fan et al. “Local patterns constrained image histograms for image retrieval.” In: *2008 15th IEEE International Conference on Image Processing*. IEEE. 2008, pp. 941–944.

[48] Timo Ojala, Matti Pietikäinen, and David Harwood. “A comparative study of texture measures with classification based on featured distributions.” In: *Pattern recognition* 29.1 (1996), pp. 51–59.

[49] Caifeng Shan, Shaogang Gong, and Peter W McOwan. "Facial expression recognition based on local binary patterns: A comprehensive study." In: *Image and vision Computing* 27.6 (2009), pp. 803–816.

[50] Di Huang et al. "Local binary patterns and its application to facial image analysis: a survey." In: *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)* 41.6 (2011), pp. 765–781.

[51] Ville Ojansivu and Janne Heikkilä. "Blur insensitive texture classification using local phase quantization." In: *International conference on image and signal processing*. Springer. 2008, pp. 236–243.

[52] Juho Kannala and Esa Rahtu. "Bsif: Binarized statistical image features." In: *Proceedings of the 21st international conference on pattern recognition (ICPR2012)*. IEEE. 2012, pp. 1363–1366.

[53] Aapo Hyvärinen and Erkki Oja. "Independent component analysis: algorithms and applications." In: *Neural networks* 13.4-5 (2000), pp. 411–430.

[54] Navneet Dalal and Bill Triggs. "Histograms of oriented gradients for human detection." In: *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*. Vol. 1. Ieee. 2005, pp. 886–893.

[55] Qais Al-Nuzaili et al. "Arabic bank cheque words recognition using Gabor features." In: *2018 IEEE 2nd International Workshop on Arabic and Derived Script Analysis and Recognition (ASAR)*. IEEE. 2018, pp. 84–89.

[56] K Fukushima GI. "A Hierarchical Neural Network Capable of Visual Pattern Recognition." In: *Neural Network* 1 (1989).

[57] Yanming Guo et al. "Deep learning for visual understanding: A review." In: *Neurocomputing* 187 (2016), pp. 27–48.

[58] Luiz G Hafemann, Robert Sabourin, and Luiz S Oliveira. "Learning features for offline handwritten signature verification using deep convolutional neural networks." In: *Pattern Recognition* 70 (2017), pp. 163–176.

[59] Lianwen Jin et al. "Online handwritten Chinese character recognition: from a bayesian approach to deep learning." In: *Advances in Chinese Document and Text Processing*. World Scientific, 2017, pp. 79–126.

[60] Fenfen Sheng et al. "End-to-end chinese image text recognition with attention model." In: *International Conference on Neural Information Processing*. Springer. 2017, pp. 180–189.

[61] Tsung-Han Chan et al. "PCANet: A simple deep learning baseline for image classification?" In: *IEEE transactions on image processing* 24.12 (2015), pp. 5017–5032.

[62] Timo Ahonen, Abdenour Hadid, and Matti Pietikainen. "Face description with local binary patterns: Application to face recognition." In: *IEEE transactions on pattern analysis and machine intelligence* 28.12 (2006), pp. 2037–2041.

[63] Karen Simonyan and Andrew Zisserman. "Very deep convolutional networks for large-scale image recognition." In: *arXiv preprint arXiv:1409.1556* (2014).

[64] Md Zahangir Alom et al. "The history began from alexnet: A comprehensive survey on deep learning approaches." In: *arXiv preprint arXiv:1803.01164* (2018).

[65] Ken Chatfield et al. "Return of the devil in the details: Delving deep into convolutional nets." In: *arXiv preprint arXiv:1405.3531* (2014).

[66] Christian Szegedy et al. "Going deeper with convolutions." In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015, pp. 1–9.

[67] Mostafa Mehdipour Ghazi, Berrin Yanikoglu, and Erchan Aptoula. "Plant identification using deep neural networks via optimization of transfer learning parameters." In: *Neurocomputing* 235 (2017), pp. 228–235.

[68] Kaiming He et al. "Deep residual learning for image recognition." In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 770–778.

[69] Salah Eddine Bekhouche et al. "Facial age estimation and gender classification using multi level local phase quantization." In: *2015 3rd International Conference on Control, Engineering & Information Technology (CEIT)*. IEEE. 2015, pp. 1–4.

[70] Salah Eddine Bekhouche et al. "Pyramid multi-level features for facial demographic estimation." In: *Expert Systems with Applications* 80 (2017), pp. 297–310.

[71] Vladimir N Vapnik. "An overview of statistical learning theory." In: *IEEE transactions on neural networks* 10.5 (1999), pp. 988–999.

[72] S Amarappa and SV Sathyanarayana. "Data classification using Support vector Machine (SVM), a simplified approach." In: *Int. J. Electron. Comput. Sci. Eng* 3 (2014), pp. 435–445.

[73] Chih-Wei Hsu and Chih-Jen Lin. "A comparison of methods for multiclass support vector machines." In: *IEEE transactions on Neural Networks* 13.2 (2002), pp. 415–425.

[74] Liang Huang et al. "An adaptive nonparametric discriminant analysis method and its application to face recognition." In: *Asian Conference on Computer Vision*. Springer. 2007, pp. 680–689.

[75] David Zhang et al. *Advanced pattern recognition technologies with applications to biometrics*. IGI Global, 2009.

[76] Tin Kam Ho. "Multiple classifier combination: Lessons and next steps." In: *Hybrid methods in pattern recognition*. World Scientific, 2002, pp. 171–198.

[77] Mohamed Cheriet et al. *Character recognition systems: a guide for students and practitioners*. John Wiley & Sons, 2007.

[78] Nafiz Arica and Fatos T Yarman-Vural. "An overview of character recognition focused on off-line handwriting." In: *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)* 31.2 (2001), pp. 216–233.

[79] Faouzi Zaiz, Mohamed Chaouki Babahenini, and Abdelhamid Djeffal. "Puzzle based system for improving Arabic handwriting recognition." In: *Engineering Applications of Artificial Intelligence* 56 (2016), pp. 222–229.

[80] Yasser Qawasmeh et al. "Local patterns for offline Arabic handwritten recognition." In: *International Journal of Advanced Intelligence Paradigms* 16.2 (2020), pp. 203–215.

[81] Aicha Korichi et al. "Arabic handwriting recognition: Between handcrafted methods and deep learning techniques." In: *2020 21st International Arab Conference on Information Technology (ACIT)*. IEEE. 2020, pp. 1–6.

[82] Asmae Lamsaf et al. "Recognition of Arabic Handwritten Text by Integrating N-gram Model." In: *The Proceedings of the Third International Conference on Smart City Applications*. Springer. 2020, pp. 1490–1502.

[83] Qi Wang et al. "Hierarchical feature selection for random projection." In: *IEEE transactions on neural networks and learning systems* 30.5 (2018), pp. 1581–1586.

[84] Qi Wang, Qiang Li, and Xuelong Li. "Hyperspectral image super-resolution using spectrum and feature context." In: *IEEE Transactions on Industrial Electronics* (2020).

[85] Ayman Abaza et al. "A survey on ear biometrics." In: *ACM computing surveys (CSUR)* 45.2 (2013), pp. 1–35.

[86] Anika Pflug and Christoph Busch. "Ear biometrics: a survey of detection, feature extraction and recognition methods." In: *IET biometrics* 1.2 (2012), pp. 114–129.

[87] Amir Benzaoui, Abdenour Hadid, and Abdelhani Boukrouche. "Ear biometric recognition using local texture descriptors." In: *Journal of electronic imaging* 23.5 (2014), p. 053008.

[88] Maarouf Korichi, Abdallah Meraoumia, and Kamal Aiadi. "A small look at the ear recognition process using a Binarized Statistical Image Features (ML-BSIF)." In: ().

[89] Lamis Ghoualmi, Amer Draa, and Salim Chikhi. "An ear biometric system based on artificial bees and the scale invariant feature transform." In: *Expert Systems with Applications* 57 (2016), pp. 49–61.

[90] Parmeshwar Birajadar et al. "Unconstrained ear recognition using deep scattering wavelet network." In: *2019 IEEE Bombay Section Signature Conference (IBSSC)*. IEEE. 2019, pp. 1–6.

[91] M Hassaballah, Hammam A Alshazly, and Abdelmgeid A Ali. "Ear recognition using local binary patterns: A comparative experimental study." In: *Expert Systems with Applications* 118 (2019), pp. 182–200.

[92] Zhichun Mu et al. "Shape and structural feature based ear recognition." In: *Chinese Conference on Biometric Recognition*. Springer. 2004, pp. 663–670.

[93] Dasari Shailaja and Phalguni Gupta. "A simple geometric approach for ear recognition." In: *9th International Conference on Information Technology (ICIT'06)*. IEEE. 2006, pp. 164–167.

[94] Hammam A Alshazly et al. "Ear biometric recognition using gradient-based feature descriptors." In: *International conference on advanced intelligent systems and informatics*. Springer. 2018, pp. 435–445.

[95] Asmaa Sabet Anwar, Kareem Kamal A Ghany, and Hesham Elmahdy. "Human ear recognition using geometrical features extraction." In: *Procedia Computer Science* 65 (2015), pp. 529–537.

[96] Kyong Chang et al. "Comparison and combination of ear and face images in appearance-based biometrics." In: *IEEE Transactions on pattern analysis and machine intelligence* 25.9 (2003), pp. 1160–1165.

[97] Loris Nanni and Alessandra Lumini. "Fusion of color spaces for ear authentication." In: *Pattern Recognition* 42.9 (2009), pp. 1906–1913.

[98] Ajay Kumar and Chenye Wu. "Automated human identification using ear imaging." In: *Pattern Recognition* 45.3 (2012), pp. 956–968.

[99] Anam Tariq and M Usman Akram. "Personal identification using ear recognition." In: *TELKOMNIKA Telecommun. Comput. Electron. Control* 10.2 (2012), pp. 321–326.

[100] Baoqing Zhang et al. "Robust classification for occluded ear via Gabor scale feature-based non-negative sparse representation." In: *Optical Engineering* 53.6 (2013), p. 061702.

[101] Akshay Kumar Goel et al. "Profit or Loss: A Long Short Term Memory based model for the Prediction of share price of DLF group in India." In: *2019 IEEE 9th International Conference on Advanced Computing (IACC)*. IEEE. 2019, pp. 120–124.

[102] Mohit Agarwal et al. "A convolution neural network based approach to detect the disease in corn crop." In: *2019 IEEE 9th international conference on advanced computing (IACC)*. IEEE. 2019, pp. 176–181.

[103] Kapil Sethi, Varun Jaiswal, and Mohammad D Ansari. "Machine learning based support system for students to select stream (subject)." In: *Recent Advances in Computer Science and Communications (Formerly: Recent Patents on Computer Science)* 13.3 (2020), pp. 336–344.

[104] Samuel Dodge, Jinane Mounsef, and Lina Karam. "Unconstrained ear recognition using deep neural networks." In: *IET Biometrics* 7.3 (2018), pp. 207–214.

[105] Hammam Alshazly et al. "Ensembles of deep learning models and transfer learning for ear recognition." In: *Sensors* 19.19 (2019), p. 4139.

[106] Yacine Khaldi and Amir Benzaoui. "A new framework for grayscale ear images recognition using generative adversarial networks under unconstrained conditions." In: *Evolving Systems* (2020), pp. 1–12.

[107] Yacine Khaldi et al. "Ear Recognition Based on Deep Unsupervised Active Learning." In: *IEEE Sensors Journal* (2021).

[108] Hiren K Mewada et al. "Wavelet features embedded convolutional neural network for multiscale ear recognition." In: *Journal of Electronic Imaging* 29.4 (2020), p. 043029.

[109] Aman Kamboj, Rajneesh Rani, and Aditya Nigam. "A comprehensive survey and deep learning-based approach for human recognition using ear biometric." In: *The Visual Computer* (2021), pp. 1–34.

[110] Hammam Alshazly et al. "Towards Explainable Ear Recognition Systems Using Deep Residual Networks." In: *IEEE Access* (2021).

[111] Amir Benzaoui, Ali Kheider, and Abdelhani Boukrouche. "Ear description and recognition using ELBP and wavelets." In: *2015 International Conference on Applied Research In Computer Science And Engineering (Icar)*. IEEE. 2015, pp. 1–6.

[112] Tian Ying, Zhang Debin, and Zhang Baihuan. "Ear recognition based on weighted wavelet transform and DCT." In: *The 26th Chinese Control and Decision Conference (2014 CCDC)*. IEEE. 2014, pp. 4410–4414.

[113] Aythami Morales et al. "Earprint recognition based on an ensemble of global and local features." In: *2015 International Carnahan Conference on Security Technology (ICCST)*. IEEE. 2015, pp. 253–258.

[114] Shabbou Sajadi and Abdolhossein Fathi. "Genetic algorithm based local and global spectral features extraction for ear recognition." In: *Expert Systems with Applications* 159 (2020), p. 113639.

[115] Earnest E Hansley, Maurício Pamplona Segundo, and Sudeep Sarkar. "Employing fusion of learned and handcrafted features for unconstrained ear recognition." In: *IET Biometrics* 7.3 (2018), pp. 215–223.

[116] Yousri Kessentini. "Modèles de Markov multi-flux pour la reconnaissance de l'écriture manuscrite multi-scripts." PhD thesis. Université de Rouen, 2009.

[117] Christian Rathgeb, Florian Struck, and Christoph Busch. "Efficient BSIF-based near-infrared iris recognition." In: *2016 Sixth International Conference on Image Processing Theory, Tools and Applications (IPTA)*. IEEE. 2016, pp. 1–6.

[118] Ramachandra Raghavendra and Christoph Busch. "Texture based features for robust palmprint recognition: a comparative study." In: *EURASIP Journal on Information Security* 2015.1 (2015), pp. 1–9.

[119] Cong Jie Ng and Andrew Beng Jin Teoh. "DCTNet: A simple learning-free approach for face recognition." In: *2015 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA)*. IEEE. 2015, pp. 761–768.

[120] Miron Ivanov. "Comparison of PCA with ICA from data distribution perspective." In: *arXiv preprint arXiv:1709.10222* (2017).

[121] Anthony R McIntosh and Bratislav Mišić. "Multivariate statistical analyses for neuroimaging data." In: *Annual review of psychology* 64 (2013), pp. 499–525.

[122] Tianyu Geng et al. "Unsupervised Feature Learning with Single Layer ICANet for Face Recognition." In: *Sensing and Imaging* 19.1 (2018), pp. 1–10.

[123] Yongqing Zhang et al. "ICANet: a simple cascade linear convolution network for face recognition." In: *EURASIP Journal on Image and Video Processing* 2018.1 (2018), pp. 1–7.

[124] Aapo Hyvärinen, Jarmo Hurri, and Patrick O Hoyer. *Natural image statistics: A probabilistic approach to early computational vision.* Vol. 39. Springer Science & Business Media, 2009.

[125] Somaya Al-Ma'adeed, Dave Elliman, and Colin A Higgins. "A data base for Arabic handwritten text recognition research." In: *Proceedings eighth international workshop on frontiers in handwriting recognition*. IEEE. 2002, pp. 485–489.

[126] Žiga Emeršič, Vitomir Štruc, and Peter Peer. "Ear recognition: More than a survey." In: *Neurocomputing* 255 (2017), pp. 26–39.

[127] Mustafa S Kadhm and Alia Karim Abdul Hassan. "Handwriting word recognition based on SVM classifier." In: *International Journal of Advanced Computer Science & Applications* 1.6 (2015), pp. 64–68.